THE ADVANCED THEORY OF STATISTICS

MAURICE G. KENDALL, M.A., Sc.D.
Chairman, Scientific Control Systems Ltd. • Formerly Professor of Statistics in the University of London • President of the Royal Statistical Society, 1960-2

and

ALAN STUART, D.Sc. (ECON.)
Professor of Statistics in the University of London

IN THREE VOLUMES

VOLUME 3

DESIGN AND ANALYSIS, AND TIME-SERIES

SECOND EDITION

CHARLES GRIFFIN & COMPANY LIMITED
LONDON

М. КЕНДАЛЛ, А. СТЬЮАРТ

МНОГОМЕРНЫЙ СТАТИСТИЧЕСКИЙ АНАЛИЗ И ВРЕМЕННЫЕ РЯДЫ

Перевод с английского
Э. Л. ПРЕСМАНА, В. И. РОТАРЯ

Под редакцией
А. Н. КОЛМОГОРОВА, Ю. В. ПРОХОРОВА

ИЗДАТЕЛЬСТВО «НАУКА»
ГЛАВНАЯ РЕДАКЦИЯ
ФИЗИКО-МАТЕМАТИЧЕСКОЙ ЛИТЕРАТУРЫ
МОСКВА 1976

517.8
К 35
УДК 519.24

Многомерный статистический анализ и временные ряды, М. Кендалл, А. Стьюарт, Главная редакция физико-математической литературы изд-ва «Наука», 1976.

Книга является последним томом трехтомного курса статистики М. Кендалла и А. Стьюарта, первый том которого вышел в 1966 г. под названием «Теория распределений», а второй — в 1973 г. под названием «Статистические выводы и связи». В книге содержатся сведения по дисперсионному анализу, планированию экспериментов, теории выборочных обследований, многомерному анализу и временным рядам. Как и первые два тома, книга содержит много практических рекомендаций и примеров их применения, а изложение сочетает более или менее подробный вывод основных результатов с относительно кратким перечислением большого количества более частных сведений.

Книга будет представлять интерес для студентов и аспирантов, специализирующихся в области математической статистики, а также для широкого круга научных работников, имеющих дело с ее приложениями

Библ. 563 назв. Илл. 23.

Морис Кендалл, Алан Стьюарт

Многомерный статистический анализ и временные ряды

М., 1976 г., 736 стр. с илл.

Редактор *С. А. Айвазян*
Техн. редактор *В. Н. Кондакова*
Корректоры *О. А. Сигал, М. Л. Медведская*

Сдано в набор 9/I 1976 г. Подписано к печати 1/VI 1976 г. Бумага 60×90¹/₁₆ тип. № 1. Физ. печ. л. 46. Усл. печ. л. 46. Уч.-изд. л. 47,89. Тираж 16 000 экз. Цена книги 3 р. 61 к. Заказ № 27.

Издательство «Наука»
Главная редакция физико-математической литературы
117071, Москва, В-71, Ленинский проспект, 15

Ордена Трудового Красного Знамени
Ленинградская типография № 2 имени Евгении Соколовой
Союзполиграфпрома при Государственном комитете Совета
Министров СССР по делам издательств, полиграфии
и книжной торговли. 198052, Ленинград, Л-52,
Измайловский пр., 29.

К $\frac{20203-080}{053(02)-76}$ 44-75

© Перевод на русский язык.
Главная редакция
физико-математической литературы
издательства «Наука», 1976

ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ К ПЕРВОМУ ИЗДАНИЮ . . . 9
ПРЕДИСЛОВИЕ КО ВТОРОМУ ИЗДАНИЮ . . . 10
СПИСОК СОКРАЩЕНИЙ . . . 10

ГЛАВЫ

35. ДИСПЕРСИОННЫЙ АНАЛИЗ В ЛИНЕЙНОЙ МОДЕЛИ: КЛАССИФИЦИРОВАННЫЕ ДАННЫЕ . . . 11
36. ДРУГИЕ МОДЕЛИ ДИСПЕРСИОННОГО АНАЛИЗА . . . 88
37. ПРЕДПОЛОЖЕНИЯ ДИСПЕРСИОННОГО АНАЛИЗА . . . 129
38. ПЛАНИРОВАНИЕ ЭКСПЕРИМЕНТОВ . . . 175
39. ТЕОРИЯ ВЫБОРОЧНЫХ ОБСЛЕДОВАНИЙ: ПЛАНИРОВАНИЕ . . . 240
40. ТЕОРИЯ ВЫБОРОЧНЫХ ОБСЛЕДОВАНИЙ: ДОПОЛНИТЕЛЬНАЯ ИНФОРМАЦИЯ . . . 301
41. ТЕОРИЯ МНОГОМЕРНЫХ РАСПРЕДЕЛЕНИЙ . . . 340
42. ПРОВЕРКА ГИПОТЕЗ В МНОГОМЕРНОМ АНАЛИЗЕ . . . 372
43. КАНОНИЧЕСКИЕ ВЕЛИЧИНЫ . . . 400
44. ДИСКРИМИНАЦИЯ И КЛАССИФИКАЦИЯ . . . 437
45. ВРЕМЕННЫЕ РЯДЫ. ОБЩАЯ ЧАСТЬ . . . 474
46. ВРЕМЕННЫЕ РЯДЫ. ТРЕНД И СЕЗОННОСТЬ . . . 505
47. СТАЦИОНАРНЫЕ ВРЕМЕННЫЕ РЯДЫ . . . 555
48. ВЫБОРОЧНАЯ ТЕОРИЯ СЕРИАЛЬНЫХ КОРРЕЛЯЦИЙ . . . 590
49. СПЕКТРАЛЬНАЯ ТЕОРИЯ . . . 620
50. ВРЕМЕННЫЕ РЯДЫ. НЕКОТОРЫЕ ДОПОЛНИТЕЛЬНЫЕ АСПЕКТЫ . . . 644
ПОСЛЕСЛОВИЕ К ТРЕТЬЕМУ ТОМУ . . . 685

Приложение . . . 686

ТАБЛИЦЫ

1. ФУНКЦИЯ ПЛОТНОСТИ НОРМАЛЬНОГО РАСПРЕДЕЛЕНИЯ . . . 686
2. ФУНКЦИЯ НОРМАЛЬНОГО РАСПРЕДЕЛЕНИЯ . . . 686
3. КВАНТИЛИ χ^2-РАСПРЕДЕЛЕНИЯ . . . 688
4а. ФУНКЦИЯ χ^2-РАСПРЕДЕЛЕНИЯ С ОДНОЙ СТЕПЕНЬЮ СВОБОДЫ ДЛЯ $0 \leqslant \chi^2 \leqslant 1$ С ШАГОМ 0,01 . . . 689
4б. ФУНКЦИЯ χ^2-РАСПРЕДЕЛЕНИЯ С ОДНОЙ СТЕПЕНЬЮ СВОБОДЫ ДЛЯ $1 \leqslant \chi^2 \leqslant 10$ С ШАГОМ 0,1 . . . 690

5. КВАНТИЛИ t-РАСПРЕДЕЛЕНИЯ 691
6. 5%-НЫЕ ТОЧКИ Z-РАСПРЕДЕЛЕНИЯ 693
7. 5%-НЫЕ ТОЧКИ F-РАСПРЕДЕЛЕНИЯ 695
8. 1%-НЫЕ ТОЧКИ Z-РАСПРЕДЕЛЕНИЯ 696
9. 1%-НЫЕ ТОЧКИ F-РАСПРЕДЕЛЕНИЯ 698
10. СИММЕТРИЧЕСКИЕ ФУНКЦИИ. ФОРМУЛЫ ПРЕДСТАВЛЕНИЙ РАСШИРЕННЫХ СИММЕТРИЧЕСКИХ ФУНКЦИЙ В ТЕРМИНАХ СУММ СТЕПЕНЕЙ И ОБРАТНЫЕ ФОРМУЛЫ 699

ЦИТИРОВАННАЯ ЛИТЕРАТУРА 700

УКАЗАТЕЛЬ 723

«...не проворным достается успешный бег, не храбрым — победа, не мудрым — хлеб, и не у разумных богатство, и не искусным — благорасположение, но время и случай для всех их».

Экклезиаст, 9.11

ПРЕДИСЛОВИЕ К ПЕРВОМУ ИЗДАНИЮ

Читателю предлагается последний том нашей монографии. Его написание заняло больше времени, чем мы ожидали. Отчасти это объясняется нашей загруженностью другой работой, а отчасти тем, что за последние годы было получено большое число результатов, относящихся к вопросам, рассматриваемым в этом томе. Этими вопросами являются дисперсионный анализ, планирование экспериментов, теория выборочных обследований, многомерный анализ и временные ряды. В бурном потоке текущих исследований становится все труднее понять, что является окончательным результатом, а что только промежуточным этапом развития теории. При решении вопроса, что включать в изложение, а что опускать, были случаи, когда мы вынуждены были вспоминать слова Теккерея о Маколее, что тот прочитывал книгу, чтобы написать предложение.

Этот том, как и первые два, самостоятелен в следующих трех отношениях: в нем содержится собственный список литературы; в приложении приведены те таблицы, которые необходимы для чтения текста; имеется собственный указатель. Теперь, когда вышла в свет «Библиография литературы по статистике» Кендалла и Дойга — нет нужды приводить исчерпывающий список литературы по всем томам. Как и ранее, в конце глав приводится большое число упражнений.

Мы снова пользуемся случаем поблагодарить м-ра Е. В. Бёрка (E. V. Burke) из издательства Charless Griffin and Company Limited за его труд, связанный с выходом в свет этой книги.

Мы благодарны также многим читателям и критикам, указавшим опечатки, неточности изложения и места, требующие пояснений, в первых двух томах, и будем рады любым подобным замечаниям к этому последнему тому.

Лондон,
август, 1966

М. Дж. К.
А. С.

ПРЕДИСЛОВИЕ КО ВТОРОМУ ИЗДАНИЮ

Это издание мало отличается от первого. Сделан ряд незначительных изменений в тексте, добавлено несколько новых упражнений и приведены многочисленные ссылки на новые исследования в быстро развивающихся областях, изучаемых в этом томе. Как всегда, нам очень помогли читатели, обратившие наше внимание на опечатки и другие неточности, и мы надеемся на такую помощь и в будущем.

Лондон,
май, 1968

М. Дж. К.
А. С.

СПИСОК СОКРАЩЕНИЙ

АОЭ	асимптотическая относительная эффективность
ВРД	выбор с равными долями
ДА	дисперсионный анализ
МГД	минимальная граница дисперсии
МД	минимальная дисперсия
МП	максимум правдоподобия
НК	наименьшие квадраты
НКО	наилучшая критическая область
ОП	отношение правдоподобия
ОХ	оперативная характеристика
ПКОВ	последовательный критерий отношения вероятностей
п. ф.	производящая функция
п. ф. м.	производящая функция моментов
п. ф. с.	производящая функция семиинвариантов
п. ф. ф. м.	производящая функция факториальных моментов
РНМ	равномерно наиболее мощный
РНМН	равномерно наиболее мощный несмещенный
СК	сумма квадратов
СНБ	сбалансированные неполные блоки
СрК	средний квадрат
ст. св.	степени свободы
ф. п.	функция плотности
ф. р.	функция распределения
х. ф.	характеристическая функция
ЧСНБ	частично сбалансированные неполные блоки

ГЛАВА 35

ДИСПЕРСИОННЫЙ АНАЛИЗ В ЛИНЕЙНОЙ МОДЕЛИ: КЛАССИФИЦИРОВАННЫЕ ДАННЫЕ

35.1 При исследовании свойств НК-оценки (см. **19.4—19.12** т. 2) в линейной модели $y = X\theta + \varepsilon$ *) как линейной несмещенной МД-оценки было отмечено, что сумма квадратов (СК) наблюдений может быть преобразована к сумме двух неотрицательных компонент

$$y'y \equiv (y - X\hat{\theta})'(y - X\hat{\theta}) + (X\hat{\theta})'(X\hat{\theta}), \tag{35.1}$$

где первое слагаемое представляет собой СК остатков (остаточную СК) от модели, подогнанной по методу НК. Второе слагаемое в правой части (35.1) представляет собой величину уменьшения СК, обусловленного подгонкой модели. Чем больше эта величина (т. е. чем меньше остаточная СК), тем лучше подогнанная модель отражает наблюдаемое соотношение между y и X.

Переписав (35.1) в виде

$$y'y \equiv (y - X\hat{\theta})'(y - X\hat{\theta}) + \hat{\theta}'(X'X)\hat{\theta} \tag{35.2}$$

и вспоминая из (19.16), что

$$V(\hat{\theta}) = \sigma^2 (X'X)^{-1}, \tag{35.3}$$

убеждаемся, что если вектор ошибок ε имеет нормальное распределение, то, согласно **15.4** (т. 1), вектор $\hat{\theta}$, будучи линейной функцией от ε, также имеет многомерное нормальное распределение со средним θ и матрицей рассеяния (35.3). Следовательно, второе слагаемое правой части в (35.2), с точностью до множителя $-(2\sigma^2)^{-1}$, совпадает с показателем экспоненты в формуле для плотности указанного многомерного нормального

*) Напомним читателю, что y и X — это соответственно $(n \times 1)$-вектор-столбец и $(n \times k)$-матрица наблюдений исследуемых переменных y и x_i $(i = 1, 2, \ldots, k)$, θ — это $(k \times 1)$-вектор-столбец неизвестных параметров и ε есть $(n \times 1)$-вектор-столбец случайных «ошибок» с математическим ожиданием $\mathbf{M}(\varepsilon) = 0$ и с матрицей рассеяния $V(\varepsilon) = \mathbf{M}(\varepsilon\varepsilon') = \sigma^2 I$. При этом НК-оценки — это оценки, полученные по методу наименьших квадратов, а МД-оценки — это оценки, имеющие минимальную дисперсию. (*Прим. ред.*)

распределения. Отсюда, согласно **24.6**, $\hat{\theta}'X'X\hat{\theta}/\sigma^2$ имеет нецентральное χ^2-распределение вида (24.18) с $\nu = k$ степенями свободы и параметром нецентральности

$$\lambda = \theta' V^{-1}(\hat{\theta})\,\theta = \theta' X' X \theta/\sigma^2. \tag{35.4}$$

Для краткости, так же как в **24.5**, это распределение будем обозначать $\chi'^2(\nu, \lambda)$.

35.2 Полученный результат позволяет проверить гипотезу о том, что $\theta = \theta_0$, и, в частности, гипотезу

$$H_0\colon\ \theta = 0. \tag{35.5}$$

Проверка последней гипотезы основана на том, что в случае (35.5) λ, определяемое (35.4), равно нулю, и соответствующее распределение становится (центральным) χ^2 с k степенями свободы (ст. св.). Как мы видели в **19.11—12** (т. 2), величина $(y - X\hat{\theta})'(y - X\hat{\theta})/\sigma^2$ имеет χ^2-распределение с $n - k$ ст. св., а если справедливо (35.5), то и $y'y/\sigma^2$ имеет χ^2-распределение с n степенями св. Значит, применима теорема Кокрэна из **15.16** (т. 1) и слагаемые в правой части (35.2) независимы. Тем самым, если справедливо (35.5), то отношение этих слагаемых (поделенных на число ст. св.)

$$F = \{\hat{\theta}'X'X\hat{\theta}/k\}/\{(y - X\hat{\theta})'(y - X\hat{\theta})/(n - k)\} \tag{35.6}$$

имеет распределение дисперсионного отношения (F-распределение) с $(k, n - k)$ ст. св.

Для исследования мощности критерия проверки гипотезы (35.5), основанного на (35.6), мы должны знать соответствующее распределение в том случае, когда (35.5) не выполняется. Чтобы показать, что оно совпадает с нецентральным F (24.105), нам нужно доказать, что числитель и знаменатель в (35.6) остаются независимыми и в случае $\theta \neq 0$. Если же мы хотим проверить более общую гипотезу о том, что $\theta = \theta_0 \neq 0$, то нам нужно знать соответствующее распределение для того, чтобы вообще строить критерий. Таким образом, нам нужно распространить теорему Кокрэна на случай нецентральных нормальных величин, т. е. нормальных величин, у которых, возможно, разные математические ожидания.

35.3 Соотношение (35.2), помимо упомянутого выше применения, открывает и другие возможности. Предположим, что $X'X$ — диагональная матрица, скажем, C, с диагональными элементами c_{ii}. Тогда второе слагаемое в (35.2) можно само представить в виде следующей суммы:

$$(X\hat{\theta})'(X\hat{\theta}) = \sum_{i=1}^{k} c_{ii}\theta_i^2. \tag{35.7}$$

При этом элементы c_{ii} положительны, поскольку $X'X$ положительно определенная, невырожденная матрица. Соотношение (35.7) дает сведение СК к сумме k членов, каждый из которых соответствует своему параметру. При этом можно поставить задачу проверки гипотезы о значении отдельного θ_i, и нас опять интересует распределение компоненты $c_{ii}\hat{\theta}_i^2$ при $\theta \neq 0$.

Если матрица $X'X$ диагональна, то диагональна и (35.3). В этом случае линейная модель называется *ортогональной*, поскольку $\hat{\theta}_i$ некоррелированы, а при нормальном ε просто независимы. Ортогональная модель уже рассматривалась в связи с теорией регрессии в **28.15—20** т. 2, где указывалось на удобство применения ортогональных полиномов. Используя только интуитивное оправдание, в (28.72) определяется (а в примере 28.3 иллюстрируется) процедура вычисления уменьшенного значения СК (см. выше рассуждения о втором слагаемом в правой части (35.1)), соответствующего каждому вновь добавляемому параметру. Далее рассуждения носят более общий характер.

Дисперсионный анализ

35.4 Изложим теперь фундаментальную концепцию, предложенную впервые Р. А. Фишером в 1920 г. Если сумма квадратов $y'y$ может быть представлена в виде суммы неотрицательных компонент, каждая из которых соответствует некоторому подмножеству параметров линейной модели, то будем говорить *о дисперсионном анализе* (ДА) по y. (Было бы более логично говорить об анализе СК, но в силу исторических причин употребляется термин дисперсионный анализ). ДА интерпретируется как метод разделения влияния на наблюдаемые значения y различных подмножеств множества параметров. Важность такого разделения во многих областях исследования делает ДА одним из важнейших методов прикладной статистики.

Разложение нецентральных квадратичных форм

35.5 Сформулируем теперь общую задачу, возникающую в дисперсионном анализе. Пусть y — вектор, образованный p независимыми нормальными величинами,

$$\mathbf{M}(y) = \mu, \quad V(y) = I, \tag{35.8}$$

и пусть

$$y'y = \sum_{i=1}^{k} y'A_i y, \tag{35.9}$$

где A_i имеет ранг r_i. При каких условиях k квадратичных форм $Q_i = y'A_i y$ будут попарно независимы и какими будут их распределения? Поскольку, согласно **24.4** (т. 2), квадратичная

форма $y'y = Q$ имеет $\chi'^2(p, \mu'\mu)$-распределение, то можно ожидать, что компоненты должны иметь распределение того же вида.

Эта задача отличается от изученной в **15.16** (т. 1) только тем, что там выполнялось равенство $\mu = 0$. Но при $\mu = 0$ мы получили, что любое из следующих трех условий:

(а) $\sum_{i=1}^{k} r_i = r$, где r — ранг Q,

(б) каждая матрица A_i идемпотентна, т. е. $A_i = A_i^2$,

(в) $A_iA_j = 0$ при всех $i \neq j$

влечет два остальных. Просмотрев доказательство этого факта в **15.17—19**, убеждаемся, что это доказательство вовсе не зависит от значения μ. Также не зависит от значения μ доказательство (в **15.13**) теоремы Крэйга о том, что Q_i и Q_j независимы тогда и только тогда, когда $A_iA_j = 0$. Отсюда следует, что (в) эквивалентно условию

(в′) формы Q_i попарно независимы.

Однако эквивалентность условия (б) условию

(б⁰) каждая форма Q_i имеет (центральное) χ^2-распределение с r_i ст. св.

опирается на результат **15.11** о том, что если $\mu = 0$, то Q_i имеют χ^2-распределение тогда и только тогда, когда A_i идемпотентны. Таким образом, для обобщения результата **15.16—19** на случай $\mu \neq 0$ остается обобщить утверждение (б⁰) (т. е. результат **15.11**) на случай $\mu \neq 0$.

35.6. По существу, единственное изменение, которое нужно внести в **15.11**, состоит в том, что полученные после канонического преобразования переменные y_i^2 в (15.43) будут, согласно **24.4** (т. 2), иметь распределение $\chi'^2(1, \mu_i^2)$. Производящая функция семиинвариантов (п. ф. с.) формы $\sum_{i=1}^{r} a_i y_i^2$ будет иметь вид не (15.45) (т. 1), а более общий вид, полученный в упражнении 24.1. Отсюда для семиинвариантов формы Q получим, что

$$\varkappa_s = 2^{s-1}(s-1)! \sum_{i=1}^{r} a_i^s (1 + s\mu_i^2) \tag{35.10}$$

(это обобщает (15.46)) и что семиинварианты $\chi'^2(\nu, \lambda)$-распределения суть

$$\varkappa_s^* = 2^{s-1}(s-1)!\,(\nu + s\lambda). \tag{35.11}$$

Для тождественного совпадения (35.10) и (35.11) нужно, чтобы при всех s выполнялись равенства

$$\sum_{i=1}^{r} a_i^s = \nu, \quad \sum_{i=1}^{r} a_i^s \mu_i^2 = \lambda. \tag{35.12}$$

Но (35.12) выполняется тогда и только тогда, когда все $a_i = 1$, так что $\nu = r$ и $\lambda = \sum_{i=1}^{r} \mu_i^2$. Поскольку a_i суть ненулевые характеристические числа матрицы A, то отсюда следует, что A идемпотентна. Таким образом, мы видим, что утверждение, эквивалентное приведенному выше утверждению (б), сводится к

(б′) каждая форма Q_i имеет $\chi'^2(r_i, \lambda_i)$-распределение, что совпадает с (б⁰) при $\mu = 0$. Более того, если совершить обратное ортогональное преобразование от канонических переменных к исходным, то сразу получим, что $\lambda_i = \mu'A_i\mu$ и $\sum_{i=1}^{k} \lambda_i = \mu'\mu$, т. е. сумма параметров нецентральности форм Q_i равна параметру нецентральности формы Q (см. упражнение 24.1 т. 2).

Таким образом, мы приходим к выводу, что если имеет место (35.8—9), то каждое из условий (а), (б), (в) влечет два других. Аналогичное утверждение справедливо для (а), (б′), (в′).

35.7 Если в (35.9) $y'y$ заменить на $Q = y'Ay$, где A — произвольная идемпотентная матрица, имеющая ранг $r < p$, то это не повлияет на результат **35.6**. Соответствующие рассуждения, проведенные в главе 15 для случая $\mu = 0$, справедливы и в случае произвольного μ.

Результат изменится совсем незначительно даже в том случае, если в (35.8) вместо условия $V(y) = I$ потребовать, чтобы компоненты y имели невырожденное многомерное нормальное распределение с дисперсионной матрицей V, которая не обязательно диагональна. Так, если

$$\mathbf{M}(y) = \mu, \quad V(y) = V, \tag{35.13}$$

и

$$Q = y'Ay = \sum_{i=1}^{k} y'A_iy = \sum_{i=1}^{k} Q_i, \tag{35.14}$$

можно положить $V = TT'$, и преобразование $y = Tz$ приводит к независимым нормальным величинам z, поскольку показатель экспоненты в формуле для плотности многомерного нормального распределения сводится к

$$y'V^{-1}y = z'T'(T')^{-1}T^{-1}Tz = z'z.$$

Далее, из (35.14) имеем

$$Q = z'T'ATz = \sum_{i=1}^{k} z'T'A_iTz = \sum_{i=1}^{k} Q_i,$$

т. е. получаются квадратичные формы, с которыми мы уже имели дело. Условие (б) в **35.5** сводится к

$$\boldsymbol{T}'\boldsymbol{A}_i\boldsymbol{T} = \boldsymbol{T}'\boldsymbol{A}_i\boldsymbol{T} \cdot \boldsymbol{T}'\boldsymbol{A}_i\boldsymbol{T}$$

или

$$\boldsymbol{A}_i\boldsymbol{V} = \boldsymbol{A}_i\boldsymbol{V} \cdot \boldsymbol{A}_i\boldsymbol{V},$$

так что $\boldsymbol{A}_i\boldsymbol{V}$, так же как $\boldsymbol{AV}$, должны быть теперь идемпотентными матрицами. Условие (в) сводится к

$$\boldsymbol{T}'\boldsymbol{A}_i\boldsymbol{T} \cdot \boldsymbol{T}'\boldsymbol{A}_j\boldsymbol{T} = \boldsymbol{0}$$

или

$$\boldsymbol{A}_i\boldsymbol{V}\boldsymbol{A}_j = \boldsymbol{0}, \qquad i \neq j.$$

Условие (а) не меняется при ортогональном преобразовании.

Теперь мы можем окончательно сформулировать общий результат.

Пусть $\boldsymbol{y}$ имеет невырожденное многомерное нормальное распределение, моменты которого удовлетворяют условию (35.13), и пусть для квадратичной формы Q имеет место представление (35.14), где матрица $\boldsymbol{AV}$ идемпотентна. Тогда Q имеет $\chi'^2(r, \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu})$-распределение, где r равно рангу матрицы $\boldsymbol{A}$, и любое из приведенных ниже условий влечет два других:

(а) $\sum_{i=1}^{r} r_i = r$;

(б) матрицы $\boldsymbol{A}_i\boldsymbol{V}$ идемпотентны, или, что эквивалентно, Q_i имеют $\chi'^2(r_i, \boldsymbol{\mu}'\boldsymbol{A}_i\boldsymbol{\mu})$-распределение, где r_i — ранг $\boldsymbol{A}_i$;

(в) $\boldsymbol{A}_i\boldsymbol{V}\boldsymbol{A}_j = \boldsymbol{0}$ или, что эквивалентно, Q_i попарно независимы.

Грейбилл и Марсалья (1957) получили несколько более общий результат, чем приведенный здесь. Банерджи (1964) упростил их доказательство.

35.8 Этот общий результат о разложении квадратичных форм нормально распределенных случайных величин дает решение задач, сформулированных в **35.2—3**, которые служили мотивировкой наших исследований. В частности, теперь доказано, что числитель и знаменатель в (35.6) независимы, а значит, их отношение имеет соответствующее нецентральное F-распределение (которое мы, как и в **24.31** т. 2, будем обозначать $F'(\nu_1, \nu_2, \lambda)$). При этом $\nu_1 = k$, $\nu_2 = n - k$, а λ задается соотношением (35.4). Аналогично для любого отдельного $\hat{\theta}_i$ в ортогональной модели из **35.3** отношение

$$F = c_{ii}\hat{\theta}_i^2/\{(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})/(n-k)\} \qquad (35.15)$$

имеет $F'(1, n-k, c_{ii}\theta_i^2/\sigma^2)$-распределение, и теперь можно использовать этот факт для проверки гипотезы о значении θ_i.

В более общей ситуации для проверки гипотезы H_0, налагающей $r \leqslant k$ ограничений, отношение СК, соответствующей H_0, к остаточной СК, умноженное на $(n-k)/r$, имеет $F'(r, n-k, \lambda)$-распределение (см. **24.29—31**). При этом особо следует отметить тот факт, что, согласно **24.29**, параметр нецентральности λ всегда имеет точно такой же вид, как и СК в числителе статистики критерия, только наблюденные значения заменяются на их математические ожидания, а затем все выражение делится на σ^2. Таким образом, λ всегда может быть очень просто получено из числителя статистики. Надо $\hat{\boldsymbol{\theta}}$ заменить на $\boldsymbol{\theta}$ и разделить полученное выражение на σ^2.

Это были примеры критерия ОП в линейной модели, полученные, вообще говоря, из канонической формы модели в **24.25—9**. Обсуждение, следующее за (24.100), указывало на то, что критерии ОП для проверки линейных гипотез, содержащих любое подмножество r из k параметров, основываются на величине уменьшения СК, обусловленной этими r параметрами, поделенной на остаточную СК. Канонический подход главы 24 оказался полезным при выводе оптимальных свойств критерия ОП в **24.36—7**. Для целей настоящего изложения обсуждавшееся эквивалентное разбиение СК является более прямым и информативным.

Напомним читателю, что в **24.32—3** приведены точные и приближенные выражения для функции мощности F-критерия ОП.

ДА для классифицированных наблюдений

35.9 Приведенное в **35.4** определение ДА применимо к любой линейной модели и охватывает применения теории регрессии в **28.12—23**. Однако термин «дисперсионный анализ», вообще говоря, используется в более узком смысле, связанном с возникновением этого метода.

В **35.4** указывалось, что ДА используется для разделения влияний на $\boldsymbol{y}$ отдельных параметров. В экспериментальных исследованиях в качестве параметров часто выступают воздействия некоторых «обработок» над переменной $\boldsymbol{y}$. Так, например, при проведении сельскохозяйственных экспериментов, откуда и возникла соответствующая терминология, $\boldsymbol{y}$ может быть, например, урожаем пшеницы с участка определенного размера, а изучаемая «обработка» состоит в применении на участке некоторого удобрения в течение сельскохозяйственного сезона. Естественно, что в эксперимент должны включаться как обработанные, так и необработанные удобрением участки. Главным

при этом является то, что такого рода эксперимент описывается в терминах общей линейной модели посредством введения индикаторной переменной x, которая равна единице, если воздействие присутствует, и нулю в противном случае.

Легко понять, что любой набор «обработок» можно описать указанным способом. Нужно только определить индикаторную переменную x для каждой участвующей в эксперименте «обработки». Так, если в примере предыдущего абзаца имеется два удобрения, то нужно определить x_1 как индикаторную переменную для первого и x_2 как индикаторную переменную для второго удобрения. При этом для участков, на которых вносятся оба удобрения, следует положить $x_1 = 1$, $x_2 = 1$; для участков, получающих только первое удобрение, положить $x_1 = 1$, $x_2 = 0$; для получающих только второе удобрение, положить $x_1 = 0$, $x_2 = 1$; а для участков, на которые удобрения не вносятся, положить $x_1 = x_2 = 0$. Для этого случая без труда можно провести анализ линейной модели, ибо по определению линейной модели элементами матрицы $\boldsymbol{X}$ могут быть любые заданные константы.

35.10 Характерной особенностью матрицы $\boldsymbol{X}$ в примерах, обсуждавшихся в **35.9**, было то, что ее элементы равнялись либо нулю, либо единице, поскольку они отмечали просто присутствие или отсутствие некоторой компоненты в «обработках». В упомянутом выше более узком смысле термин ДА используется при проведении анализа таких линейных моделей, что это ограничение справедливо для всех элементов матрицы $\boldsymbol{X}$. Небольшое количество других целых положительных значений для элементов матрицы $\boldsymbol{X}$ также возможно и в этом узком смысле ДА. Например, в эксперименте с одним удобрением, обсужденном в начале **35.9**, возможно, что некоторые участки получают двойную, некоторые одинарную дозу удобрения, а некоторые вообще не обрабатываются удобрением. В этом случае следует положить $x = 2, 1$ или 0 соответственно. Анализ этой модели все еще будет называться дисперсионным анализом. Однако недостаток такой формулировки в том, что она подразумевает два варианта воздействия на $\mathbf{M}(y)$ как двойных, так и одинарных доз, а модель линейна по β, результирующему параметру, отражающему зависимость y от x. Избежать этого можно, введя две индикаторные переменные: x_1, определяющую отсутствие или присутствие одинарной дозы, и x_2, определяющую присутствие или отсутствие двойной дозы удобрения. Такая формулировка (как могло бы показаться читателю) не сводится к модели с двумя разными удобрениями из конца **35.9**, поскольку в обсуждаемом случае ни для какого участка не может быть $x_1 = x_2 = 1$. При этом мы идем на некоторую потерю симметрии взамен отказа от линейности при изучении эффекта от внесения разных доз.

35.11 Продемонстрируем теперь, как некоторые важные случаи применения ДА формулируются в терминах линейной модели. При этом окажется, что простая структура элементов матрицы $\boldsymbol{X}$ (равных обычно нулю или единице) приводит к соответствующему упрощению самого анализа. Простейшим является случай, называемый обычно *однофакторным анализом* (классификация по одному признаку), когда наблюдения разбиты на группы в соответствии с возможным различием математических ожиданий.

Пример 35.1. Однофакторный дисперсионный анализ

Пусть выборка независимых наблюдений состоит из k групп, причем в i-й группе n_i $(i = 1, \ldots, k)$ наблюдений и $\sum_{i=1}^{k} n_i = n$. Если группы отличаются только математическими ожиданиями, то это можно записать соотношениями

$$y_{iq} = \theta_i + \varepsilon_{iq}, \qquad i = 1, \ldots, k, \qquad q = 1, \ldots, n_i,$$

которые в терминах общей линейной модели переписываются в виде

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon},$$

где

$$\underset{(n\times 1)}{\boldsymbol{y}} = \begin{bmatrix} y_{11} \\ y_{12} \\ \cdot \\ \cdot \\ \cdot \\ y_{1n_1} \\ y_{21} \\ \cdot \\ \cdot \\ \cdot \\ y_{2n_2} \\ \cdot \\ \cdot \\ \cdot \\ y_{k1} \\ \cdot \\ \cdot \\ \cdot \\ y_{kn_k} \end{bmatrix}, \qquad \underset{(k\times 1)}{\boldsymbol{\theta}} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \cdot \\ \cdot \\ \cdot \\ \theta_k \end{bmatrix},$$

и

$$\underset{(n\times k)}{X} = \begin{array}{c} \left(\begin{array}{ccccc} 1 & & & & \\ \cdot & & & & \\ \cdot & & & & \\ \cdot & & & & \\ 1 & & & & \\ & 1 & & & \\ & \cdot & & & \\ & \cdot & & & \\ & \cdot & & & \\ & 1 & & & \\ & & 1 & & \\ & & \cdot & & \\ & & \cdot & & \\ & & \cdot & & \\ & & 1 & & \\ & & & \ddots & \\ & & & & 1 \\ & & & & \cdot \\ & & & & \cdot \\ & & & & \cdot \\ & & & & 1 \end{array}\right) \end{array} \begin{array}{l} \left.\vphantom{\begin{array}{c}1\\ \cdot\\ \cdot\\ \cdot\\ 1\end{array}}\right\} n_1 \text{ строк} \\ \left.\vphantom{\begin{array}{c}1\\ \cdot\\ \cdot\\ \cdot\\ 1\end{array}}\right\} n_2 \text{ строк} \\ \left.\vphantom{\begin{array}{c}1\\ \cdot\\ \cdot\\ \cdot\\ 1\end{array}}\right\} n_3 \text{ строк} \\ \cdot \\ \cdot \\ \cdot \\ \left.\vphantom{\begin{array}{c}1\\ \cdot\\ \cdot\\ \cdot\\ 1\end{array}}\right\} n_k \text{ строк} \end{array}$$

Здесь равные нулю элементы матрицы X опущены. Далее получим

$$X'X = \begin{pmatrix} n_1 & & & 0 \\ & n_2 & & \\ & & \ddots & \\ 0 & & & n_k \end{pmatrix},$$

так что анализ ортогонален (см. 35.3), какими бы ни были значения n_i (в частности, они могут и не быть равными между собой). Далее

$$X'y = \begin{pmatrix} \sum\limits_q y_{1q} \\ \sum\limits_q y_{2q} \\ \cdot \\ \cdot \\ \cdot \\ \sum\limits_q y_{kq} \end{pmatrix},$$

так что НК-оценка параметра θ есть

$$\hat{\theta} = (X'X)^{-1} X'y = \begin{pmatrix} y_{1\cdot} \\ y_{2\cdot} \\ \cdot \\ \cdot \\ \cdot \\ y_{k\cdot} \end{pmatrix},$$

где $y_{i\cdot} = \sum\limits_{q=1}^{n_i} y_{iq}/n_i$ — среднее арифметическое *) наблюдений в i-й группе. Эта оценка согласуется с интуицией, поскольку наблюдения независимы. Наконец,

$$X\hat{\theta} = \begin{array}{c} \left(\begin{array}{c} y_{1\cdot} \\ \cdot \\ \cdot \\ \cdot \\ y_{1\cdot} \\ y_{2\cdot} \\ \cdot \\ \cdot \\ \cdot \\ y_{2\cdot} \\ \cdot \\ \cdot \\ \cdot \\ y_{k\cdot} \\ \cdot \\ \cdot \\ \cdot \\ y_{k\cdot} \end{array}\right) \end{array} \begin{array}{l} \left.\vphantom{\begin{array}{c}1\\ \cdot\\ \cdot\\ \cdot\\ 1\end{array}}\right\} n_1 \text{ строк} \\ \left.\vphantom{\begin{array}{c}1\\ \cdot\\ \cdot\\ \cdot\\ 1\end{array}}\right\} n_2 \text{ строк} \\ \cdot \\ \cdot \\ \cdot \\ \left.\vphantom{\begin{array}{c}1\\ \cdot\\ \cdot\\ \cdot\\ 1\end{array}}\right\} n_k \text{ строк} \end{array}$$

и величина уменьшения СК, обусловленного «подгонкой» модели, равна

$$S_1 = (X\hat{\theta})'(X\hat{\theta}) = \sum_{i=1}^{k} n_i y_{i\cdot}^2. \tag{35.16}$$

Вычитанием получаем из (35.1), что остаточная СК равна

$$S_R = y'y - S_1 = \sum_{i=1}^{k} \sum_{q=1}^{n_i} y_{iq}^2 - \sum_{i=1}^{k} n_i y_{i\cdot}^2 \equiv \sum_i \sum_q (y_{iq} - y_{i\cdot})^2. \tag{35.17}$$

Для проверки простой гипотезы

$$H_1\colon\ \theta = 0, \tag{35.18}$$

налагающей k ограничений, используем (35.6) и получим, что

$$F = \left(\frac{n-k}{k}\right)\frac{S_1}{S_R} \tag{35.19}$$

*) Следует обратить внимание на то, что использование прн y в качестве индекса точки указывает на усреднение, а не на суммирование, как это делалось для частот и вероятностей в главе 33 (т. 2).

имеет $F'\left(k,\ n-k,\ \sum_i n_i\theta_i^2/\sigma^2\right)$-распределение. Если справедлива гипотеза H_1, то это сводится к (центральному) $F(k,\ n-k)$-распределению.

Однако гипотеза (35.18) не представляет принципиального интереса в большинстве практических ситуаций, когда интересуются просто равенством всех θ_i, вне зависимости от их общего значения. При этом вместо (35.18) проверяется сложная гипотеза

$$H_2:\ \theta_1-\theta_k=\theta_2-\theta_k=\ \ldots\ =\theta_{k-1}-\theta_k=0, \qquad (35.20)$$

налагающая только $k-1$ ограничений. Если справедлива гипотеза (35.20), то все n наблюдений имеют одинаковое распределение с общим математическим ожиданием

$$\theta_*\equiv\sum_{i=1}^{k} n_i\theta_i/n=\begin{bmatrix} n_1/n \\ n_2/n \\ \vdots \\ n_k/n \end{bmatrix}'\theta.$$

Оценкой параметра θ_* по методу наименьших квадратов является просто полное выборочное среднее

$$\hat\theta_*=y_{..}=\frac{1}{n}\sum_{i=1}^{k}\sum_{q=1}^{n_i} y_{iq}=\frac{1}{n}\sum_i n_i y_{i.}.$$

Пусть $\mathbf{1}$ обозначает вектор $(n\times 1)$, состоящий из единиц. Тогда линейную модель на время можно переписать в следующем (сингулярном) виде:

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\theta}+\boldsymbol{\varepsilon}\equiv\mathbf{1}\theta_*+\left[\boldsymbol{X}-\mathbf{1}\begin{bmatrix} n_1/n \\ n_2/n \\ \vdots \\ n_k/n \end{bmatrix}'\right]\boldsymbol{\theta}+\boldsymbol{\varepsilon}.$$

Отсюда видно, что значение параметра θ_* (единственное ограничение) не включено в гипотезы (35.20) и что СК, относящуюся к θ_*, а именно

$$(\mathbf{1}\hat\theta_*)'(\mathbf{1}\hat\theta_*)=ny_{..}^2,$$

нужно вычесть из (35.16) (величины уменьшения СК, обусловленного подгонкой модели), чтобы получить СК, соответствующую

оставшимся $k-1$ ограничениям. Таким образом получаем

$$S_2=S_1-ny_{..}^2\equiv\sum_{i=1}^{k} n_i(y_{i.}-y_{..})^2. \qquad (35.21)$$

Остаточная СК, как и ранее, задается выражением S_R, определенным в (35.17).

Теперь, согласно 35.8, получаем статистику критерия

$$F=\left(\frac{n-k}{k-1}\right)\frac{S_2}{S_R}, \qquad (35.22)$$

которая имеет $F'\left(k-1,\ n-k,\ \sum_{i=1}^{k} n_i(\theta_i-\theta_*)^2/\sigma^2\right)$-распределение, сводящееся к центральному $F(k-1,\ n-k)$, если справедлива гипотеза H_2.

При проведении вычислений S_2 и S_R обычно переписываются в виде

$$\left.\begin{aligned} S_2&=\sum_{i=1}^{k}\frac{\left(\sum_{q=1}^{n_i} y_{iq}\right)^2}{n_i}-\frac{\left(\sum_i\sum_q y_{iq}\right)^2}{n}, \\ S_R&=\sum_i\sum_q y_{iq}^2-\sum_i\frac{\left(\sum_q y_{iq}\right)^2}{n_i}. \end{aligned}\right\} \qquad (35.23)$$

Полученные результаты можно свести в следующую таблицу:

Таблица однофакторного дисперсионного анализа

Компонента дисперсии	Число ст. св.	СК	Средний квадрат (СрК) (равен СК, поделенной на число ст. св.)
Между группами	$k-1$	S_2	$S_2/(k-1)$
Остаточная	$n-k$	S_R	$S_R/(n-k)$
	$n-1$	$S_2+S_R=\boldsymbol{y}'\boldsymbol{y}-ny_{..}^2$	
Генерального среднего	1	$S_1-S_2=ny_{..}^2$	$ny_{..}^2$
Полная	n	$S_1+S_R=\boldsymbol{y}'\boldsymbol{y}$	

(35.24)

Строка «генеральное среднее», вообще говоря, опускается, как не представляющая интереса. Критерий отношения дисперсий, основанный на отношении $n(y_{..}-\theta_*)^2$ к $S_R/(n-k)$ является, конечно, обычным t^2-критерием Стьюдента для среднего, т. е. имеет $F(1, n-k)$-распределение, если справедлива гипотеза H_2. Критерий (35.22) основан на отношении СрК между группами к остаточному СрК, в то время как (35.19) получается сложением строчек «между группами» и «генеральное среднее» и делением полученного СрК на остаточный СрК.

Тождества ДА и их геометрические интерпретации

35.12 В примере 35.1 использовалась общая теория линейной модели, однако окончательный результат можно менее формально получить следующим образом. Тождество

$$\sum_{i=1}^{k}\sum_{q=1}^{n_i}(y_{iq}-y_{..})^2=\sum_{i=1}^{k}\sum_{q=1}^{n_i}(y_{iq}-y_{i.})^2+\sum_{i=1}^{k}n_i(y_{i.}-y_{..})^2 \qquad (35.25)$$

разбивает СК отклонений наблюденных значений от их общего среднего значения на СК отклонений от среднего «внутри групп» и СК отклонений от среднего «между группами», т. е. отклонений средних значений в группах от общего среднего. Если удастся показать, что две суммы в правой части (35.25) независимы и имеют распределение типа χ'^2, то отношение второй суммы к первой является интуитивно приемлемым критерием для проверки равенства групповых математических ожиданий в совокупности. Такой подход, как и ранее, ведет к (35.22), но для того, чтобы оправдать выбор именно этой частной статистики критерия, нужна общая теория. В более сложных ситуациях подход, основанный на использовании алгебраических тождеств типа (35.25), зачастую более прост и быстрее приводит к цели, чем прямое применение теории линейной модели, однако следует быть очень осторожным при разбиении СК. В конце концов правильность подтверждается лишь совпадением с общей теорией.

35.13 Пифагорова форма соотношения (35.25) имеет то преимущество, что она наталкивает на мысль о геометрической интерпретации алгебраического разбиения СК, которое является сущностью дисперсионного анализа. В примере 11.7 мы видели, что более простое тождество, когда все наблюдения объединены в одну группу,

$$\sum_i y_i^2=\sum_i(y_i-\bar{y})^2+n\bar{y}^2 \qquad (35.26)$$

геометрически эквивалентно опусканию перпендикуляра из точки $\boldsymbol{y}=(y_1, \ldots, y_n)$ в n-мерном выборочном пространстве на вектор, имеющий равные углы с осями координат (перпендикуляр пересекается с этим вектором в точке $(\bar{y}, \ldots, \bar{y})$), и применению теоремы Пифагора к полученному прямоугольному треугольнику.

Переписывая (35.26) в более общих обозначениях, которые мы уже использовали в примере 35.1 и в (35.25), получим

$$\sum_{i=1}^{k}\sum_{q=1}^{n_i}y_{iq}^2=\sum_i\sum_q(y_{iq}-y_{..})^2+ny_{..}^2, \qquad (35.27)$$

что, как легко видеть, эквивалентно выделению в «полной» СК из (35.24) строчки, соответствующей генеральному среднему для получения левой части (35.25). Последующее разложение в (35.25) первого члена правой части (35.27) имеет аналогичную геометрическую интерпретацию. Выражение $\sum_i\sum_q(y_{iq}-y_{i.})^2$ представляет собой квадрат расстояния от $\boldsymbol{y}$ до вектора $\boldsymbol{X}\hat{\boldsymbol{\theta}}$, определенного в (35.16), а $\sum_i n_i(y_{i.}-y_{..})^2$ равно квадрату расстояния от вектора $\boldsymbol{X}\hat{\boldsymbol{\theta}}$ до вектора, имеющего равные углы с осями координат. Таким образом, с геометрической точки зрения ДА состоит в разложении вектора, соединяющего $\boldsymbol{y}$ с началом координат, на компоненты, соответствующие рассматриваемой задаче.

35.14 Учитывая тот факт, что в примере 35.1 мы получили ортогональный анализ при любом разбиении на группы, вне зависимости от размеров n_i можно попытаться исследовать более сложные системы классификации. Ниже будет показано, что ортогональность, вообще говоря, не имеет места, если группы состоят из разного числа элементов, но в случае групп одинакового размера ортогональность сохраняется. Сначала детально будет разобран случай двухфакторной классификации (по двум признакам), поскольку уже в этом случае проявляется большинство основных моментов, представляющих интерес и в общей ситуации.

Двухфакторный перекрестный ДА

53.15 Пусть выборка из n наблюдений состоит теперь не из k групп, как это было в примере 35.1, а представлена в виде $(r\times c)$-таблицы с числами наблюдений в клетках

(частотами)

$$\begin{array}{cccc|c} n_{11} & n_{12} & \cdots & n_{1c} & n_{1.} \\ n_{21} & n_{22} & \cdots & n_{2c} & n_{2.} \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ n_{r1} & n_{r2} & \cdots & n_{rc} & n_{r.} \\ \hline n_{.1} & n_{.2} & & n_{.c} & n \end{array} \qquad (35.28)$$

Хотя таблица (35.28) формально тождественна с (33.60), изучаемая задача отличается от задачи главы 33 тем, что здесь известно значение y для каждого наблюдения, в то время как там были известны только частоты, указанные в клетках таблицы. Чтобы подчеркнуть это различие, мы будем говорить о $(r \times c)$-*классификации* в отличие от используемого в главе 33 термина *категоризация*. Как и в главе 33, употребление точки вместо некоторого индекса при n означает *суммирование* по этому индексу, в то время как употребление точки вместо какого-либо индекса при y, как и в примере 35.1, обозначает *усреднение* по этому индексу. Как то, так и другое служит для упрощения дальнейших обозначений. Как заметит читатель, в соответствии с этим соглашением, полную частоту в (35.28) следовало бы обозначать $n_{..}$, однако только в этом единственном случае мы сохраним просто обозначение n для «объема выборки».

Естественно, что rc элементов таблицы (35.28) можно рассматривать как разбиение на $k = rc$ групп при однофакторном анализе (пример 35.1). Однако в случае двухфакторного перекрестного ДА обычно интересуются следующими вопросами:

(1) Имеют ли строчки (с частотами $n_{1.}, \ldots, n_{r.}$) одно и то же математическое ожидание?

(2) Имеют ли столбцы (с частотами $n_{.1}, \ldots, n_{.c}$) одно и тоже математическое ожидание?

(3) Существует ли какое-либо соотношение между математическими ожиданиями строк и столбцов?

Реже интересуются также вопросом

(4) Отличается ли математическое ожидание всего множества наблюдений от некоторого гипотетического значения?

35.16 Обозначим p-е наблюдение из i-ой строчки и j-го столбца через y_{ijp}. Тогда в соответствии с нашим соглашением об

обозначениях

$$\left.\begin{aligned} y_{ij.} &= \sum_{p=1}^{n_{ij}} y_{ijp}/n_{ij}, \\ y_{i..} &= \sum_{j=1}^{c}\sum_{p=1}^{n_{ij}} y_{ijp}/n_{i.} = \sum_{j=1}^{c} n_{ij}y_{ij.}/n_{i.} = \sum_{j=1}^{c} n_{ij}y_{ij.}\Big/\sum_{j=1}^{c} n_{ij}, \\ y_{.j.} &= \sum_{i=1}^{r}\sum_{p=1}^{n_{ij}} y_{ijp}/n_{.j} = \sum_{i=1}^{r} n_{ij}y_{ij.}/n_{.j} = \sum_{i=1}^{r} n_{ij}y_{ij.}\Big/\sum_{i=1}^{r} n_{ij}, \\ y_{...} &= \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{p=1}^{n_{ij}} y_{ijp}/n = \sum_{i=1}^{r} n_{i.}y_{i..}/n = \sum_{j=1}^{c} n_{.j}y_{.j.}/n. \end{aligned}\right\} \qquad (35.29)$$

Чтобы избежать путаницы в обозначениях, введем фиктивные переменные n_{ijp}, тождественно равные единице для всех $p = 1, \ldots, n_{ij}$. Тогда (35.29) сведется к

$$\left.\begin{aligned} y_{ij.} &= \sum_p y_{ijp}\Big/\sum_p n_{ijp}, \\ y_{i..} &= \sum_j\sum_p y_{ijp}\Big/\sum_j\sum_p n_{ijp}, \\ y_{.j.} &= \sum_i\sum_p y_{ijp}\Big/\sum_i\sum_p n_{ijp}, \\ y_{...} &= \sum_i\sum_j\sum_p y_{ijp}\Big/\sum_i\sum_j\sum_p n_{ijp}, \end{aligned}\right\} \qquad (35.30)$$

что очень легко запомнить из-за симметрии числителя и знаменателя. В (35.30) и везде далее, если не оговорено противное, суммирование по i производится от 1 до r, суммирование по j от 1 до c, а по p от 1 до n_{ij}.

35.17 При формулировке линейной модели требуется, чтобы для каждого математического ожидания, соответствующего клетке (i, j) таблицы $r \times c$, был свой параметр μ_{ij}. Однако чтобы ответить на вопросы в **35.15**, мы выразим математические ожидания μ_{ij} в терминах *):

μ_{**} — среднее, общее для всех наблюдений;
μ_{i*} — среднее, общее для всех наблюдений в i-й строке;
μ_{*j} — среднее, общее для всех наблюдений в j-м столбце.

Вместе с математическими ожиданиями μ_{ij}, общими для наблюдений из i-й строки и j-го столбца, мы имеем теперь $1 + r + c + rc$ параметров, из которых только rc могут быть линейно независимыми. Сингулярность, которую мы вводим таким выбором

*) Определяя μ_{**}, μ_{i*} и μ_{*j}, автор имеет в виду соответствующим образом взвешенные средние, см. ниже. (*Прим. ред.*).

параметров, можно легко устранить, используя либо технику пополнения из **19.14—15** т. 2, либо исключением лишних параметров, что мы и делаем ниже.

Если определено μ_{**}, то любые $r-1$ средние μ_{i*} определяют одно оставшееся; аналогично достаточно рассматривать только $c-1$ средние μ_{*j}, так как вместе с μ_{**} они однозначно определяют оставшееся. Если μ_{i*} и μ_{*j} заданы в соответствии с этим замечанием, то легко видеть, что только $(r-1)$ $(c-1)$ из μ_{ij} можно определить независимо (ср. со ст. св. в **33.29**). Таким образом, можно ограничиться $r-1$ параметрами μ_{i*} (опуская, скажем, μ_{r*}), $c-1$ параметрами μ_{*j} (опуская, скажем, μ_{*c}), и $(r-1)$ $(c-1)$ параметрами μ_{ij} (скажем, для $i=1, \ldots, r-1$ $j=1, \ldots, c-1$). Эти параметры вместе с μ_{**} образуют тот набор из rc параметров, который нужен для того, чтобы модель не была сингулярна.

Следует обратить внимание, что мы не определяли параметры μ_{**}, μ_{i*}, μ_{*j}, за исключением заявления о том, что они являются (взвешенными) средними от μ_{ij}.

35.18 Положим теперь

$$\left.\begin{aligned} \theta_{**} &= \mu_{**}, \\ \theta_{i*} &= \mu_{i*} - \mu_{**}, \\ \theta_{*j} &= \mu_{*j} - \mu_{**}, \\ \theta_{ij} &= \mu_{ij} - (\mu_{i*} + \mu_{*j}) + \mu_{**}, \end{aligned}\right\} \qquad (35.31)$$

и перепишем линейную модель в виде

$$y_{ijp} = \theta_{**} + \theta_{i*} + \theta_{*j} + \theta_{ij} + \varepsilon_{ijp} \equiv \mu_{ij} + \varepsilon_{ijp}. \qquad (35.32)$$

По понятным причинам θ_{**} называется *генеральным средним*, а θ_{i*} и θ_{*j} называются соответственно *эффектом i-й строки* и *эффектом j-го столбца*, будучи мерой отклонения в указанных строке или столбце от полного генерального среднего. Если бы отклонение математического ожидания в клетке от генерального среднего в точности равнялось сумме соответствующих эффекта строки и эффекта столбца, мы бы имели

$$\mu_{ij} - \mu_{**} = (\mu_{i*} - \mu_{**}) + (\mu_{*j} - \mu_{**}),$$

что влечет $\theta_{ij} = 0$. В этом случае обычно говорят, что i-я строка и j-й столбец «действуют аддитивно» или «не взаимодействуют». Определенная в (35.31) величина θ_{ij} является мерой отклонения от этой ситуации и называется *взаимодействием* i-й строки и j-го столбца.

35.19 Обсуждаемые в **35.17** $r+c+1$ линейные соотношения между параметрами можно теперь переписать (мы еще вернемся к этому ниже в **35.26—8**) в виде

$$\left.\begin{aligned} 0 &= \sum_{i=1}^{r} n_{i\cdot}\theta_{i*} = \sum_{j=1}^{c} n_{\cdot j}\theta_{*j} \\ &= \sum_{i=1}^{r} n_{ij}\theta_{ij}, \quad j = 1, \ldots, c-1, \\ &= \sum_{j=1}^{r} n_{ij}\theta_{ij}, \quad i = 1, \ldots, r-1, \\ &= \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}\theta_{ij}. \end{aligned}\right\} \qquad (35.33)$$

Если как и в **35.17** определить в (35.31) параметры θ только для $i=1, \ldots, r-1$ и $j=1, \ldots, c-1$, то, используя (35.33), исключенные $r+c+1$ параметры можно выразить через эти оставленные следующим образом:

$$\left.\begin{aligned} \theta_{r*} &= -\sum_{i=1}^{r-1} n_{i\cdot}\theta_{i*}/n_{r\cdot}, \\ \theta_{*c} &= -\sum_{j=1}^{c-1} n_{\cdot j}\theta_{*j}/n_{\cdot c}, \\ \theta_{rj} &= -\sum_{i=1}^{r-1} n_{ij}\theta_{ij}/n_{rj}, \qquad j = 1, \ldots, c-1, \\ \theta_{ic} &= -\sum_{j=1}^{c-1} n_{ij}\theta_{ij}/n_{ic}, \qquad i = 1, \ldots, r-1, \\ \theta_{rc} &= +\sum_{i=1}^{r-1}\sum_{j=1}^{c-1} n_{ij}\theta_{ij}/n_{rc}. \end{aligned}\right\} \qquad (35.34)$$

35.20 Теперь мы можем записать матрицу X линейной модели (35.32). Эта матрица состоит не только из нулей и единиц, поскольку запись исключенных параметров в терминах оставшихся в (35.34) включает отношение чисел n с разными индексами.

Чтобы читателю было легче понять смысл элементов матрицы в (35.35), над ее столбцами проставлены параметры, к которым относятся соответствующие элементы, а перед строчками проставлены частоты тех клеток, к которым относятся соответствующие строки. При этом указаны только ненулевые элементы матрицы X, а фигурирующие в ней векторы **1** состоят из единиц, и число их компонент равно сумме тех клеточных частот (указанных перед строками), на которые графически распространяется вектор **1** в (35.35).

$$
\underset{(n\times rc)}{\mathbf{X}} = \begin{array}{c}
\\ n_{11} \\ n_{12} \\ \vdots \\ n_{1,c-1} \\ n_{1c} \\ n_{21} \\ n_{22} \\ \vdots \\ n_{2,c-1} \\ n_{2c} \\ \vdots \\ n_{r-1,1} \\ n_{r-1,2} \\ \vdots \\ n_{r-1,c-1} \\ n_{r-1,c} \\ n_{r1} \\ n_{r2} \\ \vdots \\ n_{r,c-1} \\ n_{rc}
\end{array}
\left(\begin{array}{c|c|c|

Домножая (35.35) слева на транспонированную матрицу, получим

$$\underset{(rc\times rc)}{\boldsymbol{X}'\boldsymbol{X}} = \begin{array}{c} \\ 1 \\ (r-1) \\ (c-1) \\ (r-1)(c-1) \end{array} \begin{array}{c} \begin{array}{cccc} 1 & (r-1) & (c-1) & (r-1)(c-1) \end{array} \\ \left(\begin{array}{c|c|c|c} n & 0' & 0' & 0' \\ \hline 0 & \boldsymbol{A} & \boldsymbol{D} & 0 \\ \hline 0 & \boldsymbol{D}' & \boldsymbol{B} & 0 \\ \hline 0 & 0 & 0 & \boldsymbol{C} \end{array}\right) \end{array} \tag{35.36}$$

Здесь порядок подматриц указан числами, окаймляющими матрицу (35.36), а подматрицы $\boldsymbol{A}$, $\boldsymbol{B}$ и $\boldsymbol{C}$ являются симметричными матрицами, элементы которых, расположенные над главной диагональю, задаются соотношениями

$$\underset{(r-1)\times(r-1)}{\boldsymbol{A}} = \begin{pmatrix} n_{1\cdot} + \frac{n_{1\cdot}^2}{n_{r\cdot}} & \frac{n_{1\cdot}n_{2\cdot}}{n_{r\cdot}} & \frac{n_{1\cdot}n_{3\cdot}}{n_{r\cdot}} & \cdots & \frac{n_{1\cdot}n_{r-1\cdot}}{n_{r\cdot}} \\ & n_{2\cdot} + \frac{n_{2\cdot}^2}{n_{r\cdot}} & \frac{n_{2\cdot}n_{3\cdot}}{n_{r\cdot}} & \cdots & \frac{n_{2\cdot}n_{r-1\cdot}}{n_{r\cdot}} \\ & & \ddots & & \vdots \\ & & & & n_{r-1\cdot} + \frac{n_{r-1\cdot}^2}{n_{r\cdot}} \end{pmatrix} \tag{35.37}$$

$$\underset{(c-1)\times(c-1)}{\boldsymbol{B}} = \begin{pmatrix} n_{\cdot 1} + \frac{n_{\cdot 1}^2}{n_{\cdot c}} & \frac{n_{\cdot 1}n_{\cdot 2}}{n_{\cdot c}} & \frac{n_{\cdot 1}n_{\cdot 3}}{n_{\cdot c}} & \cdots & \frac{n_{\cdot 1}n_{\cdot c-1}}{n_{\cdot c}} \\ & n_{\cdot 2} + \frac{n_{\cdot 2}^2}{n_{\cdot c}} & \frac{n_{\cdot 2}n_{\cdot 3}}{n_{\cdot c}} & \cdots & \frac{n_{\cdot 2}n_{\cdot c-1}}{n_{\cdot c}} \\ & & \ddots & & \vdots \\ & & & & n_{\cdot c-1} + \frac{n_{\cdot c-1}^2}{n_{\cdot c}} \end{pmatrix} \tag{35.38}$$

Несколько сложнее записывается $(r-1)(c-1)\times(r-1)(c-1)$-матрица $\boldsymbol{C}$. Если снабдить ее строчки и столбцы индексами того параметра θ_{ij}, к которому они относятся (так что, например, 3-я строка и c-й столбец снабжаются индексом (13)-я строка и (21)-й столбец), то элемент из строки, снабженной индексом (kl), и столбца, снабженного индексом (mq), матрицы $\boldsymbol{C}$

(соответствующий параметрам θ_{kl} и θ_{mq}) равен

$$C_{(kl),(mq)} = \begin{cases} n_{kl} + n_{kl}^2\left(\frac{1}{n_{kc}} + \frac{1}{n_{rl}} + \frac{1}{n_{rc}}\right), & \text{если } k=m,\ l=q, \\ n_{kl}n_{kq}\left(\frac{1}{n_{kc}} + \frac{1}{n_{rc}}\right), & \text{если } k=m,\ l\neq q, \\ n_{kl}n_{ml}\left(\frac{1}{n_{rl}} + \frac{1}{n_{rc}}\right), & \text{если } k\neq m,\ l=q, \\ n_{kl}n_{mq}/n_{rc}, & \text{если } k\neq m,\ l\neq q. \end{cases} \tag{35.39}$$

Элементы $(r-1)\times(c-1)$-матрицы $\boldsymbol{D}$ из (35.36) равны

$$D_{ij} = n_{ij}\left(1 - \frac{n_{ic}}{n_{\cdot c}}\frac{n_{\cdot j}}{n_{ij}}\right) + \frac{n_{i\cdot}n_{\cdot j}}{n_{r\cdot}}\left(\frac{n_{rc}}{n_{\cdot c}} - \frac{n_{rj}}{n_{\cdot j}}\right). \tag{35.40}$$

35.21 В общем случае матрицу $\boldsymbol{X}'\boldsymbol{X}$ в (35.36) можно обратить только численно, как и в общем методе НК, но нетрудно заметить, что если $\boldsymbol{D}=0$, то матрица принимает вид

$$\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} n & & & 0 \\ & \boldsymbol{A} & & \\ & & \boldsymbol{B} & \\ 0 & & & \boldsymbol{C} \end{pmatrix}, \tag{35.41}$$

а обратная к ней равна просто

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{pmatrix} n^{-1} & & & 0 \\ & \boldsymbol{A}^{-1} & & \\ & & \boldsymbol{B}^{-1} & \\ 0 & & & \boldsymbol{C}^{-1} \end{pmatrix}. \tag{35.42}$$

Таким образом, мы приходим к необходимости изучить условия, при которых $\boldsymbol{D}=0$, т. е. каждый элемент D_{ij}, определенный в (35.40), равен нулю. Из структуры D_{ij} легко следует, что это выполняется тогда и только тогда, когда

$$\frac{n_{ij}}{n_{\cdot j}} = \frac{n_{ic}}{n_{\cdot c}}, \qquad i=1,\ldots,r-1; \qquad j=1,\ldots,c-1,$$

а также

$$\frac{n_{rj}}{n_{\cdot j}} = \frac{n_{rc}}{n_{\cdot c}}.$$

Эти условия сводятся к тому, что все клеточные частоты в (35.28) n_{ij} пропорциональны своим полным частотам столбцов $n_{\cdot j}$. Отсюда следует, что n_{ij} должны быть пропорциональны также своим полным частотам строк $n_{i\cdot}$, а значит, для того

чтобы $\boldsymbol{D}$ была нулевой матрицей, нужно, чтобы для всех i, j

$$n_{ij} = n_{i.}n_{.j}/n. \tag{35.43}$$

При этих условиях пропорциональности частот *) двухфакторный анализ становится относительно простым.

Случай пропорциональных частот

35.22 Отметим сначала, что из (35.42) следует, что НК-оценка генерального среднего θ_{**} ортогональна оценкам всех других параметров. Аналогично $r-1$ линейно независимых эффектов строк оцениваются ортогонально ко всем другим параметрам так же, как и $c-1$ линейно независимых эффектов столбцов и $(r-1)(c-1)$ линейно независимых взаимодействий. Неортогональность появляется только *внутри* последних трех групп параметров и возникает из-за того, что каждая группа получалась как линейно независимое подмножество группы, состоящей из большого числа изучаемых нами параметров.

Возможно, читатель обратил внимание на то, что мы не указали еще значение самих НК-оценок. Причина в том, что даже если выполняются условия пропорциональности частот (35.43), элементы матрицы $\boldsymbol{C}$ все еще достаточно сложны, чтобы можно было обратить эту матрицу, хотя в каждом конкретном случае можно вычислить значение $\boldsymbol{C}^{-1}$. К счастью, однако, можно использовать ортогональность, сославшись на предыдущий параграф, чтобы сначала получить НК-оценки для эффектов строк и эффектов столбцов, а затем, используя их, получить НК-оценки взаимодействий. Для этого нужно только обратить матрицы $\boldsymbol{A}$ и $\boldsymbol{B}$ из (35.37—8) и, используя (35.42), вычислить первые $1+(r-1)+(c-1)=r+c-1$ строки $(rc\times 1)$-вектора НК-оценок.

35.23 Умножением соответствующих матриц легко проверить, что матрица, обратная (35.37), равна

$$\boldsymbol{A}^{-1} = \begin{bmatrix} \frac{1}{n_{1.}}-\frac{1}{n} & -\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ & \frac{1}{n_{2.}}-\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ & & \ddots & & \vdots \\ & & & \ddots & \vdots \\ & & & & \frac{1}{n_{r-1,.}}-\frac{1}{n} \end{bmatrix}, \tag{35.44}$$

*) Как может заметить читатель, условие пропорциональности (35.43) в точности совпадает с условием, определяющим «независимость» частот в таблице сопряженности (см. **33.4** и **33.29** т. 2). Однако сходство только формально, ибо в главе 33 n_{ij} были случайными величинами, а здесь они — просто наперед заданные константы.

и аналогично для (35.38)

$$\boldsymbol{B}^{-1} = \begin{bmatrix} \frac{1}{n_{.1}}-\frac{1}{n} & -\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ & \frac{1}{n_{.2}}-\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ & & \ddots & & \vdots \\ & & & \ddots & \vdots \\ & & & & \frac{1}{n_{.c-1}}-\frac{1}{n} \end{bmatrix}. \tag{35.45}$$

Теперь нам нужны только первые $r+c-1$ строк $(rc\times 1)$-вектора $\boldsymbol{X'y}$. Из (35.35) получаем в обозначениях (35.29)

$$(\boldsymbol{X'y})_{r+c-1} = \begin{bmatrix} ny_{...} \\ n_{1.}(y_{1} - y_{r..}) \\ n_{2.}(y_{2..} - y_{r..}) \\ \cdots\cdots \\ n_{r-1,.}(y_{r-1,..} - y_{r..}) \\ n_{.1}(y_{.1.} - y_{.c.}) \\ n_{.2}(y_{.2.} - y_{.c.}) \\ \cdots\cdots \\ n_{.c-1}(y_{.c-1,.} - y_{.c.}) \end{bmatrix} \tag{35.46}$$

Используя (35.42) и (35.44—6), получим для первых $r+c-1$ компонент НК-оценки $(\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$

$$\begin{bmatrix} \hat{\theta}_{**} \\ \hat{\theta}_{1*} \\ \vdots \\ \hat{\theta}_{r-1,*} \\ \hat{\theta}_{*1} \\ \vdots \\ \hat{\theta}_{*c-1} \end{bmatrix} = \begin{bmatrix} y_{...} \\ y_{1..} - y_{...} \\ \vdots \\ y_{r-1,..} - y_{...} \\ y_{.1.} - y_{...} \\ \vdots \\ y_{.c-1,.} - y_{...} \end{bmatrix}. \tag{35.47}$$

Таким образом, значение НК-оценки генерального среднего равно просто полному выборочному среднему, а значения НК-оценок эффектов строк (столбцов) равны разности между выборочными средними соответствующих строк (столбцов) и полным выборочным средним. Из (35.47) и первых двух линейных соотношений в (35.34) сразу следует, что то же самое утверждение

справедливо и для исключенных (лишних) эффектов строки и столбца, т. е. что

$$\hat{\theta}_{r*} = y_{r..} - y_{...}, \quad \hat{\theta}_{*c} = y_{.c.} - y_{...}. \tag{35.48}$$

35.24 Подставляя (35.47—8) в определение взаимодействий θ_{ij} в (35.31) и используя то, что θ_{ij} является линейной функцией других величин (см. **19.6** т. 2), получим

$$\hat{\theta}_{ij} = \hat{\mu}_{ij} - y_{i..} - y_{.j.} + y_{...}. \tag{35.49}$$

Далее из крайнего правого соотношения (35.32) очевидно, что соответствующая НК-оценка должна иметь вид

$$\hat{\mu}_{ij} = y_{ij.},$$

и тогда (35.49) переписывается в виде

$$\hat{\theta}_{ij} = y_{ij.} - y_{i..} - y_{.j.} + y_{...}. \tag{35.50}$$

Таким образом, в случае пропорциональных частот все параметры оценены интуитивно «понятными» оценками.

Теперь, когда в нашей модели получены НК-оценки всех параметров, мы можем привести в примере 35.2 критерии проверки различных гипотез, соответствующих вопросам из **35.15**.

Пример 35.2. Двухфакторный перекрестный ДА с пропорциональными частотами

Как следует из вышесказанного, линейную модель (35.32) для двухфакторного перекрестного анализа можно представить в следующем невырожденном виде:

$$\boldsymbol{y} = \boldsymbol{X}\theta + \varepsilon, \tag{35.51}$$

где матрица $\boldsymbol{X}$ определена в (35.35), а координаты вектора-столбца θ проставлены над соответствующими столбцами матрицы $\boldsymbol{X}$ в (35.35).

Рассмотрим сначала первый из вопросов, сформулированных в (35.15). Для ответа на него нужно указать критерий для проверки гипотезы

$$H_1\colon\ \theta_{1*} = \theta_{2*} = \ \ldots\ = \theta_{r-1,*} = 0, \tag{35.52}$$

налагающей $r-1$ ограничений. Аналогично вопрос (2) из (35.15) эквивалентен гипотезе

$$H_2\colon\ \theta_{*1} = \theta_{*2} = \ \ldots\ = \theta_{*c-1} = 0, \tag{35.53}$$

налагающей $c-1$ ограничений, а вопрос (3) из **35.15** — гипотезе

$$H_3\colon\ \theta_{ij} = 0, \qquad i = 1, \ldots, r-1; \qquad j = 1, \ldots, c-1, \tag{35.54}$$

с $(r-1)(c-1)$ ограничениями. Наконец, вопрос (4) из **35.15** эквивалентен гипотезе с одним ограничением

$$H_4\colon\ \theta_{**} = 0. \tag{35.55}$$

Все четыре гипотезы (35.52—5) являются сложными. Для проверки любой из них нужно найти СК, соответствующую этой гипотезе, и использовать общую теорию, которая была развита выше и подытожена в **35.8**. Поскольку четыре гипотезы в совокупности включают все rc параметров, участвующих в модели, и не имеют общих параметров, мы видим, что величину уменьшения СК, обусловленного подгонкой модели, а именно $\theta'\boldsymbol{X}'\boldsymbol{X}\hat{\theta}$, следует разбить на компоненты, соответствующие каждой из четырех гипотез. В рассматриваемом случае это особенно просто, поскольку, как указывалось в **35.22**, все четыре группы параметров оцениваются ортогонально друг другу. Это следует из того, что $\boldsymbol{X}'\boldsymbol{X}$ имеет вид (35.41).

Перепишем теперь НК-оценки (35.47) и (35.50) в виде

$$\hat{\theta} = \begin{bmatrix} y_{...} \\ \cdots\cdots\cdots \\ y_{1..} - y_{...} \\ \vdots \\ y_{r-1,..} - y_{...} \\ \cdots\cdots\cdots \\ y_{.1.} - y_{...} \\ \vdots \\ y_{.c-1,.} - y_{...} \\ \cdots\cdots\cdots \\ y_{11.} - y_{1..} - y_{.1.} + y_{...} \\ \vdots \\ y_{r-1,c-1,.} - y_{r-1,..} - y_{.c-1,.} + y_{...} \end{bmatrix} = \begin{bmatrix} \hat{\theta}_{**} \\ \cdots \\ \hat{\theta}_{i*} \\ \cdots \\ \hat{\theta}_{*j} \\ \cdots \\ \hat{\theta}_{ij} \end{bmatrix}, \tag{35.56}$$

где соответствующие подвекторы вектора $\hat{\theta}$ имеют 1, $r-1$, $c-1$ и $(r-1)(c-1)$ компонент. Из (35.56) и (35.41) получаем разложение

$$\hat{\theta}'\boldsymbol{X}'\boldsymbol{X}\theta = n\hat{\theta}_{**}^2 + \hat{\theta}'_{i*}\boldsymbol{A}\hat{\theta}_{i*} + \hat{\theta}'_{*j}\boldsymbol{B}\hat{\theta}_{*j} + \hat{\theta}'_{ij}\boldsymbol{C}\hat{\theta}_{ij}, \tag{35.57}$$

где $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ являются подматрицами (35.41), определенными в (35.37—9). Первый член в правой части (35.57) представляет собой СК, соответствующую H_4, и мы можем записать его явно в виде

$$S_4 = ny_{...}^2. \tag{35.58}$$

Соотношение (35.37) можно записать в виде

$$\boldsymbol{A} = \boldsymbol{D}_{n_{i.}} + \boldsymbol{n}_r \boldsymbol{n}_r',$$

где $\boldsymbol{D}_{n_{i.}}$ — диагональная матрица $(r-1)\times(r-1)$ с элементами $n_{i.}$, а $\boldsymbol{n}_r$ — вектор $(r-1)\times 1$ с элементами $n_{i.}/n_{r.}^{1/2}$. Таким образом, второй член в правой части (35.57) записывается в виде

$$S_1 = \hat{\theta}_{i*}' \{\boldsymbol{D}_{n_i} + \boldsymbol{n}_r \boldsymbol{n}_r'\} \hat{\theta}_{i*} = \hat{\theta}_{i*}' \boldsymbol{D}_{n_i} \hat{\theta}_{i*} + \hat{\theta}_{i*}' \boldsymbol{n}_r (\hat{\theta}_{i*}' \boldsymbol{n}_r)' =$$

$$= \sum_{i=1}^{r-1} n_{i.} (y_{i..} - y_{...})^2 + \left\{ \sum_{i=1}^{r-1} \frac{n_{i.}}{n_{r.}^{1/2}} (y_{i..} - y_{...}) \right\}^2 =$$

$$= \sum_{i=1}^{r} n_{i.} (y_{i..} - y_{...})^2 \qquad (35.59)$$

и представляет собой СК, соответствующую H_1. Совершенно аналогично, используя вид матрицы $\boldsymbol{B}$ в (35.38), для третьего члена правой части (35.57) получаем

$$S_2 = \sum_{j=1}^{c} n_{.j} (y_{.j.} - y_{...})^2, \qquad (35.60)$$

т. е. СК, соответствующую H_2. Наконец, из (35.39) находим, что последний член в правой части (35.57) равен

$$S_3 = \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} n_{ij} (y_{ij.} - y_{i..} - y_{.j.} + y_{...})^2 +$$

$$+ \frac{1}{n_{rc}} \left\{ \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} n_{ij} (y_{ij.} - y_{i..} - y_{.j.} + y_{...}) \right\}^2 +$$

$$+ \sum_{i=1}^{r-1} \frac{1}{n_{ic}} \left\{ \sum_{j=1}^{c-1} n_{ij} (y_{ij.} - y_{i..} - y_{.j.} + y_{...}) \right\}^2 +$$

$$+ \sum_{j=1}^{c-1} \frac{1}{n_{rj}} \left\{ \sum_{i=1}^{r-1} n_{ij} (y_{ij.} - y_{i..} - y_{.j.} + y_{...}) \right\}^2 =$$

$$= \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} (y_{ij.} - y_{i..} - y_{.j.} + y_{...})^2; \qquad (35.61)$$

при этом мы использовали тот факт, что для оценок взаимодействий справедливы те же самые линейные соотношения (35.34), что и для соответствующих параметров. Величина уменьшения СК, обусловленного подгонкой модели, исчерпывается четырьмя СК, определенными в (35.58—61). Теперь нам остается найти

остаточную СК, которая здесь, как и в общем случае, есть просто разность

$$S_R = \boldsymbol{y}'\boldsymbol{y} - \hat{\theta}'\boldsymbol{X}'\boldsymbol{X}\hat{\theta} = \sum_i \sum_j \sum_p y_{ijp}^2 - (S_1 + S_2 + S_3 + S_4). \qquad (35.62)$$

Для численных расчетов остальные СК записывают обычно в виде

$$\left.\begin{aligned}
S_4 &= \Big(\sum_i \sum_j \sum_p y_{ijp}\Big)^2 / n, \\
S_1 &= \sum_i \Big\{ \Big(\sum_j \sum_p y_{ijp}\Big)^2 / n_{i.} \Big\} - \Big(\sum_i \sum_j \sum_p y_{ijp}\Big)^2 / n, \\
S_2 &= \sum_j \Big\{ \Big(\sum_i \sum_p y_{ijp}\Big)^2 / n_{.j} \Big\} - \Big(\sum_i \sum_j \sum_p y_{ijp}\Big)^2 / n, \\
S_3 &= \sum_i \sum_j \Big\{ \Big(\sum_p y_{ijp}\Big)^2 / n_{ij} \Big\} - \sum_i \Big\{ \Big(\sum_j \sum_p y_{ijp}\Big)^2 / n_{i.} \Big\} - \\
&\quad - \sum_j \Big\{ \Big(\sum_i \sum_p y_{ijp}\Big)^2 / n_{.j} \Big\} + \Big(\sum_i \sum_j \sum_p y_{ijp}\Big)^2 / n.
\end{aligned}\right\} \qquad (35.63)$$

Проверку этих соотношений мы оставляем читателю. Подставляя в (35.62) соотношения (35.63), получим

$$S_R = \sum_i \sum_j \sum_p y_{ijp}^2 - \sum_i \sum_j \Big\{ \Big(\sum_p y_{ijp}\Big)^2 / n_{ij} \Big\}. \qquad (35.64)$$

Теперь, так же как и в примере 35.1, результаты нашего анализа можно свести в таблицу.

Таблица ДА для перекрестной классификации по двум признакам с пропорциональными частотами

Компонента дисперсии	Число ст. св.	СК	СрК	Параметр λ нецентральности
Между строчками	$r-1$	S_1	$S_1/(r-1)$	$\sum_i n_{i.}\theta_{i*}^2/\sigma^2$
Между столбцами	$c-1$	S_2	$S_2/(c-1)$	$\sum_j n_{.j}\theta_{*j}^2/\sigma^2$
Взаимодействий	$(r-1)\times(c-1)$	S_3	$S_3/(r-1)(c-1)$	$\sum_i \sum_j n_{ij}\theta_{ij}^2/\sigma^2$
Остаточная	$n-rc$	S_R	$S_R/(n-rc)$	
	$n-1$			
Генерального среднего	1	S_4	S_4	$n\theta_{**}^2/\sigma^2$
Полная	n	$y'y$		

(35.65)

Из общей теории пункта **35.8** следует, что критерий ОП для любой из гипотез H_1, H_2, H_3, H_4 сводится к вычислению отношения, соответствующего СрК в (35.65) к остаточному СрК и отклонению гипотезы в случае больших значений полученного отношения. Каждое из этих отношений имеет нецентральное F-распределение с числами ст. св., указанными в соответствующих строках второго столбца таблицы, и параметром нецентральности, указанным в последнем столбце таблицы. При этом параметр нецентральности получается с помощью общего правила из **35.8**. (Ср. с примером 35.1, где критерий для генерального среднего представляет собой обычный t^2-критерий Стьюдента).

Если заранее можно предположить, что все взаимодействия равны нулю, то S_3 объединяется с S_R в качестве новой остаточной СК с $n-r-c+1$ ст. св., а для других критериев соответствующий СрК выступает в роли знаменателя. Для проверки «объемлющей» (comprehensive) гипотезы

$$H_0\colon\ \theta = 0 \tag{35.66}$$

о значениях всех параметров (из выполнения H_0 следует, что выполняются H_1, H_2, H_3 и H_4) нужно использовать отношение

$$F = \frac{(S_1+S_2+S_3+S_4)/rc}{S_R/(n-rc)}. \tag{35.67}$$

Это отношение имеет $F'\left\{rc,\ n-rc,\ \sum_i\sum_j n_{ij}(\theta_{**}+\theta_{i*}+\theta_{*j}+\theta_{ij})^2/\sigma^2\right\}$-распределение, при этом в соответствии с общим правилом параметр нецентральности получается из последнего члена в правой части (35.64) подстановкой θ вместо $\hat{\theta}$. Полученный таким образом критерий совпадает с упомянутым в **35.15**, если рассматривать нашу таблицу как разбиение на rc групп при классификации по одному признаку и применить критерий (35.19). Аналогично для проверки одновременного выполнения H_1, H_2 и H_3 (но не H_4) нужно числитель в (35.67) заменить на $(S_1+S_2+S_3)/(rc-1)$. В результате получается критерий, эквивалентный (35.22) для разбиения наблюдений на rc групп.

Случай равного числа наблюдений в клетках

35.25 В ситуации с пропорциональными частотами самым важным является сбалансированный случай, когда все числа наблюдений в клетках n_{ij} равны, скажем, заданному числу m. При этом вычисление соответствующих сумм в формулах (35.63—4) очевидным образом упрощается (см. упражнение 35.1). В этом случае становится проще найти матрицу, обратную к $\boldsymbol{C}$, из (35.39), и рассуждения, приведенные в **35.22—4**, можно провести соответственно более прямо (см. упражнение 35.2).

Помимо этого упрощения, в *сбалансированном* случае имеется еще один новый эффект, который возникает при $m=1$. В этом случае (см. упражнение 35.1) остаточная СК (35.64) тождественно равна нулю, так же как и ее число ст. св., равное $n-rc$. Эта ситуация, очевидно, требует специального изучения, поскольку все критерии, приведенные в примере 35.2, теряют всякий смысл. Нетрудно понять, что необходимость дальнейшего рассмотрения возникнет из-за того, что при $m=1$, $n=rc$ нам нужно оценить rc параметров по rc наблюдениям. Не удивительно, что мы делам это без остаточной дисперсии, ибо мы находимся в той же ситуации, что и при построении полинома степени $q-1$ (содержащего q констант) по q наблюдениям. Таким образом, мы можем оценить все rc параметров даже при $m=1$, однако при этом остаточная СК становится равной нулю.

Преодолеть указанную трудность можно, только согласившись уменьшить число параметров в модели. И в самом деле, мы исключим $(r-1)$ $(c-1)$ параметров взаимодействий и будем оценивать отавшиеся $r+c-1$ параметров. В этом случае вместо (35.64) мы получим новую остаточную СК. На самом деле, как будет показано в примере 35.3, она совпадет с прежней СК взаимодействий, определяемой согласно (35.61).

Мы не будем особо подчеркивать, что эту ограниченную модель (без параметров взаимодействий) нельзя применять для анализа данных, когда заранее известно, что имеются взаимодействия. Поэтому нецелесообразно сознательно ограничиваться одним наблюдением в клетке при перекрестной классификации, если только заранее неизвестно, что столбцы и строчки не взаимодействуют. Однако иногда соображения стоимости или времени вынуждают нас к такого рода ограничениям.

Пример 35.3. Перекрестная классификация по двум признакам для случая одного наблюдения в клетке

Если исключить из модели параметры взаимодействий, то для случая одного наблюдения в клетке получим

$$y_{ij} = \theta_{**} + \theta_{i*} + \theta_{*j} + \varepsilon_{ij},$$

где, как и прежде, чтобы избежать сингулярности, θ_{i*} определены только для $i=1,\ldots,r-1$, а θ_{*j} — для $j=1,\ldots,c-1$. Все рассуждения пунктов **35.17—19**, относящиеся к рассматриваемым нами параметрам, остаются справедливыми. Матрица $\boldsymbol{X}$, определенная в (35.35), остается той же, если ограничиться ее первыми $r+c-1$ столбцами. То же самое справедливо для главной $(r+c-1)\times(r+c-1)$-подматрицы матрицы $\boldsymbol{X}'\boldsymbol{X}$ в (35.36), причем, поскольку условия пропорциональности (35.43) выполняются, то остается справедливым равенство

$\boldsymbol{D} = 0$. Матрицы $\boldsymbol{A}$ и $\boldsymbol{B}$ из (35.37—8) и обратные к ним из (35.44—5) не изменяются. То же справедливо и для векторов (35.46—7), причем последние дают полные НК-оценки наших параметров, а не являются подвекторами соответствующих векторов оценок. Следовательно, мы можем проверить гипотезы H_1, H_2 и H_4 из (35.52—3), (35.55) точно так же, как и в примере 35.2, только S_3, которая раньше являлась СК взаимодействий, становится теперь новой остаточной СК, поскольку сумма четырех СК в приводимой ниже сокращенной таблице должна, как всегда, равняться $\boldsymbol{y}'\boldsymbol{y}$.

Таблица ДА для перекрестной классификации по двум признакам с одним наблюдением в клетке

Компонента дисперсии	Число ст. св.	СК
Между строчками	$r-1$	S_1
Между столбцами	$c-1$	S_2
Остаточная	$(r-1)(c-1)$	S_3
	$rc-1$	
Генерального среднего	1	S_4
Полная	n	$y'y$

(35.68)

Критерии для проверки гипотез H_1, H_2 и H_4 могут теперь быть получены заменой знаменателя в соответствующей F статистике на СрК $S_3/(r-1)(c-1)$.

Хотя мы исключили параметры взаимодействий θ_{ij} из модели, чтобы получить остаточную СК, тем не менее можно использовать их оценки $\hat{\theta}_{ij}$ для проверки равенства взаимодействий нулю, выделив из полученной остаточной СК соответствующую компоненту.

Рассмотрим линейную форму

$$L = \sum_i \sum_j c_{ij}\hat{\theta}_{ij},$$

где $\hat{\theta}_{ij}$ определены в (35.50) (последний индекс у игреков теперь излишен), а c_{ij} — коэффициенты, которые будут определены позже. Если все взаимодействия θ_{ij} равны нулю, то $\mathbf{M}(L) = 0$. В противном случае это, вообще говоря, неверно. Поэтому для проверки гипотезы о нулевых взаимодействиях интуитивно приемлемо использовать статистику L. Если выбрать c_{ij} так, чтобы $\sum_i c_{ij} = \sum_j c_{ij} = 0$, то из (35.50) мы видим, что $L \equiv \sum_i \sum_j c_{ij}y_{ij}$,

и, следовательно,

$$\mathbf{D}(L) = C^2\sigma^2,$$

где $C^2 = \sum_i \sum_j c_{ij}^2$, а σ^2, как обычно, дисперсия ошибок. Таким образом, если взаимодействия равны нулю, $L^2/(C^2\sigma^2)$ имеет χ^2-распределение с одной ст. св. Более того, рассматриваемая теперь остаточная СК в соответствии с (35.61) есть просто $S_3 = \sum_i \sum_j \hat{\theta}_{ij}^2$. Далее $S_3 - \{L^2/(C^2\sigma^2)\}$ не зависит от $L^2/C^2\sigma^2$, поскольку $\hat{\theta}_{ij}$ можно ортогональным преобразованием перевести в новое множество независимых стандартных нормальных величин, одной из которых является $L/C\sigma$, а $S_3 - \{L^2/(C^2\sigma^2)\}$, будучи суммой квадратов остальных величин, имеет χ^2-распределение с $(r-1)(c-1)-1 = rc-r-c$ ст. св.

Остается выбрать c_{ij}. Заметим, что они могут быть выбраны как функции от $\hat{\theta}_{i*}$ и $\hat{\theta}_{*j}$, поскольку в соответствии с (35.42) последние не зависят от $\hat{\theta}_{ij}$. Поэтому любое условное распределение L при фиксированных $\hat{\theta}_{i*}$, $\hat{\theta}_{*j}$ будет таким же, как и маргинальное распределение L, т. е. совпадать с описанным выше.

Простейшим является выбор $c_{ij} = \hat{\theta}_{i*}\hat{\theta}_{*j}$, так что мы можем положить

$$S_I = \left(\sum_i \sum_j \hat{\theta}_{i*}\hat{\theta}_{*j}y_{ij}\right)^2 \Big/ \left(\sum_i \hat{\theta}_{i*}^2 \sum_j \hat{\theta}_{*j}^2\right) = \frac{\left\{\sum_i \sum_j (y_{i.}-y_{..})(y_{.j}-y_{..})y_{ij}\right\}^2}{\sum_i (y_{i.}-y_{..})^2 \sum_j (y_{.j}-y_{..})^2}. \qquad (35.69)$$

При этом S_I/σ^2 имеет χ^2-распределение с одной ст. св., а $(S_3 - S_I)/\sigma^2$ не зависит от S_I/σ^2 и имеет χ^2-распределение с $rc-r-c$ ст. св. Их отношение $S_I/(S_3-S_I) = F$ имеет распределение дисперсионного отношения с $(1, rc-r-c)$ ст. св. и может быть использовано для порверки гипотезы о том, что все взаимодействия равны нулю. Этот критерий полной *аддитивности* эффектов был предложен Тьюки (1949). (О дальнейших обобщениях см. Шеффе (1959).) М. Гхош и Шарма (1963) изучили его мощность против альтернативной гипотезы о том, что взаимодействия имеют вид $\theta_{ij} = \alpha\theta_{i*}\theta_{*j}$. Для классификации 6×6 было найдено, что мощность имеет тот же самый порядок, что и у F-критерия для взаимодействий, полученного приравниванием смежных пар из θ_{i*} и θ_{*j}.

Выбор весов

35.26 Теперь мы должны обсудить момент, который был умышленно обойден молчанием при формулировке нашей линейной модели в **35.17—19**. Там отмечалось, что в исходной модели $r+c+1$ параметров являются лишними в том смысле, что они линейно зависят от остальных параметров. Мы избавились от лишних параметров, используя набор линейных соотношений, приведенных в (35.33). Из этих соотношений следовало (35.34), что в свою очередь определяло структуру основной матрицы (35.35). Теперь необходимо осознать, что набор соотношений, приведенный в (35.33), по существу, произволен. Например, в первом из приведенных там соотношений мы *положим* равным нулю конкретную линейную функцию $\sum_i n_i\theta_{i*}$, используя в качестве весов маргинальные частоты строк $n_{i.}$. Это может казаться естественным, но никоим образом не является необходимым. Мы могли бы вместо этого взять равные веса так, чтобы $\sum_i \theta_{i*}=0$, или даже произвольные веса w_i так, чтобы $\sum_i w_i\theta_{i*}=0$.

Если бы все множество из n наблюдений являлось простой случайной выборкой из некоторой совокупности, то наблюдаемые значения $n_{i.}/n$ были бы оценками относительных частот соответствующих строк в совокупности. В этом случае имело бы смысл определять эффекты строк, используя эти веса для выражения линейной зависимости. Аналогично $\sum_j n_{.j}\theta_{*j}=0$ и другие соотношения из (35.33) имели бы смысл в том же контексте. В связи с этим мы назовем соответствующие веса *частотными весами* (*frequency weights*).

35.27 Однако во многих экспериментальных ситуациях нет смысла рассматривать наблюдения как случайную выборку из некоторой совокупности. Так бывает, например, когда $r\times c$ перекрестная классификация проводится умышленно для подробного рассмотрения изучаемой переменной y. В этом случае становится трудно интерпретировать использование наблюденных частот в качестве весов. Может даже оказаться бессмысленным рассматривать какие-либо веса как единственно правильные в том смысле, что именно они отражают распределение изучаемой совокупности. Например, если имеется 2×3 перекрестная классификация для изучения влияния двух различных доз удобрения A и трех различных доз удобрения B на урожай (хлеба), то ее имеет смысл рассматривать просто как изучение влияния соответствующих эффектов и взаимодействий, а вовсе не как представление какой-либо совокупности. В этом состоит существенная разница между «экспериментальным» и «выборочным» подходом к имеющимся данным. Мы вернемся к последнему в главах 38—9.

Таким образом, в экспериментальных исследованиях обычно (если отсутствует какая-либо известная соответствующая система весов), вместо (35.33) используют равные веса. В оставшейся части главы под системой *равных весов* мы будем понимать выполнение соотношений (35.33), но только со всеми символами $n_{i.}$, $n_{.j}$, n_{ij}, замененными на единицы. Следует обратить внимание, что хотя в сбалансированном случае (см. **35.25**) по существу уже использовалась система равных весов, поскольку все частоты были равны, тем не менее общие результаты, относящиеся к случаю пропорциональных частот (в **35.22—4** и упражнении 35.2), не останутся верны, если только не использовать частотные веса.

Теперь мы собираемся вернуться к обсуждению общего случая непропорциональных частот, от рассмотрения которого мы отказались в **35.21**. Различия между этими двумя системами весов выступают здесь особенно резко хотя бы потому, что в этом самом общем случае может оказаться малое число очень больших частот, которые являются преобладающими в системе частотных весов и могут затмить истинную картину.

35.28 Выбор весов в (35.33) в общем влияет на оценки всех параметров: генерального среднего, эффектов строк и столбцов, взаимодействий. Однако, как показал Шеффе (1959), если все истинные значения взаимодействий равны нулю при какой-либо системе весов, то они равны нулю и при любой другой системе. В самом деле, если $\theta_{ij}^{(1)}=0$ для одной из систем весов, то, согласно (35.31), это эквивалентно тому, что

$$\mu_{ij}=\mu_{i*}^{(1)}+\mu_{*j}^{(1)}-\mu_{**}^{(1)}.$$

Тогда, согласно (35.31), для любой другой системы весов взаимодействия равны

$$\begin{aligned}\theta_{ij}^{(2)}&=\mu_{ij}-\mu_{i*}^{(2)}-\mu_{*j}^{(2)}+\mu_{**}^{(2)}=\\&=(\mu_{i*}^{(1)}+\mu_{*j}^{(1)}-\mu_{**}^{(1)})-(\mu_{i*}^{(2)}+\mu_{*j}^{(2)}-\mu_{**}^{(2)})=\\&=(\mu_{i*}^{(1)}-\mu_{i*}^{(2)})+(\mu_{*j}^{(1)}-\mu_{*j}^{(2)})-(\mu_{**}^{(1)}-\mu_{**}^{(2)}),\end{aligned}$$

т. е. имеют вид

$$\theta_{ij}^{(2)}=a_i+b_j+c.$$

Но из определения взаимодействий в **35.18** легко понять, что если бы они могли быть представлены в виде суммы компонент, относящихся к строке, к столбцу, и константы, то эти компоненты следовало бы включить в эффект строки, эффект столбца

и генеральное среднее соответственно и положить взаимодействия равными нулю. Таким образом, мы имеем $\theta_{ij}^{(2)}=0$ для всех i и j. Следовательно, если гипотеза H_3 из (35.54) справедлива, то она справедлива при любой системе весов.

Непропорциональные частоты

35.29 Рассмотрим сначала, как и раньше, случай частотных весов (35.33). Условие пропорциональности (35.43) не выполняется, поэтому не все элементы матрицы $\boldsymbol{D}$ из (35.36) обращаются в нуль и упрощенный анализ пунктов **35.22—5** больше неприменим. Тем не менее даже в этом самом общем случае остается верным соотношение (35.36), которое записывается в виде

$$\boldsymbol{X}'\boldsymbol{X}=\begin{pmatrix} n & & \\ & \begin{pmatrix} \boldsymbol{A} & \boldsymbol{D} \\ \boldsymbol{D}' & \boldsymbol{B} \end{pmatrix} & \\ & & \boldsymbol{C} \end{pmatrix}.$$

При этом (35.41) получалось отсюда как частный случай при $\boldsymbol{D}=0$. Как легко проверить простым умножением,

$$\boldsymbol{E}=\begin{pmatrix} \boldsymbol{A} & \boldsymbol{D} \\ \boldsymbol{D}' & \boldsymbol{B} \end{pmatrix}^{-1}=\begin{pmatrix} (\boldsymbol{A}-\boldsymbol{D}\boldsymbol{B}^{-1}\boldsymbol{D}')^{-1} & -(\boldsymbol{A}-\boldsymbol{D}\boldsymbol{B}^{-1}\boldsymbol{D}')^{-1}\boldsymbol{D}\boldsymbol{B}^{-1} \\ -(\boldsymbol{B}-\boldsymbol{D}'\boldsymbol{A}^{-1}\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{A}^{-1} & (\boldsymbol{B}-\boldsymbol{D}'\boldsymbol{A}^{-1}\boldsymbol{D})^{-1} \end{pmatrix}, \tag{35.70}$$

так что матрицу, обратную к $\boldsymbol{X}'\boldsymbol{X}$, можно теперь представить в виде

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}=\begin{pmatrix} n^{-1} & & \\ & \boldsymbol{E} & \\ & & \boldsymbol{C}^{-1} \end{pmatrix}, \tag{35.71}$$

В результате того, что матрица $\boldsymbol{D}$ не обращается в нуль, изменяются оценки вектора параметров $\boldsymbol{\theta}$, полученные методом НК, поскольку, хотя подвектор из (35.46) не изменяется, главная диагональная подматрица $(r+c-1)\times(r+c-1)$ из (35.42) заменяется теперь на соответствующую подматрицу из (35.71). Если (35.46) кратко записать в виде

$$\begin{pmatrix} ny\ldots \\ \boldsymbol{v}_{r-1} \\ \boldsymbol{v}_{c-1} \end{pmatrix},$$

где векторы $\boldsymbol{v}$ являются подвекторами (35.46), а индекс указывает на их размерность, то, используя (35.70—1), получим

следующее обобщение соотношения (35.47):

$$((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y})_{r+c-1}=\begin{pmatrix} n^{-1} & 0 \\ 0 & \boldsymbol{E} \end{pmatrix}\begin{pmatrix} ny\ldots \\ \boldsymbol{v}_{r-1} \\ \boldsymbol{v}_{c-1} \end{pmatrix}=$$

$$=\begin{pmatrix} y\ldots \\ (\boldsymbol{A}-\boldsymbol{D}\boldsymbol{B}^{-1}\boldsymbol{D}')^{-1}(\boldsymbol{v}_{r-1}-\boldsymbol{D}\boldsymbol{B}^{-1}\boldsymbol{v}_{c-1}) \\ (\boldsymbol{B}-\boldsymbol{D}'\boldsymbol{A}^{-1}\boldsymbol{D})^{-1}(\boldsymbol{v}_{c-1}-\boldsymbol{D}'\boldsymbol{A}^{-1}\boldsymbol{v}_{r-1}) \end{pmatrix}. \tag{35.72}$$

Отсюда, проводя соответствующие вычисления, получаем значения оценок $\hat{\theta}_{i*}$, $\hat{\theta}_{*j}$, в то время как по интуитивно понятным соображениям для $\hat{\theta}_{**}$ всегда имеет место равенство $\hat{\theta}_{**}=y\ldots$. Так же как и в **35.24**, из определения взаимодействий в (35.31) следует, что значения их оценок удовлетворяют равенствам

$$\hat{\theta}_{ij}=y_{ij\cdot}-\hat{\theta}_{i*}-\hat{\theta}_{*j}-y\ldots. \tag{35.73}$$

Таким образом, определены НК-оценки всех параметров. Обобщением разложения (35.57) является

$$\hat{\boldsymbol{\theta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\theta}}=n\hat{\theta}_{**}^2+\begin{pmatrix} \hat{\theta}_{i*} \\ \hat{\theta}_{*j} \end{pmatrix}'\begin{pmatrix} \boldsymbol{A} & \boldsymbol{D} \\ \boldsymbol{D}' & \boldsymbol{B} \end{pmatrix}\begin{pmatrix} \hat{\theta}_{i*} \\ \hat{\theta}_{*j} \end{pmatrix}+\hat{\boldsymbol{\theta}}_{ij}'\boldsymbol{C}\hat{\boldsymbol{\theta}}_{ij}. \tag{35.74}$$

Пример 35.4. Двухфакторная перекрестная классификация с непропорциональными частотами и частотными весами

Из (35.74) следует, что каждая из гипотез H_3 и H_4 из (35.54—5) проверяется точно так же, как и в примере 35.2, хотя СК, соответствующую взаимодействиям, нужно находить теперь, проводя вычисления в соответствии с последним членом в правой части (35.74). СрК для генерального среднего и для взаимодействий (имеющие соответственно 1 и $(r-1)(c-1)$ ст. св.) нужно делить на остаточный СрК и вне зависимости от эффектов строк и столбцов получатся величины с нецентральными F-распределениями *).

Средний член в правой части (35.74) (обозначим его M) представляет собой СК, соответствующую объединению эффектов строк и столбцов. Таким образом, проверка *одновременного*

*) Следует обратить особое внимание на то, что если бы в (35.33) вместо частотных использовались бы равные веса, то СК, соответствующая взаимодействиям, не была бы отдельной компонентой в (35.74), а была бы связана с СК эффектов столбцов и строк так же, как последние связаны между собой. Тем не менее в силу результатов **35.28**, если H_3 верна для частотных весов, то она выполнялась бы и для любой другой системы весов.

выполнения H_1 и H_2 из (35.52—3) сводится к вычислению M. Однако на практике обычно интересуются отдельной проверкой гипотез об эффектах строк и столбцов. Так СК, соответствующая, например, эффектам строк, получается вычислением уменьшения остаточной СК, получающейся в результате предварительной оценки всех параметров, за исключением θ_{i*}, а затем оценки всех параметров, включая θ_{i*}. В результате получается СК, соответствующая θ_{i*}. Аналогично можно вычислить СК для θ_{*j}. Поскольку в общем случае эффекты строк и столбцов не ортогональны, эти две СК в сумме не будут равняться M.

Если можно предположить, что все взаимодействия равны нулю, то ситуация упрощается (см. упражнение 35.4).

35.30 Использование вместо (35.33) системы равных весов при проверке эффектов строк и столбцов делает эти вычисления значительно проще, чем для частотных весов, рассмотренных в примере 35.4. Можно поступить следующим образом.

Проведем сначала анализ $(r \times c)$-перекрестной классификации, как если бы это была однофакторная классификация с $k = rc$. Мы получим ДА с $n - rc$ ст. св. для остаточной СК. Оставшиеся rc ст. св. соответствуют суммарному влиянию классификации по строкам и столбцам. Используя пример 35.1, получим следующую таблицу ДА:

ДА для произвольной $(r \times c)$-перекрестной классификации

Компонента дисперсии	Число ст. св.	СК
Обусловленная классификацией в целом	rc	$\sum_i \sum_j n_{ij} y_{ij}^2$.
Остаточная	$n - rc$	$\sum_i \sum_j \sum_p (y_{ijp} - y_{ij\cdot})^2$
Полная	n	$\sum_i \sum_j \sum_p y_{ijp}^2$

(35.75)

Очевидно, что любая клетка, для которой $n_{ij} = 1$, не может внести вклад в остаточный СрК (ср. 35.25 и пример 35.3, где для *всех* клеток $n_{ij} = 1$). Чтобы расщепить rc ст. св. соответственно разбиению общей СК на компоненты, относящиеся к эффектам строк, столбцов, взаимодействиям и генеральному среднему, остается проанализировать средние значения в клетках $y_{ij\cdot}$.

35.31 Из модели (35.32) следует, что $y_{ij\cdot}$ удовлетворяют

$$y_{ij\cdot} = \theta_{**} + \theta_{i*} + \theta_{*j} + \theta_{ij} + \varepsilon_{ij}, \tag{35.76}$$

где ошибки ε_{ij} не коррелированы и имеют нулевые математические ожидания и дисперсии σ^2/n_{ij}. Усредним теперь $y_{ij\cdot}$ по столбцу, т. е. возьмем *невзвешенное* среднее $\frac{1}{c}\sum_{j=1}^{c} y_{ij\cdot} = \bar{y}_i$. Из (35.76) следует, что

$$\bar{y}_i = \theta_{**} + \theta_{i*} + \frac{1}{c}\sum_j \theta_{*j} + \frac{1}{c}\sum_{i,j} \theta_{ij} + \varepsilon_i, \tag{35.77}$$

где ε_i — не коррелированы и имеют нулевые математические ожидания и дисперсии $(\sigma^2/c^2)\sum_j n_{ij}^{-1} = \sigma^2 V_i$. Если вместо частотных весов в (35.33) использовать равные веса, то обе суммы, фигурирующие в правой части (35.77), обратятся в нуль, и мы будем иметь

$$\bar{y}_i = \theta_{**} + \theta_{i*} + \varepsilon_i. \tag{35.78}$$

Но (35.78) можно рассматривать как модель однофакторной классификации (пример 35.1) с одним наблюдением в каждой группе, с той разницей, что дисперсии ошибок различны. Если определить $z_i = \bar{y}_i/V_i^{1/2}$, то получим $\mathbf{M}(z_i) = (\theta_{**} + \theta_{i*})/V_i^{1/2}$, и условия примера 35.1 будут теперь выполнены. Влияние этой замены на анализ сводится к замене определенной в (35.21) величины S_2, которая в рассматриваемой нами ситуации равнялась бы $\sum_i \left(\bar{y}_i - \frac{1}{r}\sum_i \bar{y}_i\right)^2$, на ту же самую сумму, но с коэффициентом V_i^{-1} при каждом члене, т. е.

$$S_2 = \sum_{i=1}^{r} \left(\sum_{j=1}^{c} n_{ij}^{-1}/c^2\right)^{-1} (\bar{y}_i - \bar{y})^2,$$

где

$$\bar{y} = \sum_{i=1}^{r} \left(\sum_{j=1}^{c} n_{ij}^{-1}/c^2\right)^{-1} \bar{y}_i \Big/ \sum_{i=1}^{r} \left(\sum_{j=1}^{c} n_{ij}^{-1}/c^2\right)^{-1}$$

— взвешенное среднее величин $\bar{y}_i$ с весами V_i^{-1}.

Таким образом, мы получаем следующую таблицу:

ДА для $(r \times c)$-перекрестной классификации с равными весами

Компонента дисперсии	Число ст. св.	СК
Обусловленная строчками	$r-1$	$\sum_{i=1}^{r} V_i^{-1}(\bar{y}_i - \bar{y})^2$
Обусловленная оставшейся классификацией	$\frac{r(c-1)+1}{rc}$	(получается как разность)
Остаточная	$n-rc$	(совпадает с (35.75))
Полная	n	$\sum_i \sum_j \sum_p y_{ijp}^2$

(35.79)

Совершенно аналогично можно провести разбиение СК, относящейся к классификации, для классификации по столбцам. Таким образом, получаются критерии для эффектов строк и столбцов в общем случае. Правда, в силу неортогональности, СК, соответствующие строкам и столбцам, нельзя складывать, для того чтобы затем вычитанием получить СК взаимодействий. Однако если r или c равны двойке, то указанным методом нетрудно получить и критерий для взаимодействий (см. упражнение 35.5).

35.32 Система равных весов, применение которой позволило провести ДА (35.79), естественно употребляется в описанных ситуациях, когда n наблюдений сводятся к rc средним, а затем эти средние анализируются, как если бы они сами были наблюдениями. Ничто не мешает провести полный анализ, используя систему равных весов вместо весов из **35.29**. На самом деле часто так и поступают. При этом полученные результаты, вообще говоря, отличаются от результатов **35.29**.

35.33 Изложенный в **35.31** метод восходит к Иэйтсу (1934), который назвал его *методом взвешенных квадратов от средних значений*. Так же как Снедекор (1946) и Р. Андерсон и Банкрофт (1952), он рассматривал также и другие, более приближенные методы анализа. У всех упомянутых авторов приведены численные примеры (см. упражнение 35.7). Дальнейшее теоретическое исследование случая непропорциональных частот содержится в книге Шеффе (1959). В частности, там допускаются произвольные системы весов в (35.33).

Пустые клетки при двухфакторной перекрестной классификации

35.34 При исследовании двухфакторной перекрестной классификации в **35.15—33** мы неявно предполагали, что все клетки в таблице (35.28) содержат по крайней мере одно наблюдение, т. е. $n_{ij} > 0$. На практике это предположение может часто не выполняться в результате случайности, неудачного эксперимента или по другим причинам. Теперь нам нужно рассмотреть, какое влияние оказывает на анализ данных наличие пустых клеток в классификации. Если имеется хотя бы одна пустая клетка, то перекрестная классификация называется *неполной*. До сих пор мы рассматривали только полные классификации.

Очевидно, в общем случае, когда соответствующее взаимодействие θ_{ij} не равно нулю, мы не можем оценить среднее значение μ_{ij}, поскольку никакой информации о значении θ_{ij} из других клеток таблицы мы получить не можем. Из определения (35.31) и линейных соотношений (35.33) следует, что в общем случае, когда при перекрестной классификации имеется хотя бы одна пустая клетка, нельзя оценить ни одну из величин θ_{ij}, θ_{i*}, θ_{*j}, или θ_{**}. Однако даже в этом случае очень просто оценить дисперсию ошибки. Если обозначить $[rc]$ число клеток, содержащих наблюдения, то мы получим следующую таблицу, обобщающую (35.75):

ДА для произвольной $(r \times c)$-перекрестной классификации *)

Компонента дисперсии	Число ст. св.	СК
Обусловленная классификацией	$[rc]$	$\sum_i \sum_j n_{ij} y_{ij.}^2$
Остаточная	$n-[rc]$	$\sum_i \sum_j \sum_p (y_{ijp} - y_{ij.})^2$
Полная	n	$\sum_i \sum_j \sum_p y_{ijp}^2$

(35.80)

35.35 Если в пустых клетках θ_{ij} равно нулю, то упомянутая в **35.34** трудность исчезает. Если мы хотим проверить H_3 из (35.54) (гипотезу о том, что *все* взаимодействия равны нулю), то мы можем поступить точно так же, как и в примере 35.4, оценив

*) В (35.80) суммирование распространяется только на $[rc]$ непустых клеток.

оставшиеся параметры, рассчитав соответствующую им СК и таким образом получить остаточную СК. Разность между последней и остаточной СК из (35.80) будет соответствовать взаимодействиям и иметь $[rc]-r-c+1$ ст. св. Аналогично, если можно *постулировать* выполнение H_3, то эффекты строк и столбцов можно проверять точно так, как и в случае полной классификации, методами, изложенными в упражнении 35.4. Шеффе (1959) дает более детальное изложение.

Здесь мы не будем больше обсуждать этот вопрос. Ниже, в пунктах **37.50—6**, приводятся общие методы анализа линейной модели, когда есть пропуски наблюдений.

Иерархические классификации

35.36 Двухфакторная перекрестная классификация, которую мы рассматривали в **35.15—35**, не является единственным интересным обобщением однофакторной классификации из примера 35.1. Предположим, что внутри каждой из рассмотренных там k групп снова имеется однофакторная классификация наблюдений. Пусть n_1 наблюдений из первой группы разбиты на l_1 подгрупп с числами наблюдений (частотами) $n_{11}, \ldots, n_{1l_1}$, где $\sum_{h=1}^{l_1} n_{1h} = n_1$. Аналогично вторая группа разбита на l_2 подгрупп с частотами $n_{21}, \ldots, n_{2l_2}$, $\sum_{h=1}^{l_2} n_{2h} = n_2$, и т. д. вплоть до k-й группы, где имеется l_k подгрупп с частотами $n_{k1}, \ldots, n_{kl_k}$, $\sum_{h=1}^{l_k} n_{kh} = n_k$. Мы получим лучшее соответствие с принятым нами соглашением об обозначениях, если заменим обозначение исходных частот в группах n_i из примера 35.1 на $n_{i\cdot}$, чтобы отметить, что внутри каждой исходной группы суммируются частоты n_{ih} соответствующих подгрупп. В этих обозначениях получим

$$\sum_{h=1}^{l_i} n_{ih} = n_{i\cdot}.$$

Описанная ситуация носит название иерархической *) двухфакторной классификации наблюдений. Этот термин употребляется, чтобы подчеркнуть, что внутри каждой исходной группы имеются подгруппы, в отличие от деления каждого набора данных из столбца на общие для всех столбцов группы, соответствующие строчкам при двухфакторной перекрестной классификации.

*) Другой термин «гнездовая классификация» («nested classification») кажется более уместно употреблять в тех случаях, когда в каждой исходной группе имеется равное число подгрупп. В связи с этим, несмотря на внешнюю привлекательность этого термина, мы не будем его употреблять.

Пример 35.5. ДА для двухфакторной иерархической классификации.

В примере 35.1 мы уже определили k параметров по одному для каждой группы. Для исследования изменения средних значений θ_{ih} для l_i подгрупп внутри i-й группы *) мы используем только l_i-1 линейно независимых параметров, поскольку можно положить

$$\sum_{h=1}^{l_i} n_{ih}\theta_{ih} = 0$$

(ср. **35.17—19** для перекрестной классификации), так что, так же как в (35.34), получим

$$\theta_{il_i} = -\frac{1}{n_{il_i}} \sum_{h=1}^{l_i-1} n_{ih}\theta_{ih}. \qquad (35.81)$$

Теперь можно обобщить линейную модель примера 35.1. Будем p-е наблюдение из h-й подгруппы i-й группы обозначать y_{ihp} и положим $l = \sum_{i=1}^{k} l_i$. Тогда

$$\underset{(n\times 1)}{\boldsymbol{y}} = \begin{pmatrix} y_{111} \\ \cdot \\ \cdot \\ \cdot \\ y_{11n_{11}} \\ y_{121} \\ \cdot \\ \cdot \\ \cdot \\ y_{12n_{12}} \\ \cdot \\ \cdot \\ \cdot \\ y_{1l_11} \\ \cdot \\ \cdot \\ \cdot \\ y_{1l_1n_{1l_1}} \\ y_{211} \\ \cdot \\ \cdot \\ \cdot \\ y_{kl_kn_{kl_k}} \end{pmatrix}, \qquad \underset{(l\times 1)}{\boldsymbol{\theta}} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \cdot \\ \cdot \\ \cdot \\ \theta_k \\ \theta_{11} \\ \cdot \\ \cdot \\ \cdot \\ \theta_{1,\,l_1-1} \\ \theta_{21} \\ \cdot \\ \cdot \\ \cdot \\ \theta_{2,\,l_2-1} \\ \cdot \\ \cdot \\ \cdot \\ \theta_{k1} \\ \cdot \\ \cdot \\ \cdot \\ \theta_{k,\,l_k-1} \end{pmatrix}$$

*) Эти θ_{ih} не имеют никакого отношения к параметрам взаимодействий при перекрестной классификации. В случае иерархической классификации проблемы взаимодействий не возникает.

и

$$\underset{(n\times l)}{\boldsymbol{X}} = \left[\begin{array}{c|cccc} \underset{(n\times k)}{\boldsymbol{X}} & \underset{(n_1.\times(l_1-1))}{\boldsymbol{X}} & & & \\ & & \underset{(n_2.\times(l_2-1))}{\boldsymbol{X}} & & 0 \\ & 0 & & \ddots & \\ & & & & \underset{(n_k.\times(l_k-1))}{\boldsymbol{X}} \end{array}\right] \qquad (35.82)$$

При этом $(n\times k)$ подматрица $\boldsymbol{X}$ в (35.82) та же самая, что и в примере 35.1. Как сразу следует из (35.81), остальные подматрицы имеют вид

$$\underset{(n_i.\times(l_i-1))}{\mathbf{X}} = \left[\begin{array}{ccccccc} 1 & & & & & & \\ \vdots & & & 0 & & & \\ 1 & & & & & & \\ 0 & 1 & & & & & \\ \vdots & \vdots & & 0 & & & \\ 0 & 1 & & & & & \\ 0 & 0 & 1 & & & & \\ \vdots & \vdots & \vdots & 0 & & & \\ 0 & 0 & 1 & & & & \\ & & & \ddots & & 1 & \\ & 0 & & & & \vdots & \\ & & & & & 1 & \\ -\frac{n_{i1}}{n_{il_i}}\mathbf{1} & -\frac{n_{i2}}{n_{il_i}}\mathbf{1} & & & & -\frac{n_{i,l_i-1}}{n_{il_i}}\mathbf{1} & \end{array}\right] \begin{array}{l} \left.\vphantom{\begin{array}{c}1\\ \vdots\\ 1\end{array}}\right\} n_{i1} \text{ строк} \\ \left.\vphantom{\begin{array}{c}1\\ \vdots\\ 1\end{array}}\right\} n_{i2} \text{ строк} \\ \left.\vphantom{\begin{array}{c}1\\ \vdots\\ 1\end{array}}\right\} n_{i3} \text{ строк} \\ \vdots \\ \left.\vphantom{\begin{array}{c}1\\ \vdots\\ 1\end{array}}\right\} n_{i,l_i-1} \text{ строк} \\ \left.\vphantom{\begin{array}{c}1\\ \vdots\\ 1\end{array}}\right\} n_{il_i} \text{ строк} \end{array} \qquad (35.83)$$

Из (35.82—83) теперь получаем

$$\underset{(l\times l)}{\boldsymbol{X}'\boldsymbol{X}} = \left[\begin{array}{cccccc} \begin{pmatrix} n_1. & & & 0 \\ & n_2. & & \\ & & \ddots & \\ 0 & & & n_k. \end{pmatrix} & & & & 0 \\ & \boldsymbol{A}_1 & & & \\ & & \boldsymbol{A}_2 & & \\ & & & \ddots & \\ 0 & & & & \boldsymbol{A}_k \end{array}\right], \qquad (35.84)$$

где

$$\underset{((l_i-1)\times(l_i-1))}{\boldsymbol{A}_i} = \begin{bmatrix} n_{i1}+\frac{n_{i1}^2}{n_{il_i}} & \frac{n_{i1}n_{i2}}{n_{il_i}} & \cdots & \frac{n_{i1}n_{i,l_i-1}}{n_{il_i}} \\ & \ddots & & \vdots \\ & & \ddots & \vdots \\ & & & n_{i,l_i-1}+\frac{n_{i,l_i-1}^2}{n_{il_i}} \end{bmatrix}$$

имеет тот же самый вид, что и в (35.37) и, значит, обратная к ней имеет тот же вид, что (35.44):

$$\boldsymbol{A}_i^{-1} = \begin{bmatrix} \frac{1}{n_{i1}}-\frac{1}{n_{i.}} & -\frac{1}{n_{i.}} & -\frac{1}{n_{i.}} & \cdots & -\frac{1}{n_{i.}} \\ & \frac{1}{n_{i2}}-\frac{1}{n_{i.}} & -\frac{1}{n_{i.}} & \cdots & -\frac{1}{n_{i.}} \\ & & \ddots & & \vdots \\ & & & \ddots & \vdots \\ & & & & \frac{1}{n_{i,l_i-1}}-\frac{1}{n_{i.}} \end{bmatrix}. \qquad (35.85)$$

Таким образом, из (35.84—5) получаем НК-оценки

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \begin{bmatrix} y_{1..} \\ y_{2..} \\ \vdots \\ y_{k..} \\ y_{11.}-y_{1..} \\ y_{12.}-y_{1..} \\ \vdots \\ y_{1,l_1-1,.}-y_{1..} \\ y_{21.}-y_{2..} \\ \vdots \\ y_{k,l_k-1,.}-y_{k..} \end{bmatrix}, \qquad (35.86)$$

и, значит, СК, соответствующая подогнанной модели, равна

$$\hat{\boldsymbol{\theta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\theta}} = \sum_{i=1}^{k} n_{i.}y_{i..}^2 + \sum_{i=1}^{k}\sum_{h=1}^{l_i} n_{ih}(y_{ih.}-y_{i..})^2. \qquad (35.87)$$

Первый член правой части (35.87) совпадает с выражением для S_1, определенным соотношением (35.16) примера 35.1. По сути дела, нам теперь остается разбить дополнительную СК, второй член в правой части (35.87). Поскольку S_1 представляет СК, соответствующую k параметрам исходных групп $\theta_1, \theta_2, \ldots, \theta_k$, то второй член в (35.87) представляет собой СК, соответствующую введенным теперь $\sum_{i=1}^{k}(l_i - 1) = l - k$ линейно независимым параметрам подгрупп. Результаты можно свести в таблицу

ДА для двухфакторной иерархической классификации

Компонента дисперсии	Число ст. св.	СК
Обусловленная группами	k	$\sum_i n_{i.} y_{i..}^2$
Между подгруппами внутри групп	$l-k$	$\sum_i \sum_h n_{ih}(y_{ih.} - y_{i..})^2$
Остаточная	$n-l$	$\sum_i \sum_h \sum_p (y_{ihp} - y_{ih.})^2$
Полная	n	$y'y$

(35.88)

Остаточная СК в (35.88) получена вычитанием. Первая строчка, так же как и в примере 35.1, может быть разбита на две компоненты «между групп» и «генеральное среднее».

Как следует из общей теории, отношение СрК «между подгруппами внутри групп» к остаточному СрК имеет нецентральное F-распределение с $(l-k, n-l)$ ст. св. и параметром нецентральности $\lambda = \sum_i \sum_h n_{ih}\theta_{ih}^2$, который равен нулю, если внутри каждой группы все средние значения подгрупп равны между собой. Тем самым для проверки этой гипотезы получается центральный F-критерий.

35.37 Иерархическая структура может быть исследована более подробно с подподгруппами и даже подподподгруппами. В результате получаются трехфакторная и соответственно четырехфакторная классификации, на практике встречающиеся довольно редко. Читателю должно быть понятно, что нет нужды заново проводить довольно утомительные выкладки метода НК для получения соответствующих результатов. В самом деле, по существу в примере 35.5 внутри каждой из k групп проводилось разбиение СК на две компоненты с $l_i - 1$ и $n_{i.} - l_i$ ст. св. соответственно. Затем компоненты суммировались по всем группам и получались члены из таблицы (35.88) с $l-k$ и $n-l$ ст. св. Тот же самый процесс разбиения можно теперь провести внутри каждой подгруппы и т. д. В упражнении 35.3 читателю предлагается провести трехфакторный ДА. Шеффе (1959) приводит теоретические выкладки для трехфакторного случая.

Многофакторные классификации

35.38 В предыдущем пункте мы указали схему применения многофакторной иерархической классификации. Это подводит нас к рассмотрению общих многофакторных классификаций.

Заметим сначала, что если мы рассматриваем трехфакторную классификацию, то при этом возможна «смешанная» классификация, которая является частично иерархической и частично перекрестной. Такая ситуация возникает, когда двухфакторная иерархическая классификация соответствует одному из факторов (скажем, строчкам) в $(r \times c)$-перекрестной классификации. В обозначениях примера 35.5 мы будем иметь $r = l$. ДА проводится в два этапа. Сперва анализируется перекрестная классификация с помощью обсуждавшихся выше методов. Полная СК разлагается на обычные пять компонент. Коротко мы представим это в виде следующей таблицы, указывая только число ст. св.

СК	Число ст. св.		
Генерального среднего	1		
Строчек	$l-1$	Строчки групп	$k-1$
		Строчки подгрупп	$l-k$
Столбцов	$c-1$		
Взаимодействий	$(l-1)(c-1)$	Взаимодействия с группами	$(k-1)(c-1)$
		Взаимодействия с подгруппами	$(l-k)(c-1)$
Остаточная	$n-lc$		
Полная	n		

На втором шагу каждая СК, которой соответствует иерархическая классификация, разбивается на две части в соответствии с тем, как указано в правой части таблицы. При этом

разбиение СК строк получается простым использованием примера 35.5 (стоит напомнить, что компонента генерального среднего при этом уже выделена из первой строчки (35.88) при анализе перекрестной классификации). Однако простейший способ разбиения СК строк и СК взаимодействий состоит в следующем. Надо объединить данные из подгрупп внутри одной группы и провести пересчет СК для строк и взаимодействий, используя объединенные данные. Таким образом получаются искомые компоненты СК с $(k-1)$ и $(k-1)(c-1)$ ст. св. соответственно. Если анализ ортогонален, то СК подгрупп получаются простым вычитанием.

Шеффе (1959) приводит теоретические выкладки для случая, когда при иерархической классификации внутри каждой группы имеется равное число подгрупп, а число наблюдений в каждой клетке $(l \times c)$-таблицы одинаково.

35.39 Предположим теперь, что вместо вложения двухфакторной иерархической классификации в двухфакторную перекрестную классификацию, как это было в **35.38**, у нас имеется новая однофакторная классификация внутри каждой клетки таблицы двухфакторной классификации. Если в каждой клетке проведена *одна и та же* однофакторная классификация, очевидно, что получается трехфакторная перекрестная классификация. Как мы сейчас увидим, при этом не только заново возникают все задачи формулировки линейной модели, обсужденные в **35.15—19** для двухфакторного случая, но и необходимы некоторые обобщения.

Трехфакторная перекрестная классификация

35.40 Следуя терминологии, уже использованной при обсуждении таблиц с тремя входами для категоризованных данных в **33.58**, рассмотрим теперь выборку из n наблюдений, представленную в виде $(r \times c \times l)$-таблицы с r строками, c столбцами и l «слоями» с числами наблюдений в клетках (частотами) n_{ijk}, где $i=1, \ldots, r$, $j=1, \ldots, c$, $k=1, \ldots, l$. При этом p-е наблюдение в (i, j, k)-й клетке обозначается y_{ijkp}, $p=1, \ldots, n_{ijk}$. Так же как в **35.15**, мы используем для обозначений следующее соглашение. Точка вместо какого-либо индекса при n означает *суммирование*, а при y — *усреднение* по этому индексу, причем вместо $n_{\ldots}$ пишется просто n.

Очевидно, что все вопросы из **35.15** имеют смысл в этой более общей ситуации для всех трех классификаций, так что теперь можно интересоваться генеральным средним, эффектами строк, эффектами столбцов, эффектами слоев (три множества *главных эффектов*), а также взаимодействиями между любой парой при классификации по строкам, столбцам и слоям. Однако из-за добавления дополнительного фактора появляется новая деталь, поскольку теперь нас может интересовать, зависят ли, например, сами взаимодействия между рядами и строчками от классификации по слоям. При этом получаются взаимодействия высших порядков, которые мы и собираемся теперь определить.

35.41 Обобщая соотношения (35.32), запишем линейную модель в виде

$$y_{ijkp} = \mu_{ijk} + \varepsilon_{ijkp} \equiv$$
$$\equiv \theta_{***} + \theta_{i**} + \theta_{*j*} + \theta_{**k} + \theta_{ij*} + \theta_{i*k} + \theta_{*jk} + \theta_{ijk} + \varepsilon_{ijkp}. \quad (35.89)$$

При этом среднее значение наблюдений в (i, j, k)-й клетке μ_{ijk} состоит из восьми компонент. Первые семь групп: генеральное среднее θ_{***}, эффекты строк θ_{i**}, эффекты столбцов θ_{*j*}, эффекты слоев θ_{**k}, взаимодействия столбцов и строк θ_{ij*}, взаимодействия строк и слоев θ_{i*k}, взаимодействия столбцов и слоев θ_{*jk}, определяются точно так же, как и в **35.17—18** для двухфакторной перекрестной классификации с добавлением лишней звездочки в индексе. Последнее множество параметров взаимодействий в (35.89), а именно θ_{ijk}, определяется путем обобщения аргументов пункта **35.18**. Если бы отклонение величины μ_{ijk} от генерального среднего в точности равнялось сумме определенных ранее трех соответствующих главных эффектов и трех соответствующих взаимодействий, то мы бы имели

$$\mu_{ijk} - \mu_{***} = \theta_{i**} + \theta_{*j*} + \theta_{**k} + \theta_{ij*} + \theta_{i*k} + \theta_{*jk} =$$
$$= (\mu_{i**} - \mu_{***}) + (\mu_{*j*} - \mu_{***}) + (\mu_{**k} - \mu_{***}) +$$
$$+ (\mu_{ij*} - \mu_{i**} - \mu_{*j*} + \mu_{***}) + (\mu_{i*k} - \mu_{i**} - \mu_{**k} + \mu_{***}) +$$
$$+ (\mu_{*jk} - \mu_{*j*} - \mu_{**k} + \mu_{***}),$$

или $\theta_{ijk} = 0$, где (ср. с определением θ_{ij} в (35.31))

$$\theta_{ijk} = \mu_{ijk} - (\mu_{ij*} + \mu_{i*k} + \mu_{*jk}) +$$
$$+ (\mu_{i**} + \mu_{*j*} + \mu_{**k}) - \mu_{***}. \quad (35.90)$$

Определенную в (35.90) величину θ_{ijk} мы назовем *взаимодействием второго порядка* между строками, столбцами и слоями. С этой точки зрения мы бы могли теперь взаимодействия в двухфакторной перекрестной классификации назвать *взаимодействиями первого порядка*. Для полной согласованности терминологии мы бы могли сами главные эффекты называть *взаимодействиями нулевого порядка*.

Переписывая (35.90) в виде

$$\theta_{ijk} = \{\mu_{ijk} - (\mu_{i*k} + \mu_{*jk}) + \mu_{**k}\} -$$
$$- \{\mu_{ij*} - (\mu_{i**} + \mu_{*j*}) + \mu_{***}\}, \quad (35.91)$$

и вспоминая определение θ_{ij} из (35.31), получим, что фактически θ_{ijk} равно разности между взаимодействием (первого порядка) столбцов и строк в k-м слое и тем же самым взаимодействием для объединения всех слоев. Поскольку θ_{ijk} определяется симметрично относительно строк, столбцов и слоев, то (35.91) можно в эквивалентной форме переписать как разность между взаимодействиями строк и слоев в j-м столбце и в объединении столбцов или как разность между взаимодействием столбцов и слоев в i-й строке и в объединении всех строк. По этой причине слово «взаимодействие» также уместно и для θ_{ijk}, как и для θ_{ij} в **35.18**.

35.42 Аналогично **35.17** теперь, при трехфакторной перекрестной классификации, мы можем иметь только rcl линейно независимых параметров, по одному на каждую клетку таблицы. Однако в модели (35.89) имеется

$$1+r+c+l+rc+rl+cl+rcl$$

параметров, и для того чтобы не получилась сингулярная модель, нужно исключить лишние параметры. Так же, как в **35.17**, мы исключим последние члены в множествах параметров главных эффектов, уменьшим множества параметров взаимодействий первого порядка до $(r-1)(c-1)$ и т. д. Таким образом, мы получим

$$1+(r-1)+(c-1)+(l-1)+(r-1)(c-1)+ \\ +(r-1)(l-1)+(c-1)(l-1)$$

параметров, за исключением взаимодействий второго порядка. Простым сложением читатель легко проверит, что добавление $(r-1)(c-1)(l-1)$ этих параметров доводит общее число параметров до требуемого числа rcl. Это очевидно также и из того факта, что rcl есть число тех μ_{ijk}, которые могут быть независимо определены, когда известны другие параметры.

35.43 Теперь ничто, кроме сложных алгебраических выкладок, не мешает нам проводить анализ совершенно аналогично тому, как это делалось в двухфакторном случае. В каждом конкретном случае не возникает вычислительных трудностей при подгонке линейной модели (35.85), оценке ее rcl параметров и доведении до конца ДА. Однако даже в двухфакторном случае существенные упрощения в алгебраических выкладках были только тогда, когда частоты в каждой клетке были пропорциональны произведению маргинальных частот, как это требовалось в (35.43). Аналогично для выполнения условия ортогональности теперь необходимо, чтобы для каждой клетки частота была пропорциональна произведению всех соответствующих маргинальных частот (см. Себер (1964а)). На практике это условие пропорциональности выполняется в случае

равного числа наблюдений во всех клетках таблицы, т. е. в сбалансированном случае.

В пропорциональном случае общие принципы анализа очень просты. В **35.39** мы видели, что трехфакторную перекрестную $(r\times c\times l)$-классификацию можно рассматривать как классификацию, порожденную введением новой классификации (с l уровнями) для каждой клетки существующей $(r\times c)$-перекрестной классификации. Отсюда следует, что ДА можно выполнять в два этапа, точно так же, как для «смешанной» классификации в **35.38**. Сперва $(r\times c)$-перекрестная классификация рассматривается как однофакторная классификация с rc уровнями и проводится двухфакторная $(rc\times l)$-перекрестная классификация для изучаемых нами наблюдений. Из (35.65) мы получаем следующую схематическую таблицу ДА.

СК	Число ст. св.		
Генерального среднего $(r\times c)$-перекрестной классификации	1		
	$rc-1$	строчки	$r-1$
		столбцы	$c-1$
		взаимодействия строчек и столбцов	$(r-1)(c-1)$
Классификации по слоям	$l-1$		
Взаимодействий $(r\times c)$-перекрестной классификации с l слоями	$(rc-1)\times(l-1)$	взаимодействия строчек и слоев	$(r-1)(l-1)$
		взаимодействия столбцов и слоев	$(c-1)(l-1)$
		взаимодействия второго порядка строчек столбцов и слоев	$(r-1)(c-1)(l-1)$
Остаточная	$n-rcl$		

(35.92)

На втором этапе СК, соответствующая $(r\times c)$-перекрестной классификации, разбивается на три части в соответствии с тем, как указано в правой колонке таблицы. Простейший метод проведения соответствующего разбиения (так же, как и в **35.38**) состоит в следующем. Сперва объединяются все столбцы внутри строк и в соответствии с этим пересчитываются две СК. В результате получаются нужные компоненты с $r-1$ и $(r-1)(l-1)$ ст. св. Если операцию объединения применить к исходным строкам внутри столбцов и снова пересчитать две СК, то получатся требуемые компоненты с $(c-1)$ и $(c-1)(l-1)$ ст. св.

Поскольку анализ ортогонален, то две оставшиеся СК (с $(r-1)(c-1)$ и $(r-1)(c-1)(l-1)$ ст. св.) получаются простым вычитанием.

Численные расчеты ДА в общем случае непропорциональных частот очень трудоемки даже при использовании ЭВМ. Пирс (1963) дает обзор общей ситуации. Г. Фримэн и Джефферс (1962) приводят методы анализа трехфакторной перекрестной классификации в неортогональном случае. Стивенс (1953) использует для этого случая итеративные методы. Браду (1965) решает задачи для простейшего случая, когда все взаимодействия всех порядков (кроме нулевого) предполагаются равными нулю (см. также Рис (1966)).

Габриэл (1963) приводит обзор теории для анализа средних значений в клетке, когда дисперсии обратно пропорциональны размеру выборки в клетке, специально исследуя случай, когда y принимает только значения 0 и 1 и средние значения в клетке становятся частотами. Приближенные методы анализа средних в клетках содержатся в упражнении 37.7.

Пример 35.6. Сбалансированная трехфакторная перекрестная классификация

В случае, когда числа наблюдений в клетках равны одному и тому же числу $m > 1$, то снова очевидно, что $(r \times c \times l)$-перекрестную классификацию можно на первом этапе подсчетов ДА из (35.92) рассматривать как $(r \times c) \times l$ или $(r \times l) \times c$ или $(c \times l) \times r$. Из симметрии следует, что подсчет СК для каждого из трех множеств главных эффектов и для каждого из трех множеств взаимодействий первого порядка проводится точно так же, как и в таблицах двухфакторной перекрестной классификации. Нужно только объединить все уровни третьего фактора. Остаточная СК, очевидно, также не изменяет свой вид. Используя введенные выше обозначения для трехфакторного случая, из (35.63—4) и упражнения 35.1 получим следующие выражения для СК, соответствующих компонентам таблицы (35.92), где индекс при S указывает теперь на число ст. св.

Генеральное среднее:

$$S_1 = \Big(\sum_i \sum_j \sum_k \sum_p y_{ijkp}\Big)^2 \Big/ (rclm).$$

Эффекты строк:

$$S_{(r-1)} = \sum_i \Big(\sum_j \sum_k \sum_p y_{ijkp}\Big)^2 \Big/ (clm) - S_1.$$

Эффекты столбцов:

$$S_{(c-1)} = \sum_j \Big(\sum_i \sum_k \sum_p y_{ijkp}\Big)^2 \Big/ (rlm) - S_1. \qquad (35.93)$$

Эффекты слоев:

$$S_{(l-1)} = \sum_k \Big(\sum_i \sum_j \sum_p y_{ijkp}\Big)^2 \Big/ (rcm) - S_1.$$

Взаимодействия строк и столбцов:

$$S_{(r-1)(c-1)} = \sum_i \sum_j \Big(\sum_k \sum_p y_{ijkp}\Big)^2 \Big/ (lm) - S_{(r-1)} - S_{(c-1)} - S_1.$$

Взаимодействия строк и слоев:

$$S_{(r-1)(l-1)} = \sum_i \sum_k \Big(\sum_j \sum_p y_{ijkp}\Big)^2 \Big/ (cm) - S_{(r-1)} - S_{(l-1)} - S_1.$$

Взаимодействия столбцов и слоев:

$$S_{(c-1)(l-1)} = \sum_j \sum_k \Big(\sum_i \sum_p y_{ijkp}\Big)^2 \Big/ (rm) - S_{(c-1)} - S_{(l-1)} - S_1.$$

Остаточная СК:

$$S_{rcl(m-1)} = \sum_i \sum_j \sum_k \sum_p y^2_{ijkp} - \sum_i \sum_j \sum_k \Big(\sum_p y_{ijkp}\Big)^2 \Big/ m.$$

Поскольку восемь компонент (35.93) в сумме с СК взаимодействий второго порядка дают, как обычно, $\sum_i \sum_j \sum_k \sum_p y^2_{ijkp}$, то упомянутая оставшаяся СК получается простым вычитанием:

Взаимодействия второго порядка:

$$S_{(r-1)(c-1)(l-1)} = \sum_i \sum_j \sum_k \Big(\sum_p y_{ijkp}\Big)^2 \Big/ m - S_{(r-1)(c-1)} - S_{(r-1)(l-1)} - S_{(c-1)(l-1)} - S_{(r-1)} - S_{(c-1)} - S_{(l-1)} - S_1. \qquad (35.94)$$

Сведем теперь результаты (35.92—94) в следующую таблицу ДА:

Компонента дисперсии, обусловленная	Число ст. св.	СК, определяемая (35.93—4)
Эффектами строчек	$r-1$	$S_{(r-1)}$
Эффектами столбцов	$c-1$	$S_{(c-1)}$
Эффектами слоев	$l-1$	$S_{(l-1)}$
Взаимодействиями строчек и столбцов	$(r-1)(c-1)$	$S_{(r-1)(c-1)}$
Взаимодействиями строчек и слоев	$(r-1)(l-1)$	$S_{(r-1)(l-1)}$
Взаимодействиями столбцов и слоев	$(c-1)(l-1)$	$S_{(c-1)(l-1)}$
Взаимодействиями строчек столбцов и слоев	$(r-1)(c-1)(l-1)$	$S_{(r-1)(c-1)(l-1)}$
Генеральным средним	$rcl-1$ 1	S_1
Классификацией Остаточная	rcl $rcl(m-1)$	$S_{rcl(m-1)}$
Полная	$rclm = n$	$y'y$

(35.95)

Как и раньше, любая гипотеза, соответствующая восьми строчкам (35.95), которые образуют часть СК «классификации», может быть проверена подсчетом отношения, соответствующего СрК к остаточному СрК. Каждое отношение имеет нецентральное F-распределение, которое становится центральным в том случае, если проверяемая гипотеза справедлива.

Многофакторные перекрестные классификации

35.44 Читателю должно быть теперь понятно, как повторным применением аргументов, использованных при получении трехфакторного анализа из двухфакторного, можно теперь анализ трехфакторной перекрестной классификации обобщить на четырехфакторную классификацию и на классификации с большим числом факторов. В сбалансированном случае симметрия таблицы (35.95) и соотношения (35.93—4) подсказывают более прямое обобщение на случай классификаций высших порядков. Следует отметить, что постоянное соответствие числа ст. св. у СК и числа линейно независимых параметров, которые относятся к этой СК, является, как нетрудно понять, простым следствием геометрической интерпретации, аналогичной **35.13.** Шеффе (1959) проводит детальное рассмотрение этого вопроса.

При обобщениях на многофакторный случай возникает необходимость определения взаимодействий третьего и более высоких порядков в тех случаях, когда в модели требуется их исследовать. Однако поскольку эти взаимодействия становятся все слабее, а трудности их интерпретаций возрастают, то часто их просто игнорируют при проведении анализа, объединяя относящуюся к ним СК с остаточной.

Объединение критериев ДА

35.45 Предположим сперва совсем в общем виде, что для проверки k различных гипотез мы имеем k критериев и что соответствующие им статистики стохастически независимы. Для получения критерия одновременного выполнения всех k гипотез мы можем использовать результат упражнения 16.4 в духе упражнения 30.9 т. 2. Применяя вероятностное интегральное преобразование к статистике каждого критерия и подправляя его так, чтобы критические значения каждой преобразованной статистики P_i соответствовали ее *малым* значениям, мы получим, что $-2\sum_{i=1}^{k}\log P_i = P$ имеет χ^2-распределение с $2k$ ст. св.; при этом критическими являются большие значения P. Для P мы используем критерий размера α при любом размере составляющих критериев *). Однако если статистики не являются независимыми, то мы сталкиваемся с теми же трудностями, что и в контексте критериев согласия в **30.36**, и приведенный составной критерий бесполезен, поскольку для нахождения распределения величины P необходимо знать совместное распределение исходных статистик.

Другой простой общий подход к объединению независимых критериев основывается на том факте, что если i-й критерий имеет размер α_i, то вероятность отклонения хотя бы одной из проверяемых гипотез, при условии, что все они справедливы, равна просто $1-\prod_{i=1}^{k}(1-\alpha_i)$. Если все α_i равны заданному числу α, то это сводится к

$$P_k(\alpha) = 1-(1-\alpha)^k, \tag{35.96}$$

что приблизительно равно αk, если α мало (как это обычно и бывает на практике).

35.46 Таким образом, если бы k критериев в таблице ДА были независимы, то мы могли бы использовать (35.96) для выбора их общего размера при изучении множества дисперсионных отношений как целого. Так, если бы мы имели четыре критерия и потребовали бы, чтобы полный размер (т. е. размер составного критерия) был 0,05, то для вычисления размера каждого из исходных критериев нам нужно было бы решить уравнение

$$0{,}05 = 1-(1-\alpha)^4$$

или приблизительно положить $\alpha = 0{,}05/4$. Однако критерии в таблицах ДА, которые мы рассматривали, никогда не являются независимыми критериями. Хотя различные СК в таблице могут быть независимы друг от друга, но у всех выведенных нами критериев в знаменателе соответствующей статистики стоит остаточная СК и, следовательно, эти критерии статистически зависимы. Так, например, если остаточная СК велика, то это уменьшает значение всех статистик одновременно.

35.47 Однако, к счастью, как показали Хартли (1955), соотношение (35.96) приближенно остается справедливым. Предположим, что наблюдаются $k+1$ независимых средних квадратов $s^2, s_1^2, \ldots, s_k^2$ соответственно со ст. св. $\nu, \nu_1, \ldots, \nu_k$. Пусть

*) Под вероятностным интегральным преобразованием автор подразумевает переход от статистики γ_i к статистике $P_i = F_i(\gamma_i)$, где $F_i(x)$ — функция распределения случайной величины γ_i. Формулируемый результат основан на свойствах независимых наблюдений из равномерного на отрезке [0, 1] распределения (см. **24.11** и **30.36** в т. 2). Широко употребительным в статистической литературе синонимом понятия «размер критерия» является термин «уровень значимости критерия». (*Прим. ред.*).

G, G_1, ..., G_k их функции распределения, а $g = G'$ — плотность распределения s^2. Пусть k значений F_i определяются, как решения при фиксированном α уравнений

$$\mathbf{P}\{s_i^2/s^2 \leqslant F_i\} = 1 - \alpha,$$

так что F_i является $100(1-\alpha)$-процентной квантилью распределения отношения s_i^2/s^2.

Вероятность того, что ни одно из отношений s_i^2/s^2 не превосходит соответствующего F_i, равна

$$P(1) = \mathbf{P}\left\{\frac{s_1^2}{s^2} \leqslant F_1, \ldots, \frac{s_k^2}{s^2} \leqslant F_k\right\} = \int_0^\infty \left\{\prod_{i=0}^{k} G_i(xF_i)\right\} g(x)\,dx. \quad (35.97)$$

Мы обозначили (35.97) через $P(1)$, поскольку соответствующее выражение совпадает со значением при $\theta = 1$ функции

$$P(\theta) = \int_0^\infty \prod_{i=1}^{k} [(1-\alpha) + \theta\{G_i(xF_i) - (1-\alpha)\}]\, g(x)\,dx, \quad (35.98)$$

откуда сразу следует, что $P(0) = (1-\alpha)^k$. Для разложения $P(1)$ по формуле Тейлора в окрестности нуля мы исследуем соответствующие производные. Имеем

$$P'(\theta) = \sum_{i=1}^{k} \int_0^\infty \{G_i(xF_i) - (1-\alpha)\} \prod_{\substack{j=1\\ j\neq i}}^{k} [(1-\alpha) + \theta\{G_j(xF_j) - (1-\alpha)\}]\, g(x)\,dx,$$

так что

$$P'(0) = (1-\alpha)^{k-1} \sum_i \int_0^\infty \{G_i(xF_i) - (1-\alpha)\}\, g(x)\,dx = 0,$$

поскольку

$$\int_0^\infty G_i(xF_i)\, g(x)\,dx = 1 - \alpha. \quad (35.99)$$

Таким образом, разложение Тейлора имеет вид

$$P(1) - (1-\alpha)^k = \frac{1}{2} P''(\bar\theta) = \frac{1}{2} \sum_{\substack{i,j=1\\ i\neq j}}^{k} \int_0^\infty \{G_i(xF_i) - (1-\alpha)\}\{G_j(xF_j) - (1-\alpha)\} \times \prod_{\substack{l=1\\ l\neq i,j}}^{k} [(1-\alpha) + \bar\theta\{G_l(xF_l) - (1-\alpha)\}]\, g(x)\,dx,$$

где $0 \leqslant \bar\theta \leqslant 1$. Каждый член в квадратных скобках принимает значения между нулем и единицей, поскольку этим же свойством обладают α, $\bar\theta$ и G_l. Таким образом, если заменить все эти члены единицей, то мы получим неравенство

$$|P(1) - (1-\alpha)^k| \leqslant \frac{1}{2} \sum\sum_{i\neq j} \int_0^\infty |G_i(xF_i)| - (1-\alpha)|\,|G_j(xF_j) - (1-\alpha)|\, g(x)\,dx,$$

из которого, применяя неравенство Коши — Буняковского *), получим следующее неравенство:

$$|P(1) - (1-\alpha)^k| \leqslant \frac{1}{2} \sum\sum_{i\neq j} \left[\int_0^\infty \{G_i(xF_i) - (1-\alpha)\}^2 g(x)\,dx \int_0^\infty \{G_j(xF_j) - (1-\alpha)\}^2 g(x)\,dx\right]^{1/2},$$

и далее

$$|P(1) - (1-\alpha)^k| \leqslant \frac{1}{2} k(k-1) \max_i \int_0^\infty \{G_i(xF_i) - (1-\alpha)\}^2 g(x)\,dx.$$

В силу (35.99) это можно переписать в виде

$$|P(1) - (1-\alpha)^k| \leqslant \frac{1}{2} k(k-1) \max_i [\mathbf{D}\{G_i(s^2F_i)\}]. \quad (35.100)$$

35.48 Хартли (1955) подсчитал верхнюю границу, даваемую (35.100) для широкой области значений k, ν и ν_i, когда $(1-\alpha)$ близко к 0,05. Соответствующее значение убывает до нуля при возрастании ν до бесконечности, но возрастает с ростом k и ν_i.

*) В оригинале это неравенство называется неравенством Коши — Шварца. (*Прим. ред.*).

Для достаточно неблагоприятного случая $k=10$, $\nu=30$, $\max_i \nu_i = 10$, верхняя граница равна 0,0138. Если при этом ν увеличить до 60, то верхняя граница уменьшится до 0,0050. При $k=5$, $\nu=60$, $\max_i \nu_i = 5$ имеем 0,0014. Поскольку истинная разность, вообще говоря, существенно меньше, чем верхняя граница (троекратная замена соответствующих членов на максимально возможные повышает границу), мы можем сделать вывод, что приближение величины (35.97) с помощью $(1-\alpha)^k$ вполне удовлетворительно в достаточно широкой области, особенно тогда, когда ν (число ст. св. остаточной СК) велико.

Таким образом, мы видим, что зависимость размера $P_k(\alpha) = 1 - P(1)$ критерия для множества дисперсионных отношений из таблицы ДА от номинального размера α критерия для каждого отдельного дисперсионного отношения из таблицы, приблизительно задается соотношением (35.96). Иначе, если мы перепишем (35.96) в виде

$$1-\alpha = \{1 - P_k(\alpha)\}^{1/k}$$

и подставим α вместо $P_k(\alpha)$ и $\beta(k)$ вместо α, то получим

$$\beta(k) = 1 - (1-\alpha)^{1/k}, \qquad (35.101)$$

так что критерий размера α для множества k дисперсионных отношений получается в том случае, если для каждого дисперсионного отношения взят критерий размера $\beta(k)$, где $\beta(k)$ определяется по α в соответствии с (35.101).

Первыми работами по этому вопросу были работы Хартли (1938) и Финни (1941). Кокрэн (1941) и Дарлинг (1952) рассматривали отношение наибольшего среднего квадрата к сумме всех средних квадратов в случае, когда все ν_i равны.

35.49 Из результатов **35.47—8** следует, что отдельные критерии размера (35.101) для k дисперсионных отношений с общей остаточной СК дают составной критерий приблизительно размера α. Однако тот факт, что стоящая в знаменателе СК является общей для всех k критериев, наводит на мысль, что в этом случае, возможно, неэффективно брать отдельные критерии, поскольку значение статистики любого одного из критериев дает некоторую информацию о значении других. Хартли (1955) предложил ступенчатую процедуру, которая учитывает этот факт.

Пусть H_i — вероятностное интегральное преобразование величины s_i^2/s^2 к равномерному распределению на (0, 1) (ср. **1.27** т. 1). Упорядочим H_i так, чтобы $H_{(i)}$ являлось i-м наименьшим из них. Поскольку $\mathbf{P}\{H_i \geqslant 1-\delta\} = \delta$, то критерий размера δ для H_i получается сравнением значения H_i с $1-\delta$.

Первый шаг состоит в проверке $H_{(k)}$, наибольшего из H_i с размером, равным, как и в (35.101), $1-(1-\alpha)^{1/k} = \beta(k)$. Если $H_{(k)} < 1-\beta(k)$, то сразу принимается гипотеза об однородности всех наблюдений. Если $H_{(k)} \geqslant 1-\beta(k)$, то соответствующая гипотеза об «отсутствии эффектов» в таблице ДА отклоняется, но продолжается проверка $H_{(k-1)}$ с размером $\beta(k-1)$. Если $H_{(k-1)} < 1-\beta(k-1)$, то принимаются $(k-1)$ оставшихся гипотез об отсутствии эффектов в таблице. Если же $H_{(k-1)} > 1-\beta(k-1)$, то соответствующие гипотезы отклоняются и переходят к проверке $H_{(k-2)}$ с размером $\beta(k-2)$. Так продолжается до тех пор, пока либо найдется такое i, что $H_{(i)} < 1-\beta(i)$ (в этом случае принимаются i остающихся гипотез), либо, если $H_{(1)} \geqslant 1-\beta(1)$, то все k гипотез об «отсутствии эффектов» отклоняются.

35.50 Нетрудно показать, что этот ступенчатый критерий имеет размер α. Предположим, что из k дисперсионных отношений c первых соответствуют «нулевым эффектам», а $k-c$ оставшихся — ненулевым, и что максимальным среди c преобразованных значений H_i (соответствующих нулевым эффектам) является $H_{(l)}$. Если какая-либо гипотеза о «нулевом эффекте» отклоняется процедурой проверки, то отклоняется и гипотеза, соответствующая $H_{(l)}$, поскольку она рассматривается первой при ступенчатой процедуре. Таким образом, для вероятности того, что какая-либо гипотеза о «нулевом эффекте» отклоняется, имеет место

$$P = \mathbf{P}\{H_{(i)} \geqslant 1-\beta(i),\ i = l, l+1, \ldots, k\} \leqslant \mathbf{P}\{H_{(l)} \geqslant 1-\beta(l)\}.$$

Поскольку $l \geqslant c$ и $\beta(i)$ — убывающая функция i, то справедливо

$$P \leqslant \mathbf{P}\{H_{(l)} \geqslant 1-\beta(c)\} = \alpha, \qquad (35.102)$$

поскольку, если $H_{(l)}$, наибольшая среди множества из c дисперсионных отношений, проверяется с размером $\beta(c)$, то в соответствии с (35.101) составной критерий имеет размер α. Из (35.102) следует, что размер последовательного критерия никогда не может превосходить α. Если $c=k$, то $l=k$ и (35.102) превращается в равенство.

В случае, когда все средние квадраты имеют равное число ст. св., Хартли (1955) приводит некоторые соображения о мощности соответствующего критерия.

Множественные сравнения

35.51 Каждый из критериев дисперсионного отношения в таблице ДА служит для проверки гипотезы о множестве параметров, например об эффектах строк или о взаимодействиях между строками и столбцами. Однако на практике часто бывает недостаточно знать, например, что эффекты строк θ_{i*}

различны. Нужно знать, какое из θ_{i*} можно считать большим, чем все другие, или, что встречается более часто, можно ли сказать, что θ_{i*} распадаются на отдельные группы.

НК-оценки $\hat{\theta}_{i*}$ являются, конечно, несмещенными линейными МД-оценками соответствующих параметров, и, используя их, можно строить оценки любой разности $\theta_{i*}-\theta_{s*}$. Но обычно мы не можем выделить разности, априори представляющие интерес. Таким образом, мы сталкиваемся с проблемой изучения некоторого числа критериев о разностях, которые не являются независимыми. Рассуждения **35.45—6** применяются здесь с очевидным изменением, хотя изучаемая задача является более сложной. В то время как в **35.47—50** мы имели дело с объединением критериев для проверки гипотез о множествах параметров, теперь нас интересует подробное рассмотрение отдельных множеств, например эффектов строк.

В известном смысле эта задача *множественного сравнения*, как она называется, является более сложным вариантом задачи «выпадающих» наблюдений, обсуждавшейся в **32.23—8**. Только вместо того, чтобы иметь дело со «сдвигом расположения» («location — shift») для одного или нескольких наблюдений, теперь мы интересуемся, будут ли средние значения (θ_{i*}) наблюденных величин ($\hat{\theta}_{i*}$) разбиваться на отдельные группы. В работе Тьюки (1953) содержится обзор результатов.

Критерий МЗР

35.52 Для определенности мы рассмотрим однофакторную классификацию примера 35.1, хотя ничто не изменилось бы, если бы мы рассматривали любое множество эффектов из таблицы ДА. В примере 35.1 наблюдаемые групповые средние $y_{i\cdot}$ являются НК-оценками для k групповых параметров θ_i (каждый из которых включает генеральное среднее как общее слагаемое). Если F-критерий (35.22) отклоняет гипотезы (35.20) о том, что все θ_i равны, то мы сталкиваемся с необходимостью указать, какие подмножества θ_i можно рассматривать как однородные и какие нет.

Самой простой является самая старая процедура («Стьюдент», 1908), состоящая в применении t-критерия Стьюдента для двух выборок к каждой из $k(k-1)/2$ возможных пар $y_{i\cdot}$, $y_{j\cdot}$, $i \neq j$. Если каждый из критериев имеет размер α, то очень мало можно сказать о размере составного критерия, поскольку попарные критерии не являются независимыми, и нет никакой простой формулы для объединения, аналогичной (35.96).

Применение этого составного критерия сводится к вычислению оцениваемой стандартной ошибки разности между двумя

средними (с использованием остаточной СК) и к сравнению наблюденного значения разности с табличным. При этом у распределения Стьюдента берется соответствующее (остаточной СК) число ст. св. Если числа наблюдений в группах (n_i) одинаковы (и равны, скажем, N), то необходимо вычислить только одну оценку стандартной ошибки вида $(2s^2/N)^{1/2}$, где s^2 — несмещенная оценка дисперсии ошибок σ^2. При этом мы получаем *минимальную значимую разность* (соответственно множество стандартных отклонений для двустороннего критерия размера α) и сравниваем с ней каждую из $k(k-1)/2$ наблюденных разностей. В дальнейшем эту процедуру мы будем иногда называть *МЗР-критерием* (критерий минимальной значимой разности). В общем случае о МЗР критерии нельзя сказать много больше того, что если истинные групповые средние не отличаются, то примерно α из всех пар будут ошибочно признаны «неоднородными».

Простая модификация МЗР-критерия, предложенная Фишером (1935), состоит в замене размера α составляющих критериев на $\alpha\Big/\binom{k}{2}$. Это приводит к тому, что число пар, ошибочно отклоняющих однородность (в случае, когда все групповые средние в самом деле равны), будет не $\frac{1}{2}k(k-1)\alpha$, а просто α.

Используя исходный (или модифицированный) МЗР критерий, следует помнить, что упомянутая доля ожидаемых ошибок является безусловной, не учитывающей того факта, что критерий применяется после того, как (и потому что) исходный F-критерий отклонил гипотезы об однородности.

Ступенчатые (пошаговые) и одновременные процедуры

35.53 Так же, как задача о выпадающем наблюдении (outlier problem) из **32.23—5**, задача множественного сравнения, имеющая с ней большое сходство, часто исследуется с помощью критериев выборочного размаха.

Простейшим является критерий Тьюки (1951, 1952) стьюдентизированного размаха. k групповых средних, которые мы теперь будем обозначать $\bar{x}_i$, $i=1, \ldots, k$, являются (в случае справедливости гипотезы об однородности) случайной выборкой размера k из генеральной совокупности, имеющей нормальное распределение с дисперсией σ^2/N, которая независимо оценивается величиной s^2/N, где N, как и ранее, число наблюдений в каждой группе. Упорядочим теперь групповые средние и переобозначим их так, что $\bar{x}_{(1)} \leqslant \bar{x}_{(2)} \leqslant \ldots \leqslant \bar{x}_{(k-1)} \leqslant \bar{x}_{(k)}$. Любой паре $\bar{x}_{(i)}$ и $\bar{x}_{(j)}$, $i<j$, отвечает разность $\bar{x}_{(j)}-\bar{x}_{(i)}$, которая соответствует размаху для подмножества из $j-i+1$

упорядоченных групповых средних с номерами от i до j. Это подмножество считается неоднородным (и экстремальные групповые средние рассматриваются как принадлежащие различным совокупностям), если $(\bar{x}_j - \bar{x}_i)/(s/N^{1/2})$ превосходит $100(1-\alpha)$-процентное значение q_α стьюдентизированного размаха для k наблюдений (см. **32.25**). Поскольку из того, что соответствующее неравенство не выполняется для $\bar{x}_{(k)} - \bar{x}_{(1)}$, следует, что оно не может выполняться для размаха любого подмножества, делаем вывод, что процедура имеет полный размер α.

На практике $\bar{x}_{(k)}$ последовательно сравнивается с $\bar{x}_{(1)}$, $\bar{x}_{(2)}$ и т. д. до тех пор, пока не получится утверждение об однородности. Если $\bar{x}_{(k)} - \bar{x}_{(1)}$ однородно, то процедура проверки сразу заканчивается. В противном случае средние $\bar{x}_{(1)}$, $\bar{x}_{(2)}$, ..., которые признаны не совпадающими с $\bar{x}_{(k)}$, теперь сравниваются с $\bar{x}_{(k-1)}$, снова последовательно начиная с $\bar{x}_{(1)}$. Затем все средние, которые признаны отличными от $\bar{x}_{(k-1)}$, сравниваются с $\bar{x}_{(k-2)}$ и т. д. до тех пор, пока никакое утверждение о «неоднородностях» не станет невозможно. Такая процедура, в которой решение о каждом подмножестве зависит от предыдущего решения, относящегося к большему подмножеству, называется *ступенчатой* или *пошаговой* (*step-by-step* или *stepwise*).

35.54 Другая *одновременная процедура,* использующая не размах, а суммы квадратов, была предложена Габриэлем (1964). В множестве, состоящем из k групп, выбирается любое подмножество, содержащее по крайней мере две группы (таких подмножеств имеется $2^k - k - 1$). Для *каждого* такого подмножества подсчитывается СК «между группами» в соответствии с (35.21), и результат сравнивается с *фиксированным* критическим значением, полученным из распределения дисперсионного отношения с $(k-1, n-k)$ ст. св. При этом размеры выборок n_i не обязаны быть равными. Такая процедура ведет к транзитивным суждениям в том смысле, что подмножество не может быть признано однородным, если однородными не являются все содержащиеся в нем меньшие подмножества (см. упражнение 35.16). Очевидно, что обычный критерий дисперсионного отношения является компонентой этой процедуры, соответствуя тому случаю, когда подмножество совпадает со всем множеством из k групп. Отсюда следует, что полный размер критерия равен α.

В случае, когда все n_i равны, критерий Тьюки из **35.53** можно модифицировать так, чтобы он был «одновременной процедурой» того же типа, что и у Габриэля. Для этого нужно применять критерий размаха уже к каждому подмножеству групповых средних, а не только к подмножествам с упорядоченными номерами (от i до j). (Полученный критерий обладает следующим дополнительным свойством: если какое-либо подмножество, состоящее более чем из двух групп, признается неоднородным, то по крайней мере одно из его подмножеств тоже должно быть признано неоднородным).

Описанные выше метод Габриэля и метод Тьюки — Габриэля обладают тем свойством, что вероятность ошибочности суждения о том, что подмножество неоднородно, убывает с размером подмножества. В случае, когда $k = 8$ и остаточная СК имеет 40 ст. св., это явление более заметно для первого метода (у Габриэля (1964) приведены соответствующие таблицы). При этом метод Тьюки — Габриэля более прост при вычислениях, особенно для больших k, однако его можно применять только при равных n_i.

35.55 Вместо того чтобы использовать фиксированное критическое значение стьюдентизированного размаха для k наблюдений, как это делается в **35.53**, можно сравнивать $(\bar{x}_{(j)} - \bar{x}_{(i)})/(s/N^{1/2})$ со стьюдентизированным размахом для $j - i + 1$ наблюдений, что еще раньше предложили делать Ньюмэн (1939) и Кейлс (1952). При этом возможна ситуация, когда в соответствии с этим критерием множество из q следующих друг за другом упорядоченных групповых средних признается «однородным», в то время как множество из p следующих друг за другом групповых средних, принадлежащих упомянутому выше множеству, признается «неоднородным». Ступенчатая процедура Ньюмэна — Кейлса утверждает неоднородность пары групповых средних только в том случае, если каждое подмножество следующих друг за другом групповых средних, содержащее эту пару, признано неоднородным с использованием определенного выше критерия стьюдентизированного размаха, учитывающего число элементов в подмножестве.

Процедура вычислений такая же, что и в конце **35.53**, за исключением того, что критическое значение в критерии стьюдентизированного размаха теперь не фиксировано, как это было раньше, а разное для составляющих критериев. При этом полный размер критерия снова, как легко видеть, равен α, поскольку, для того чтобы можно было делать вывод о неоднородности для какой-либо разности $\bar{x}_{(j)} - \bar{x}_{(i)}$, нужно сначала сделать соответствующий вывод о неоднородности для $\bar{x}_{(k)} - \bar{x}_{(1)}$.

Дункан (1952, 1955, 1957) предлагает процедуру, которая по существу является модификацией процедуры Ньюмэна — Кейлса. Каждая разность $\bar{x}_{(j)} - \bar{x}_{(i)}$ сравнивается с $100(1-\alpha_{j-i+1})$-процентным значением стьюдентизированного размаха для $j - i + 1$ наблюдений, где числа α_{j-i+1} определяются по $j - i$ следующим образом:

$$\alpha_{j-i+1} = 1 - (1 - \alpha_2)^{j-i}. \qquad (35.103)$$

Основанием служит (35.96) и тот факт, что критерий для $j-i+1$ наблюдений эквивалентен $j-i$ отдельным критериям для пар наблюдений.

При этом вероятность ошибки перераспределяется среди компонент критерия, падая в случае, когда сравниваемые групповые средние близки друг к другу после упорядочивания (см. **35.54**). Дункан (1955) приводит таблицы для случаев, когда размер общего критерия $\alpha = 0{,}05$ и $\alpha = 0{,}01$.

35.56 Дункан (1955), Тьюки (1953) и Шеффе (1959) рассматривают некоторые другие процедуры множественного сравнения. Хартер (1957) проводит сравнение мощностей, а Хартли (1955) кратко изучает мощность метода Ньюмэна — Кейлса. Габриэл (1964) сравнивает ступенчатые методы Ньюмэна — Кейлса и Дункана с двумя одновременными процедурами, описанными в **35.54**, и доказывает, что при заданном размере общего составного критерия и любом последовательном методе вероятность неверного суждения о неоднородности любого подмножества не может быть меньше, чем для одновременной процедуры, основанной на той же самой статистике.

Мак Дональд и В. Томпсон (1967) развивают методы множественных сравнений ранжированных (упорядоченных) сумм для случая одно- и двухфакторной классификации.

Одновременные доверительные интервалы для разностей и сравнений

35.57 Как указал Тьюки (1951), из критерия стьюдентизированного размаха в **35.53** сразу получаются доверительные интервалы одновременно для всех $\frac{1}{2}k(k-1)$ разностей между истинными групповыми средними $(\theta_i - \theta_j)$. В самом деле, какими бы ни были θ_i, случайные величины $\bar{x}_i - \theta_i$ независимы и имеют одинаковое нормальное распределение с нулевым математическим ожиданием и дисперсией σ^2/N. Вероятность того, что стьюдентизированный размах не превзойдет величины q_α, определенной в **35.53**, равна $1-\alpha$. Отсюда следует, что одновременно для всех $i \neq j$

$$\mathsf{P}\left\{\left|\frac{(\bar{x}_i-\theta_i)-(\bar{x}_j-\theta_j)}{s/N^{1/2}}\right| \leqslant q_\alpha; \quad i, j = 1, \ldots, k;\ i \neq j\right\} = 1-\alpha, \tag{35.104}$$

так что с вероятностью $1-\alpha$ неравенства

$$(\bar{x}_i - \bar{x}_j) - q_\alpha s/N^{1/2} \leqslant \theta_i - \theta_j \leqslant (\bar{x}_i - \bar{x}_j) + q_\alpha s/N^{1/2} \tag{35.105}$$

выполняются одновременно для всех $i \neq j$.

В упражнении 35.11 показано, что метод обобщается на многомерный нормальный вектор с одинаково (отрицательно) коррелированными компонентами $\bar{x}_i$.

35.58 Метод, изложенный в **35.57**, позволяет нам формулировать утверждения обо всех $\frac{1}{2}k(k-1)$ разностях $\theta_i - \theta_j$ одновременно с известным общим коэффициентом доверия $1-\alpha$. Однако во многих применениях ДА нас могут интересовать не только разности, но также и другие линейные комбинации величин θ_i с постоянными коэффициентами, равными в сумме нулю. Такие линейные комбинации, называемые *сравнениями* (*contrasts*), определяется следующим образом:

$$\psi = \sum_{i=1}^{k} c_i\theta_i, \quad \sum_{i=1}^{k} c_i = 0. \tag{35.106}$$

Самым простым полезным сравнением, отличным от разности между любыми θ_i и θ_j, является разность между усредненным значением θ_i для любого подмножества, содержащего p из k параметров и усредненным значением оставшихся $k-p$ параметров. Как легко видеть, взаимодействия (определенные в **35.18**, **35.41**) также являются сравнениями.

Метод, указанный в **35.57**, легко видоизменить так, чтобы не только для разностей величин θ_i, но и для каждого сравнения указывался одновременно соответствующий интервал. Поскольку число сравнений бесконечно, то получается большое увеличение общности метода.

35.59 Обозначим $z_i = \bar{x}_i - \theta_i$, и пусть $\sum_{i=1}^{k} c_i = 0$. Рассмотрим максимально возможное значение суммы $\sum c_i z_i$. Поскольку $\sum c_i = 0$, то сумма положительных c_i, так же как и сумма отрицательных, равна $\frac{1}{2}\sum_i |c_i|$. Отсюда получается, что

$$\sum_i c_i z_i \leqslant \left(\frac{1}{2}\sum_i |c_i|\right) \max |z_i - z_j|,$$

т. е.

$$\sum_i c_i(\bar{x}_i - \theta_i) \leqslant \left(\frac{1}{2}\sum_i |c_i|\right) \max |(\bar{x}_i - \theta_i) - (\bar{x}_j - \theta_j)|. \tag{35.107}$$

Возвращаясь к (35.104), мы видим, что из (35.107) следует, что при любом выборе c_i с $\sum_i c_i = 0$

$$\mathsf{P}\left\{\left|\frac{\sum_i c_i(\bar{x}_i - \theta_i)}{s/N^{1/2}}\right| \leqslant \left(\frac{1}{2}\sum_i |c_i|\right) q_\alpha\right\} = 1-\alpha,$$

а значит, с вероятностью $1-\alpha$ неравенства

$$\sum_i c_i\bar{x}_i - \left(\frac{1}{2}\sum_i |c_i|\right) q_\alpha s/N^{1/2} \leqslant \sum_i c_i\theta_i \leqslant$$

$$\leqslant \sum_i c_i\bar{x}_i + \left(\frac{1}{2}\sum_i |c_i|\right) q_\alpha s/N^{1/2} \qquad (35.108)$$

выполняются одновременно для всех сравнений $\psi = \sum c_i\theta_i$. Метод снова обобщается на многомерный нормальный вектор с одинаково отрицательно коррелированными компонентами $\bar{x}_i$ (см. упражнение 35.11).

35.60 В **35.59** доверительные интервалы, справедливые одновременно для всех сравнений, были получены использованием довольно грубых неравенств. Поэтому не удивительно, что они не являются самыми лучшими из возможных доверительных интервалов для всех сравнений. Мы получим более точные интервалы, используя совсем иной подход.

Оценкой любого сравнения (35.106) является

$$\hat{\psi} = \sum_i c_i\hat{\theta}_i. \qquad (35.109)$$

Очевидно, что

$$\mathsf{M}(\hat{\psi}) = \psi,$$

и если $\hat{\theta}_i$ нормально распределены, то и $\hat{\psi}$ имеет нормальное распределение. Если мы рассмотрим теперь любое множество из $r(\leqslant k)$ оцениваемых сравнений, которые мы запишем в виде $\hat{\boldsymbol{\psi}} = \boldsymbol{C}\hat{\boldsymbol{\theta}}$, то будем иметь вектор, имеющий многомерное нормальное распределение (см. **15.4** т. 1) с вектором средних $\boldsymbol{\psi} = \boldsymbol{C}\boldsymbol{\theta}$ и дисперсионной матрицей

$$\boldsymbol{V} = \boldsymbol{V}(\hat{\boldsymbol{\psi}}) = \boldsymbol{C}\boldsymbol{V}(\hat{\boldsymbol{\theta}})\boldsymbol{C}'. \qquad (35.110)$$

В обсуждавшемся выше случае однофакторной классификации с одинаковыми частотами, равными N, матрица $\boldsymbol{V}(\hat{\boldsymbol{\theta}})$ диагональна и ее диагональные элементы равны σ^2/N, так что

$$\boldsymbol{V} = \frac{\sigma^2}{N}\boldsymbol{C}\boldsymbol{C}'. \qquad (35.111)$$

Предполагается, что $\boldsymbol{V}$ не вырождена.

35.61 Из результатов **15.10** теперь следует, что квадратичная форма

$$Q = (\hat{\boldsymbol{\psi}} - \boldsymbol{\psi})'\boldsymbol{V}^{-1}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi})$$

имеет χ^2-распределение с числом ст. св., равным r, где r — ранг матрицы $\boldsymbol{V}$. Не зависящая от Q остаточная СК (деленная на σ^2) имеет распределение того же вида, скажем, с ν ст. св. Тогда отношение

$$F = (Q/r)/(s^2/\sigma^2),$$

где s^2, как обычно, равно остаточной СК, деленной на ν, имеет распределение дисперсионного отношения с (r, ν) ст. св. В простейшем случае (35.111) отсюда получаем статистику

$$F = (\hat{\boldsymbol{\psi}} - \boldsymbol{\psi})'(\boldsymbol{C}\boldsymbol{C}')^{-1}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi})/(rs^2/N).$$

Если $100(1-\alpha)$-процентное значение этого F-распределения обозначить $F_{\alpha, r, \nu}$, то получим

$$\mathsf{P}\{(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi})'(\boldsymbol{C}\boldsymbol{C}')^{-1}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi})/(s^2/N) \leqslant rF_{\alpha, r, \nu}\} = 1-\alpha. \qquad (35.112)$$

В общем случае соответствующий результат получается, если вместо (35.111) использовать соотношение (35.110).

35.62 Поскольку матрица $\boldsymbol{V}$ должна быть невырожденной, то ее ранг r не может быть больше q, где q — максимальное число линейно независимых сравнений величин θ_i (как следует из примера 35.1, в случае однофакторной классификации имеем $q = k-1$). В общем случае q совпадает с числом ст. св. соответствующей СК в таблице ДА. Соотношение (35.112) при $r = q$ справедливо для *любого* множества из q линейно независимых сравнений, но отсюда не следует, что соответствующее утверждение выполняется одновременно для *каждого* такого множества. Однако Шеффе (1953, 1959) показал, используя геометрический подход, что вероятность того, что величины всех сравнений одновременно удовлетворяют неравенствам

$$(\hat{\psi} - \psi)^2 \leqslant qF_{\alpha, q, \nu}\hat{V}(\hat{\psi}), \qquad (35.113)$$

равна $1-\alpha$. Здесь $\hat{V}(\hat{\psi})$ есть оценка дисперсии величины $\hat{\psi}$, где σ^2 оценено величиной s^2 с ν ст. св. В упражнении 35.12 содержится аналитическое, а в упражнении 35.19 чрезвычайно простое алгебраическое доказательство этого факта.

Шеффе (1953, 1959) идет дальше и вычисляет, что интервалы для всех сравнений, задаваемые (35.113), вообще говоря, меньше, чем интервалы, задаваемые (35.108), за исключением двух случаев. Во-первых, когда все сравнения сводятся к разностям. Тогда (35.108) сводится к (35.105), специально выведенному для разностей. Во-вторых, когда имеется очень мало ненулевых c_i. Более того, в методе Шеффе не обязательно требовать, чтобы все $(\bar{x}_i - \theta_i)$ имели равную дисперсию, а это требование лежит в основе рассуждений **35.57**.

35.63 Возвращаясь к критерию дисперсионного отношения для полной гипотезы о том, что θ_i равны ((35.22) в случае однофакторной классификации), мы видим, что эта гипотеза (см. (35.20)) утверждает, что q линейно независимых сравнений

равны все нулю. Отсюда следует, что *все возможные* сравнения равны нулю, поскольку каждое сравнение можно рассматривать как линейную комбинацию заданных линейно независимых. Таким образом, составной критерий логически эквивалентен проверке гипотезы о том, что каждое из бесконечного числа возможных сравнений равно нулю, т. е. проверке того, будет ли хотя бы один из бесконечного числа интервалов, задаваемых (35.113), не содержать нуль (см. также упражнения 35.12 и 35.19). Это свойство распространяется сразу на процедуру одновременной проверки Габриэля (1964) из **35.54.** При этом подмножество считается неоднородным тогда и только тогда, когда некоторое сравнение для элементов этого множества имеет интервал (35.113), не покрывающий нуль.

В этом основная польза метода всех сравнений Шеффе. Если составной критерий отвергает гипотезу однородности, то метод всех сравнений можно использовать для рассмотрения любых сравнений с целью обнаружить, являются ли они истинной причиной отклонения гипотезы, и построения для них доверительных интервалов. Причем эти сравнения не нужно выделять заранее. Естественный путь разыскания сравнений, ответственных за отклонение полной однородности, состоит в первоначальном рассмотрении $\frac{1}{2}k(k-1)$ разностей. Все это можно проделать, не влияя на размер составного критерия. Если читатель вернется теперь назад к исходному обсуждению целей множественного сравнения в **35.51**, то, возможно, он согласится, что метод всех сравнений Шеффе очень близок к достижению поставленных там целей.

Габриэл (1967) приводит общую теорию для одновременных процедур проверки.

35.64 Данн (1961) рассматривает процедуру, промежуточную между построением доверительного интервала для одного сравнения и построением доверительных интервалов для всех сравнений. В ее методе требуется, чтобы априори были выделены m сравнений, представляющих особый интерес. Получающиеся интервалы (основанные на t-статистике Стьюдента) короче, чем интервалы, полученные как методом Шеффе, так и методом Тьюки, если k (число параметров) больше двух, а m не очень велико. Выгода увеличивается при увеличении k или числа ст. св. остатка, или коэффициента доверия $1-\alpha$. В упражнении 35.14 приведен очень простой результат, который лежит в основе этого метода. Сиотани (1964) улучшает процедуру.

Борер (1967) дает улучшение метода Шеффе в случае, когда рассматриваются сравнения для переменных, про которые известно, что все они положительны.

Упорядоченная и метрическая классификации

35.65 В этой главе рассматривалась совсем общая классификация переменных. Не делалось никаких предположений, например о том, упорядочены ли каким-либо образом группы классификации. Однако если имеется полная информация о принципах классификации, то СК в таблице ДА могут быть разбиты дальше на соответствующие компоненты. Например, если известно, что группы соответствуют «равноотстоящим» («equally-spaced») значениям изучаемой переменной, то можно использовать ортогональные полиномы, рассмотренные в **28.18—20** т. 2, чтобы поставить в соответствие отдельную ст. св. линейному, квадратичному, кубическому эффектам и эффектам более высоких степеней для переменной, по которой производится классификация. Если необходимо, то таким способом можно исчерпать все $k-1$ ст. св. Используемый метод в точности совпадает с методом, рассмотренным в примере 28.3.

В более сложных классификациях взаимодействия, так же как эффекты строк (или другие), могут быть разбиты аналогичным образом, если все изучаемые переменные разбиваются на группы в соответствии с расположением. Методы вычислений приведены у Р. Андерсона и Банкрофта (1952), а также в таблицах Фишера и Иэйтса.

35.66 Бартоломью (1961) рассматривает случай упорядоченной классификации как альтернативу гипотезы об однородности групп. Получается та же ситуация, что и рассмотренная в **31.74.** Если все n_i равны, то оказывается, что свободный от распределения критерий, основанный на (31.151), имеет асимптотически бóльшую мощность, чем критерий ОП, если все θ_i равноотстоящие, и меньшую мощность, если имеет место другой крайний случай, когда равны все θ_i, за исключением одного (см. упражнение 35.15, а также статью Чако (1963)). Шорак (1967) обобщает результат на более сложные ситуации, включая обсуждение свободных от распределения критериев для альтернатив об упорядочениях.

Ковариационный анализ

35.67 Естественное обобщение ДА возникает тогда, когда при анализе классифицируемых наблюдений, которые мы рассматриваем в этой главе, наряду с наблюденными значениями величины y мы узнаем значение одной (или большего числа) переменной x, про которую известно (или предполагается), что она влияет на значения y. Если бы данные не были классифицированы, то мы должны были бы провести обычный регрессионный анализ y относительно x, но теперь мы хотим

исследовать влияние на y как классификации (возможно сложной), так и измеренных значений x. Такой анализ преследует не только одну цель.

Вообще говоря, регрессионные методы признаны выявить (и устранить) влияние значений x на y, так что после того, как учтено влияние x, можно свободно анализировать влияние на y классификации. Так бывает, если причины, влияющие на x, предшествуют причинам, влияющим на y (т. е. предшествуют испытаниям, приводящим к классификации). Например, если измеряется влияние методов обучения на знание детьми в школе какого-либо предмета, то в качестве x следует взять начальный уровень знаний этого предмета, а в качестве y — конечный. В качестве x можно взять также общий уровень знаний. По-видимому, в качестве переменных x стоит рассматривать как общий начальный уровень знаний, так и начальный уровень знаний данного предмета. Такой анализ можно назвать анализом, призванным устранить влияние x. Его методы призваны гарантировать чистое сравнение испытаний и также уменьшить остаточную дисперсию, вызванную тем, что x принимает разные значения.

Понятно, что если испытания, приводящие к классификации, не влияют на значения x, как это было в приведенном примере, то интерпретация становится проще, но анализ можно проводить и в других случаях.

Можно, например, в отличие от описанной ситуации (или в дополнение к ней) интересоваться, влияет ли классификация, как таковая, на регрессию y относительно x. В рассмотренном примере мы бы могли спросить, будет ли регрессия конечного уровня знаний относительно начального уровня знаний одной и той же при применении любого метода обучения. Целью анализа в этом случае является не устранение влияния, а особый интерес к соотношению между переменными.

35.68 Эту часть предмета называют *ковариационным анализом*, поскольку в регрессионных вычислениях участвуют разбиения произведений y и x точно так же, как в ДА — разбиения СК. Переменные x обычно называются *сопутствующими*, подчеркивая тем самым, что основной интерес представляет y.

Обширный обзор применений ковариационного анализа содержится в семи статьях специального выпуска Biometrics (т. 13, выпуск 3, сентябрь 1957). Д. Кокс (1958) дает доступное введение в предмет. Ниже мы рассмотрим только теоретические аспекты.

35.69 Мы уже видели, что как регрессионный анализ, так и ДА можно рассматривать в рамках линейной модели. Поэтому очевидно, что ковариационный анализ, являющийся их смесью, также можно рассматривать в этих рамках. Как теперь понятно, простота интерпретации в том случае, когда испытания не влияют на сопутствующие переменные, соответствует теперь упрощению линейной модели в случае некоррелированных множеств регрессоров. Аналогично можно сформулировать линейную модель для любой ситуации ковариационного анализа.

Однако этой утомительной процедуры можно избежать, если известна форма ДА, который (как говорят) вложен в ковариационный анализ. В этом случае можно провести расширение ДА для проведения связанного с ним ковариационного анализа, взяв просто соответствующее расширение линейной модели. Более того, оказывается, что это расширение является весьма общим. Оно позволяет вводить дополнительные параметры в существующую линейную модель. Как указывает сам заголовок, приведенные ниже алгебраические выкладки не ограничиваются ситуациями ДА.

Расширение линейной модели для учета новых параметров

35.70 Пусть рассматривается линейная модель (вырожденная или невырожденная)

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \tag{35.114}$$

и НК-оценка параметров θ, полученная методами главы 19, равна

$$\hat{\hat{\boldsymbol{\theta}}} = \boldsymbol{T}\boldsymbol{y}, \tag{35.115}$$

где, разумеется, в невырожденном случае $\boldsymbol{T} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Тогда остаточная СК равна

$$(\boldsymbol{y} - \boldsymbol{X}\hat{\hat{\boldsymbol{\theta}}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\hat{\boldsymbol{\theta}}}) = \{(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{T})\boldsymbol{y}\}'\{(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{T})\boldsymbol{y}\} = \\ = \boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{T})\boldsymbol{y}, \tag{35.116}$$

поскольку $\boldsymbol{T}\boldsymbol{X} = \boldsymbol{I}$ в силу несмещенности $\hat{\hat{\boldsymbol{\theta}}}$. Матрица $(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{T})$ идемпотентна.

35.71 Рассмотрим теперь расширенную модель

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

или

$$\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}. \tag{35.117}$$

Если бы $\boldsymbol{\beta}$ было известным, то, согласно (35.114—15), мы имели бы решение

$$\hat{\hat{\boldsymbol{\theta}}} = \boldsymbol{T}(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}), \tag{35.118}$$

а, значит, решение по методу НК модели (35.117) относительно $\hat{\boldsymbol{\theta}}$ и $\hat{\boldsymbol{\beta}}$ может быть получено решением только относительно $\hat{\boldsymbol{\beta}}$

модели

$$y = X \cdot T(y - Z\beta) + Z\beta + \varepsilon$$

или

$$(I - XT)\,y = (I - XT)\,Z\beta + \varepsilon. \qquad (35.119)$$

(35.119) представляет собой линейную модель, и мы предположим, что она не вырождена. Тогда из главы 19 получим

$$\hat{\beta} = \{Z'(I - XT)\,Z\}^{-1} Z'(I - XT)\,y, \qquad (35.120)$$

$$V(\hat{\beta}) = \sigma^2 \{Z'(I - XT)\,Z\}^{-1}. \qquad (35.121)$$

35.72 Уменьшение остаточной СК, обусловленное расширением модели, равно

$$\{(I - XT)\,Z\hat{\beta}\}'\,\{(I - XT)\,Z\hat{\beta}\} = \hat{\beta}'Z'(I - XT)\,y. \qquad (35.122)$$

Сравнение с (35.116) показывает, что в обоих этих соотношениях содержится одна и та же матрица $(I - XT)$. (35.122) отличается только тем, что эта матрица слева умножается не на y', а на $\hat{\beta}'Z'$. Это соображение позволяет упростить вычисление (35.122). Вместо квадратичной формы от y нужно рассмотреть совокупность билинейных форм, получающихся в результате подстановки в (35.116) вместо y' по очереди всех столбцов матрицы Z. Эти билинейные формы нужно расположить в виде столбца и умножить слева на $\hat{\beta}'$. В результате получится (35.122).

Разность между (35.116) и (35.122)

$$(y - Z\hat{\beta})'(I - XT)\,y \qquad (35.123)$$

является остаточной СК, когда в качестве подогнанной рассматривается расширенная модель (35.117).

Легко видеть, что если минимизировать СК при каких-либо ограничениях на элементы вектора θ, то уменьшение будет иметь точно такой же вид, что и (35.122). Единственное изменение сводится к замене $(I - XT)$ на матрицу Q, соответствующую квадратичной форме минимизируемой СК. Аналогами (35.122—3) являются тогда $\hat{\beta}'Z'Qy$ и $(y - Z\hat{\beta})'Qy$ соответственно.

35.73 Применение результатов **35.70—2** в случае ковариационного анализа, очевидно. Фигурирующим в ДА суммам квадратов соответствуют суммы произведений в (35.122). Шеффе (1959) приводит правила вычислений и примеры. У Р. Андерсона и Банкрофта (1952), так же как в упомянутом выпуске Biometrics (сентябрь 1957), приведены вычисления для конкретных ситуаций.

УПРАЖНЕНИЯ

35.1 Проверить, что если $n_{ij} = m$, то СК в (35.63—4) сводятся к

$$S_4 = \Big(\sum_i \sum_j \sum_p y_{ijp}\Big)^2 \Big/ (rcm),$$

$$S_1 = \sum_i \Big(\sum_j \sum_p y_{ijp}\Big)^2 \Big/ (cm) - S_4,$$

$$S_2 = \sum_j \Big(\sum_i \sum_p y_{ijp}\Big)^2 \Big/ (rm) - S_4,$$

$$S_3 = \sum_i \sum_j \Big(\sum_p y_{ijp}\Big)^2 \Big/ m - (S_1 + S_2 + S_4),$$

$$S_R = \sum_i \sum_j \sum_p y_{ijp}^2 - (S_1 + S_2 + S_3 + S_4),$$

и показать, что если $m = 1$, то эти СК принимают вид (суммирование по p теперь излишне)

$$S_4 = \Big(\sum_i \sum_j y_{ij}\Big)^2 \Big/ (rc),$$

$$S_1 = \sum_i \Big(\sum_j y_{ij}\Big)^2 \Big/ c - S_4,$$

$$S_2 = \sum_j \Big(\sum_i y_{ij}\Big)^2 \Big/ r - S_4,$$

$$S_3 = \sum \sum y_{ij}^2 - (S_1 + S_2 + S_4),$$

$$S_R = 0.$$

35.2 Проверить, что если $n_{ij} = m$, то матрица C, определяемая в (35.39), принимает вид

$$\underset{(r-1)(c-1)\times(r-1)(c-1)}{C} = m \begin{pmatrix} 2E & E & E & \dots & E \\ E & 2E & E & \dots & E \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & E \\ E & \dots & \dots & E & 2E \end{pmatrix},$$

где

$$\underset{(c-1)\times(c-1)}{E} = \begin{pmatrix} 2 & 1 & \dots & 1 \\ 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \dots & 1 & 2 \end{pmatrix}.$$

Показать, что

$$C^{-1} = \frac{1}{rm} \begin{pmatrix} (r-1)E^{-1} & -E^{-1} & \dots & -E^{-1} \\ -E^{-1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & -E^{-1} \\ -E^{-1} & \dots & -E^{-1} & (r-1)E^{-1} \end{pmatrix},$$

где

$$E^{-1} = \frac{1}{c}\begin{pmatrix} c-1 & -1 & \dots & -1 \\ -1 & c-1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & -1 \\ -1 & \dots & -1 & c-1 \end{pmatrix},$$

и вывести отсюда, что НК-оценки величин θ_{ij} задаются соотношением (35.50).

35.3 Обобщить таблицу ДА из примера 35.5 на случай трехфакторной иерархической классификации и привести три критерия проверки гипотез о том, что нет различия а) между групповыми средними, б) между средними для подгрупп внутри групп, в) между средними подподгрупп внутри подгрупп.

35.4 В примере 35.4 показать, что если постулировать выполнение H_3 из (35.54), т. е. считать известным, что нет взаимодействий, то СК, соответствующая эффектам строк, равна $M - S_2$, а СК, соответствующая эффектам столбцов, равна $M - S_1$, где S_1 и S_2 определены в (35.59—60). Показать, что в этом случае остаточный СрК имеет $(n - r - c + 1)$ ст. св.

35.5 Показать, что если в $(r \times 2)$-перекрестной классификации разности клеточных средних $d_i = y_{i1.} - y_{i2.}$ анализировать методом взвешенных квадратов средних значений, примененным в **35.31** к $\bar{y}_i$, то СК $\sum_i W_i \left(d_i - \sum_i W_i d_i \Big/ \sum_i W_i\right)^2$, где $W_i = n_{i1}^{-1} + n_{i2}^{-1}$, приводит к критерию для проверки гипотезы о том, что все взаимодействия равны нулю.

(Иэйтс, 1934.)

35.6 В $(2 \times 2 \times 2 \times \ldots = 2^m)$-перекрестной классификации показать, что если образовать $(2 \times 2^{m-1})$-таблицу для любого из признаков (скажем, A) против всех возможных комбинаций других, то невзвешенное среднее разностей строчно-клеточных средних доставляет эффективную оценку и критерий для эффекта A. Показать, что это обобщает **35.31** и что аналогично можно проверять любые взаимодействия.

(Иэйтс, 1934.)

35.7 Ниже в таблице приведены данные Брандта, использованные Иэйтсом (1938) для (8×2)-перекрестной классификации 533 сданных на убой свиней по полу и породе. В таблице даны клеточные частоты n_{ij} и суммарные по клетке значения изучаемой переменной (логарифм процентного содержания бекона в туше).

Порода \ Пол	Женский n_{i1}	Женский $\sum_p y_{i1p}$	Мужской n_{i2}	Мужской $\sum_p y_{i2p}$
I	83	66,55	89	181,04
II	51	98,69	141	281,43
III	13	25,90	17	34,20
IV	4	7,62	9	17,58
V	8	14,64	4	8,20
VI	15	28,11	32	64,42
VII	35	66,90	47	90,52
VIII	12	23,32	23	46,70
Всего	171	331,73	362	724,09

Полная СК, которую нельзя получить из этой таблицы, была 13,0142. Используя **35.31** и упражнение 35.5, показать, что (пренебрегая 1 ст. св. для генерального среднего) ДА имеет вид

Компонента дисперсии	Число ст. св.	СК	СрК
Обусловленная породой	7	0,6056	0,0865
Обусловленная полом	1	0,3032	0,3032
Взаимодействий пола и породы	8	1,0415	
	7	0,2300	0,0329
Полная между классами	15	1,2715	0,0848
Остаточная	517	11,7427	0,0227
Полная минус генеральное среднее	532	13,0142	

Показать, что как пол, так и порода влияют на содержание бекона, но что они не взаимодействуют.

35.8 Показать, что если в примере 35.6 в клетке имеется только одно наблюдение, то таблица (35.95) ДА остается верной с остаточной СК, тождественно равной нулю. Если считать, что в линейной модели нет взаимодействий второго порядка, то СК с $(r-1)(c-1)(l-1)$ ст. св. становится остаточной СК для проверки всех других параметров (см. упражнение 35.3).

35.9 Показать, что ДА для $(r \times c)$-перекрестной классификации с m наблюдениями в каждой клетке можно формально вывести из $(r \times c \times m)$-перекрестной классификации с ровно одним наблюдением в клетке, для которой классификация по слоям соответствует повторению опытов и соответствующие ей главные эффекты и все взаимодействия положены равными нулю.

35.10 Для определенной в **35.54** процедуры Неймана — Кейлса показать, что если для некоторого подмножества, включающего p из k групп, истинные средние равны, в то время как все другие группы имеют другие средние, то вероятность неверно признать неоднородной пару из упомянутого подмножества не превосходит α.

Показать, что если имеется m таких подмножеств с равными истинными средними, то вероятность неправильно признать неоднородной пару из любого такого подмножества не превосходит $m\alpha$.

(Хартли, 1955.)

35.11 Показать, что если в **35.57** k величин $\bar{x}_i - \theta_i$ имеют совместное многомерное нормальное распределение с дисперсиями σ^2/N и всеми ковариациями, равными $-\lambda^2\sigma^2/N$, а вектор x_0 не зависит от $\bar{x}_i$ и имеет нормальное распределение с нулевым средним и дисперсией $\lambda^2\sigma^2/N$, то величины $z_i = \bar{x}_i - \theta_i + x_0$ независимы и нормальны. Применяя к z_i метод **35.57**, показать, что вероятность одновременного выполнения неравенств

$$\bar{x}_i - \bar{x}_j - q_\alpha \{(1+\lambda^2)\, s^2/N\}^{1/2} \leqslant \theta_i - \theta_j \leqslant \bar{x}_i - \bar{x}_j + q_\alpha \{(1+\lambda^2)\, s^2/N\}^{1/2}$$

для всех $k(k-1)/2$ разностей $\theta_i - \theta_j$ равна $1-\alpha$. Показать, что точно так же обобщается результат **35.59**.

(Шеффе, 1959.)

35.12 В однофакторной классификации из примера 35.1 без ограничения общности положим $\sigma = 1$ и будем считать $\sum_{i=1}^{k} n_i\hat{\theta}_i$ началом координат. Показать, что значение $t^2 = \left(\sum_i c_i\hat{\theta}_i\right)^2 \Big/ \left(\sum_i c_i^2/n_i\right)$ достигает максимума при таком выборе c_i, что $c_i \propto n_i\hat{\theta}_i$, так что $\sum_i c_i = 0$, а $|t|$ является наибольшим возможным при данных наблюдениях значением отношения сравнения к соответствующей стандартной ошибке. Показать далее, что $t^2 = S_2$, где S_2 — определенная в (35.21) СК, стоящая в числителе составного критерия дисперсионного отношения, определенного в (35.22). Так что составной критерий, по существу, является критерием максимально возможного при данных наблюдениях сравнения.

(Этот результат является общим — см. Шеффе, 1959, и Габриэл, 1964.)

35.13 Вводя фиктивный параметр $\theta_0 \equiv 0$ с оценкой $\hat{\theta} \equiv 0$, показать, что всем линейным комбинациям $\psi = \sum c_i\theta_i$ (где $\sum_i c_i$ не обязательно равна нулю) можно сопоставить доверительные интервалы (35.113), увеличив q на единицу. Показать также, что аналогично можно использовать метод (35.108), увеличив k на единицу и заменив $\frac{1}{2}\sum_i |c_i|$ на

$$\max\left\{\sum_i c_i,\ c_i > 0;\ \left|\sum_i c_i\right|,\ c_i < 0\right\}.$$

(Тьюки, 1953; Шеффе, 1959.)

35.14 В **35.64** рассмотрим k событий, не являющихся независимыми, с равными вероятностями P_1 осуществления. Показать, что вероятность P_k того, что все они осуществятся, удовлетворяет неравенству $P_k \geqslant 1 - k(1-P_1)$. Пусть $P_1 = 1-\alpha$ — вероятность того, что стьюдентовская t-статистика (с ν ст. св.) лежит в интервале $(-t_\alpha, t_\alpha)$. Показать, что если m линейных комбинаций $\lambda_s = \sum_{i=1}^{k} c_{si}\theta_i$ $(s = 1, \ldots, m)$ оценить величинами $l_s = \sum_i c_{si}\hat{\theta}_i$, то $(l_s - \lambda_s)/\{\hat{D}(l_s)\}^{1/2}$ имеет для каждого s распределение Стьюдента с ν ст. св. Показать отсюда, что

$$P\{l_s - t_\alpha[\hat{D}(l_s)]^{1/2} \leqslant \lambda_s \leqslant l_s + t_\alpha[\hat{D}(l_s)]^{1/2}\} \geqslant 1-\alpha.$$

(Данн, 1961.)

35.15 Рассмотрим свободную от распределения статистику U для проверки k выборок против альтернатив об упорядочениях определенную в (31.151) т. 2, и конкурирующую статистику U', определяемую аналогично, но с заменой U_{pq} на

$$U'_{pq} = \sum_{i=1}^{n_p}\sum_{j=1}^{n_q}(x_{pi} - x_{qj}) \equiv n_p n_q(\bar{x}_p - \bar{x}_q)$$

так, что

$$U' = \sum_{p=1}^{k}\sum_{q=p+1}^{k} n_p n_q(\bar{x}_p - \bar{x}_q).$$

Для нормально распределенных наблюдений x_{is} со средними $\mathbf{M}(x_{is}) = \theta_i$ и равными дисперсиями σ^2 показать, что если $n_1 = \ldots = n_k = N$, то асимптотические функции мощности U и U' при проверке равенства θ_i против альтернативной гипотезы о том, что θ_i равноотстоящие, имеют вид

$$G\{\Delta(3/\pi)^{1/2} - \lambda_\alpha\} \quad \text{и} \quad G\{\Delta - \lambda_\alpha\},$$

где G — стандартная нормальная ф. р., $\Delta^2 = \sum_{i=1}^{k}(\theta_i - \bar{\theta})^2/\sigma^2$, а $G\{-\lambda_\alpha\} = \alpha$, как и в главе 25. Вывести, что АОЭ статистики U по сравнению с U' равна $3/\pi$. Показать, что последний результат справедлив при любом соотношении между упорядоченными θ_i.

(Бартоломью, 1961.)

35.16 Пусть в **35.54** R и P — любые подмножества из k групп такие, что R содержит P. Показать, что (35.21), вычисленная для R, не может быть меньше, чем та же СК, вычисленная для P.

35.17 Показать, что одновременная процедура проверки из **35.54** обладает тем свойством, что для любого однородного подмножества вероятность быть неправильно признанным неоднородным не превосходит α.

(Габриэл, 1964.)

35.18 В **35.54** определяется новая ступенчатая процедура, основанная на СК (35.21), но применяемая в духе последнего абзаца из **35.53**. Она имеет тот же самый общий размер α, что и одновременная процедура проверки, описанная в **35.54**. Показать, что для такой ступенчатой процедуры критическое значение для СК должно возрастать с ростом размера проверяемого подмножества, а, значит, множество может быть признано однородным в ступенчатом методе, только если оно признано таковым в одновременной процедуре.

(Габриэл, 1964.)

35.19 Используя неравенство Коши, показать в **35.62—3**, что

$$\max_{c_i}\left\{\sum_{i=1}^{k} c_i(\hat{\theta}_i - \theta_i)\right\}^2 = \sum_{i=1}^{k} c_i^2 \sum_{i=1}^{k}(\hat{\theta}_i - \theta_i)^2,$$

а, значит, квадрат разности между наблюденным значением сравнения и его математическим ожиданием имеет для того сравнения, для которого он максимален, то же самое распределение, что и $\left(\sum_i c_i^2\right) qs^2F$, где F — статистика критерия для проверки составной гипотезы о том, что все θ_i равны. Вывести отсюда (35.113).

(Это доказательство принадлежит Бэлцу и Ферхагену.)

ГЛАВА 36

ДРУГИЕ МОДЕЛИ ДИСПЕРСИОННОГО АНАЛИЗА

36.1 В предыдущей главе рассматривалось применение общей линейной модели к анализу наблюдений, расквалифицированных на группы. Основным в рассуждениях было предположение, явно записанное в линейной модели, о том, что принадлежность наблюдений к какой-либо группе влияет только на среднее значение этих наблюдений и никакого другого влияния не оказывает. Таким образом, в случае линейной модели дисперсионный анализ (ДА) можно описать как анализ *средних*, который проводится посредством вычисления некоторых сумм квадратов (СК) от наблюденных значений.

Примечательно, что к совершенно аналогичному подсчету СК (а в простейшем случае даже тождественному) приводит исследование ситуации совершенно другого типа. На начальной стадии изучения это подобие в проведении анализа даже затмевало существенное различие между лежащими в основе математическими моделями. Это различие впервые было явно указано Эйзенхартом (1947). При этом ДА, основанный на методе НК для общей линейной модели, изложенный в главе 35, назывался моделью I ДА. Это название будет употребляться и в дальнейшем. Перейдем теперь к исследованию другой строго определенной математической модели, которую будем называть модель II ДА.

Читателю должно быть понятно, что термин «дисперсионный анализ», применявшийся раньше к модели I (см. **35.4**, **35.9—10**), теперь нужно понимать в более широком смысле. Вообще, определим ДА как анализ данных, вне зависимости от того расквалифицированы они или нет (см. **35.15**), проводимый с помощью разбиения СК на компоненты, соответствующие различным факторам, которые могут действовать в комбинации или каждый в отдельности.

Модель II: компоненты дисперсии

36.2 Рассмотрим теперь вместо общей линейной модели (19.8) следующую внешне похожую на нее модель

$$\underset{(n\times 1)}{\boldsymbol{y}} = \underset{(n\times 1)}{\mathbf{1}\theta} + \underset{(n\times p)}{\boldsymbol{X}}\ \underset{(p\times 1)}{\boldsymbol{u}} + \underset{(n\times 1)}{\boldsymbol{\varepsilon}} \tag{36.1}$$

Так же, как и в главе 35 (и ранее в упражнении 19.1 т. 2), в (36.1) мы выделили «генеральное среднее» θ, которое теперь пишется без индекса. Как и ранее, $\mathbf{1}$ обозначает вектор, компоненты которого равны единице, а $\boldsymbol{X}$ — заданную постоянную матрицу, в то время как $\boldsymbol{\varepsilon}$ — вектор, соответствующий ошибкам при наблюдении. Важнейшим изменением является замена вектора параметров θ из (19.8) на вектор $\boldsymbol{u}$, компонентами которого являются p *случайных величин*. Таким образом, в (36.1) утверждается, что значение y_i $(i = 1, \ldots, n)$ складывается из генерального среднего θ плюс линейная комбинация p случайных величин u_j плюс член ε_i, соответствующий ошибке. При этом y_i имеет теперь уже не одну случайную компоненту, как это было в (19.8), а $(p+1)$.

Так же, как и в (19.9—10) для общей линейной модели, предполагается, что

$$\mathsf{M}(\boldsymbol{\varepsilon}) = \mathbf{0}; \quad \boldsymbol{V}(\boldsymbol{\varepsilon}) = \mathsf{M}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma_\varepsilon^2 \boldsymbol{I}. \tag{36.2}$$

При этом дисперсию ошибки мы обозначили σ_ε^2, добавив соответствующий индекс для того, чтобы было отличие от дисперсий наших новых случайных величин. Далее, поскольку мы выделили генеральное среднее, то будем предполагать, что

$$\mathsf{M}(\boldsymbol{u}) = \mathbf{0}, \tag{36.3}$$

и что

$$\begin{gathered} \mathsf{D}(u_j) = \sigma_j^2; \quad \operatorname{cov}(u_j, u_l) = 0, \quad l \neq j; \\ \operatorname{cov}(u_j, \varepsilon_i) = 0 \quad \text{для всех} \quad i, j. \end{gathered} \tag{36.4}$$

Таким образом, все наши случайные величины имеют нулевое математическое ожидание и попарно не коррелированы.

36.3 Параметрами, представляющими интерес в модели (36.1) являются σ_j^2, дисперсии величин u_j и дисперсия ошибок σ_ε^2. Размерность задачи может быть уменьшена в том случае, если заранее известно, что некоторые из σ_j^2 могут быть равными. Предположим, что имеется k *различных* дисперсий σ_j^2, где $k \leqslant p$, и перепишем (36.1) в виде

$$\boldsymbol{y} = \mathbf{1}\theta + \sum_{j=1}^{k} \boldsymbol{X}_j \boldsymbol{u}_j + \boldsymbol{\varepsilon}, \tag{36.5}$$

где $\boldsymbol{u}_j$ являются теперь векторами (подвекторы вектора $\boldsymbol{u}$) с p_j компонентами $\left(\sum_{j=1}^{k} p_j = p\right)$, которые в свою очередь являются некоррелированными случайными величинами с нулевыми математическими ожиданиями и дисперсиями σ_j^2. Матрица $\boldsymbol{X}_j$ представляет собой $(n \times p_j)$ подматрицу матрицы $\boldsymbol{X}$.

Для y из (36.5) пусть W обозначает матрицу $\mathsf{M}(yy')$, а V_y — дисперсионную матрицу, которая предполагается невырожденной. Тогда, используя (36.2—4), получим

$$V_y = W - \theta^2 11' = \sum_{j=1}^{k} \sigma_j^2 X_j X_j' + \sigma_\varepsilon^2 I. \qquad (36.6)$$

Таким образом, по существу, мы желаем оценить параметры, которые являются коэффициентами у $k+1$ компоненты дисперсионной матрицы наблюдений. В соответствии с этим (36.5) часто называют моделью ДА для *компонент дисперсии*, другое название для «модели II».

Соотношение (36.6) указывает на существенную разницу между (36.5) и общей линейной моделью. Наблюдения теперь не являются некоррелированными (V_y не диагональна), поскольку компоненты этих наблюдений являются линейными функциями от одних и тех же случайных величин u_j.

Пример 36.1

Рассмотрим модель (36.5) с $k=1$, $(n \times p)$-матрица X_1 является матрицей вида X из примера 35.1, и u_1 представляет собой $(p \times 1)$-вектор. В результате получается задача однофакторной классификации для изучаемой модели, в которой наблюдения представлены в виде p групп, причем q-е наблюдение в i-й группе равно

$$y_{iq} = \theta + u_i + \varepsilon_{iq},$$

где величины u_i (p компонент вектора u_1) и ε_{iq} не коррелированы, имеют нулевые математические ожидания и дисперсии

$$\mathsf{D}(u_i) = \sigma_1^2 \quad \text{для всех} \quad i,$$
$$\mathsf{D}(\varepsilon_{iq}) = \sigma_\varepsilon^2 \quad \text{для всех} \quad i,\ q.$$

Отсюда немедленно получаем, что

$$\mathsf{D}(y_{iq}) = \sigma_1^2 + \sigma_\varepsilon^2,$$

так что участвующие в модели два параметра в буквальном смысле являются компонентами дисперсии наблюдений. Отсюда и возник термин «компоненты дисперсии», который мы отнесли к общей модели (36.5).

36.4 Исследование свойств модели (36.5) нужно начинать с самого начала, ибо метод НК, используемый в общей линейной модели, теперь неприменим. Наше изложение следует Грейбиллу и Халтквисту (1961).

Соотношение (36.5) можно более симметрично переписать в виде

$$y = \sum_{j=0}^{k+1} X_j u_j, \qquad (36.7)$$

где мы положили

$$X_0 \equiv 1, \quad u_0 \equiv \theta \text{ (скаляр)}, \quad X_{k+1} \equiv I, \quad u_{k+1} = \varepsilon.$$

В (36.7) фигурируют $k+2$ параметра, а именно θ и $k+1$ дисперсий, участвующих в формулировке модели. Предположения (36.2—4) переписываются в виде

$$\begin{aligned} &\mathsf{M}(u_j) = 0, \quad V(u_j) = \mathsf{M}(u_j u_j') = \sigma_j^2 I, \\ &\mathsf{M}(u_j u_i') = 0, \quad i \neq j; \quad i,\ j = 1, \ldots, k+1, \end{aligned} \qquad (36.8)$$

где $\sigma_{k+1}^2 \equiv \sigma_\varepsilon^2$. Формально можно записать

$$\sigma_0^2 = \mathsf{M}(u_0 u_0') = \theta^2, \qquad (36.9)$$

так что параметрами теперь являются σ_j^2 ($j = 0, 1, \ldots, k+1$). Соотношение (36.6) можно теперь переписать в виде

$$W = \mathsf{M}(yy') = \sum_{j=0}^{k+1} \sigma_j^2 A_j, \qquad (36.10)$$

где $A_j = X_j X_j'$, а

$$V_y = \sum_{j=1}^{k+1} \sigma_j^2 A_j. \qquad (36.11)$$

Несмещенные квадратичные оценки параметров

36.5 При изучении общей линейной модели в (19.19) были найдены условия, которые должны выполняться, чтобы линейная функция наблюдений могла служить несмещенной оценкой линейной функции параметров. В модели для компонент дисперсии (36.7) параметры являются коэффициентами в разложении (36.10) матрицы W, являющейся матрицей вторых моментов от наблюдений. Естественно пытаться искать квадратичные оценки этих параметров. Ниже мы показываем, что необходимым и достаточным условием того, что σ_s^2 допускают несмещенные оценки квадратичного вида $y'C_s y$, является линейная независимость матриц A_s.

36.6 Предположим сперва, что существуют матрицы C_s такие, что

$$\mathsf{M}(y'C_s y) = \sigma_s^2, \qquad s = 1, \ldots, k+1. \qquad (36.12)$$

Используя (36,7—8), отсюда получаем, что

$$\sigma_s^2 = \mathsf{M}\left\{\left(\sum_j X_j u_j\right)' C_s \left(\sum_j X_j u_j\right)\right\} = \mathsf{M}\left\{\sum_j u_j' X_j' C_s X_j u_j\right\}. \qquad (36.13)$$

Как и в (19.38),

$$\mathsf{M}(u_j' B u_j) = \sigma_j^2 \operatorname{tr} B, \qquad j = 1, \ldots, k+1, \qquad (36.14)$$

и поскольку при $j = 0$ это тривиально выполняется, то (36.13) принимает вид

$$\sigma_s^2 = \sum_{j=0}^{k+1} \sigma_j^2 \operatorname{tr} X_j' C_s X_j = \sum_j \sigma_j^2 \operatorname{tr}(X_j X_j' C_s) = \sum_j \sigma_j^2 \operatorname{tr}(A_j C_s). \qquad (36.15)$$

Приравнивая коэффициенты при σ_j^2 в (36.15), получаем

$$\begin{aligned} &\operatorname{tr}(A_s C_s) = 1, \\ &\operatorname{tr}(A_j C_s) = 0, \qquad j \neq s. \end{aligned} \qquad (36.16)$$

Пусть существуют константы $l_0, \ldots, l_{k+1}$ такие, что

$$\sum_{j=0}^{k+1} l_j A_j = 0. \qquad (36.17)$$

Если бы не все l_j равнялись нулю, то это означало бы линейную зависимость A_j. Но из (36.16—17) следует, что

$$l_s = \sum_j l_j \operatorname{tr}(A_j C_s) = \operatorname{tr}\left(C_s \sum_j l_j A_j\right) = 0.$$

Таким образом, мы видим, что (36.17) может выполняться только в случае, когда все l_j равны нулю, т. е. из (36.12) следует линейная независимость матриц A_s.

36.7 Для доказательства обратного утверждения представим $n(n+1)/2$ элементов, расположенных на и над главной диагональю матриц yy' и A_j в (36.10) в виде векторов w^*, a_j^* ($j = 0, 1, \ldots, \ldots, k+1$) с одним и тем же (произвольным) порядком следования компонент. Тогда из (36.10) следует

$$\mathsf{M}(w^*) = \sum_j \sigma_j^2 a_j^*. \qquad (36.18)$$

Если через A^* обозначить $[n(n+1)/2 \times (k+2)]$-матрицу, столбцы которой совпадают с векторами a_j^*, а через σ^2 — вектор с компонентами σ_j^2, то (36.18) примет вид

$$\mathsf{M}(w^*) = A^* \sigma^2. \qquad (36.19)$$

Поскольку теперь A_j предполагаются линейно независимыми, то векторы a_j^*, образованные из элементов этих матриц, тоже будут линейно независимыми, а значит, матрица A^*, образованная из a_j^*, имеет ранг $k+2$ и можно указать $k+2$ линейно независимые строки, которые образуют невырожденную подматрицу A^{**}. Пусть w^{**} — соответствующий подвектор вектора w^*. Из (36.19) следует

$$\mathsf{M}(w^{**}) = A^{**} \sigma^2,$$

откуда

$$(A^{**})^{-1} \mathsf{M}(w^{**}) = \sigma^2,$$

или

$$\mathsf{M}\{(A^{**})^{-1} w^{**}\} = \sigma^2. \qquad (36.20)$$

Тем самым получена несмещенная квадратичная оценка вектора параметров.

Всюду в дальнейшем будет предполагаться, что матрицы A_s линейно независимы.

Достаточные статистики в нормальном коммутативном случае

36.8 До сих пор не делалось никакого предположения о виде распределения случайных величин u_j из нашей модели. Рассмотрим теперь случай, когда каждый вектор u_j ($j = 1, \ldots, k+1$) имеет многомерное нормальное распределение. Отсюда и из предположений о ковариациях (36.8) и обращения в нуль средних следует, что p случайных величин u_j ($j = 1, \ldots, p$) являются независимыми с нормальным распределением и нулевым математическим ожиданием.

Одно лишь предположение нормальности не дает нам возможности проводить дальнейшее исследование. Чтобы получить существенное продвижение, нужно наложить дополнительное предположение о коммутативности

$$A_j A_i = A_i A_j, \qquad i, j = 0, 1, \ldots, k+1. \qquad (36.21)$$

Поскольку A_j симметричны, то всегда справедливо

$$A_j A_i = A_j' A_i' = (A_i A_j)'.$$

Таким образом, (36.21) сводится к требованию о том, что наряду с A_j матрицы $A_i A_j$ также являются симметричными. В примере 36.2 показывается, что это требование в самом деле ограничительно.

Пример 36.2

Рассмотрим снова ситуацию однофакторного анализа примера 36.1 с $k = 1$. Как всегда $A_0 = 11'$ и из примера 35.1 имеем

$$A_1 = \begin{pmatrix} (11')_{n_1} & & & 0 \\ & (11')_{n_2} & & \\ & & \ddots & \\ 0 & & & (11')_{n_p} \end{pmatrix},$$

где индекс указывает число строк и столбцов в подматрице. Простым умножением убеждаемся, что все элементы первых n_1 столбцов матрицы A_0A_1 равны n_1, все элементы следующих n_2 столбцов равны n_2, и, продолжая далее, получим в конце, что все элементы последних n_p столбцов равны n_p. Таким образом, A_0A_1 симметрична (A_0 и A_1 коммутируют) только в случае, когда все n_i равны. Поскольку $X_2 = I$, то также и $A_2 = I$ и всегда коммутируют. Таким образом, рассматриваемая модель включает в себя однофакторный анализ только в сбалансированном случае, когда числа наблюдений во всех p группах одинаковы. Это существенное отличие от примера 35.1 для модели I, где числа наблюдений в группах не имели никакого значения.

36.9 Из примера 36.2 следует, что только для сбалансированного случая можно ожидать выполнения (36.21). Продолжим наши рассмотрения, помня об этом ограничении.

Из нормальности и независимости p случайных величин u_j ($j = 1, \ldots, p$) следует, что (коррелированные) величины $y_1, \ldots, y_n$ имеют многомерное нормальное распределение с

$$\mathrm{M}(y) = X_0u_0 = 1\theta$$

и дисперсионной матрицей V_y, задаваемой (36.11). Квадратичная форма в показателе экспоненты для плотности их совместного нормального распределения равна, таким образом,

$$Q = (y - 1\theta)' V_y^{-1} (y - 1\theta), \tag{36.22}$$

и, согласно **15.10**, имеет распределение типа χ^2 с n степенями свободы.

36.10 Далее в силу условия коммутативности (36.31), существует ортогональная матрица P, которая одновременно диагонализирует все A_j, так что

$$PA_jP' = D_j, \tag{36.23}$$

где D_j — диагональные матрицы. Более того, P можно выбрать так, чтобы все элементы одной из ее строк (скажем, первой, которую мы обозначим P_1) равнялись $n^{-1/2}$ и чтобы

$$P_1 1 = n^{1/2}; \quad P_j 1 = 0, \quad j \neq 1, \tag{36.24}$$

где P_j — любое множество строк матрицы P, не содержащее P_1.

Из (36.10—11) следует, что P диагонализирует также W и V_y, так что на главных диагоналях матриц PWP' и PV_yP' стоят соответственно характеристические числа W и V_y. Из (36.24) сразу следует, что эти два множества характеристических чисел (которые все положительны) совпадают, за исключением первого элемента. Если характеристическими числами W являются λ_j ($j = 1, \ldots, n$), то характеристическими числами V_y будут $\lambda_1^* = \lambda_1 - n\theta^2$ и λ_j при $j > 1$. Эти характеристические числа являются, конечно, функциями параметров σ_j^2.

Если s — число различных характеристических чисел матрицы W, а s^* — матрицы V_y, то либо $s^* = s - 1$, если λ_1 однократно, а λ_1^* нет; либо $s^* = s + 1$, если однократно λ_1^*, а λ_1 нет; либо $s^* = s$, если λ_1 и λ_1^* одновременно являются или не являются однократными характеристическими числами. Грейбилл и Халтквист (1961) показали, что $s \geqslant k + 2$ и что при дополнительных условиях $s^* \leqslant 1 + \mathrm{rank}(X_0 \vdots X_1 \vdots \ldots \vdots X_k)$ *).

36.11 В наборе чисел $\lambda_2, \ldots, \lambda_n$ обозначим $\lambda_2^*, \ldots, \lambda_t^*$ все различные значения, а $n_2, \ldots, n_t$ — их кратности. При этом, очевидно, $t = s$ или $t = s + 1$. Поскольку $PP' = I$, то (36.22) эквивалентно

$$Q = (Py - P1\theta)' (PV_yP')^{-1} (Py - P1\theta). \tag{36.25}$$

Разобьем теперь P на $P_1, \ldots, P_t$, где, как и раньше, P_1 — первая строчка матрицы P, а P_j ($j > 1$) являются матрицами порядка $(n_j \times n)$. Используя (36.24), соотношение (36.25) можно теперь переписать в виде

$$Q = \begin{pmatrix} P_1y - n^{1/2}\theta \\ P_2y \\ \vdots \\ P_ty \end{pmatrix}' \begin{pmatrix} 1/\lambda_1^* & & \ldots & 0 \\ & I_2/\lambda_2^* & & \\ & & \ddots & \\ 0 & & \ldots & I_t/\lambda_t^* \end{pmatrix} \begin{pmatrix} P_1y - n^{1/2}\theta \\ P_2y \\ \vdots \\ P_ty \end{pmatrix}, \tag{36.26}$$

где I_j — единичная матрица порядка n_j.

Из (36.26) сразу следует, что t статистик

$$P_1y, \quad y'P_j'P_jy, \quad j = 2, \ldots, t,$$

достаточны для $k + 2$ параметров модели, поскольку никакие другие статистики не входят в функцию правдоподобия. Интуитивно понятно, что они образуют *минимальное* множество доста-

*) Здесь и в дальнейшем под $(X \vdots Y \vdots Z \vdots \ldots)$ автор подразумевает матрицу, составленную с помощью последовательного присоединения к столбцам матрицы X столбцов матриц Y, Z и т. д. (*Прим. ред.*).

точных статистик. Если λ_1^* не совпадает ни с каким λ_j, то доказательство этого факта проводится непосредственно методом **23.18**. В противном случае доказательство проведено Грейбиллом и Халтквистом (1961), которые используют также приводимый ниже в **36.16** результат Гаучи (1959) для того, чтобы показать, что если $t = s = k + 2$ (в соответствии с **36.10** — минимально возможное значение), то множество минимальных достаточных статистик полно.

36.12 Из (36.26) непосредственно следует, что статистика $\boldsymbol{P}_1\boldsymbol{y}$ имеет одномерное нормальное распределение со средним $n^{1/2}\theta$ и дисперсией $\lambda_1^* = \boldsymbol{P}_1\boldsymbol{V}_y\boldsymbol{P}_1'$. Далее, каждая из оставшихся $t-1$ компонент минимального достаточного множества представляет собой квадратную форму $Q_j = \boldsymbol{y}'\lambda_j^{-1}\boldsymbol{P}_j'\boldsymbol{P}_j\boldsymbol{y}$, $j = 2, \ldots, t$, от случайных величин, имеющих многомерное нормальное распределение. Каждая матрица $\lambda_j^{-1}\boldsymbol{P}_j'\boldsymbol{P}_j\boldsymbol{V}_y$ идемпотентна, поскольку $\boldsymbol{P}_j\boldsymbol{V}_y\boldsymbol{P}_j' = \lambda_j\boldsymbol{I}_j$. Более того, поскольку $\boldsymbol{P}_j\boldsymbol{V}_y\boldsymbol{P}_l' = 0$ $(j \neq l)$, то сумма Q_j

$$Q^* = \sum_{j=2}^{t} \boldsymbol{y}'\lambda_j^{-1}\boldsymbol{P}_j'\boldsymbol{P}_j\boldsymbol{y}$$

также обладает тем свойством, что соответствующая ей матрица

$$\sum_{j=2}^{t} \lambda_j^{-1}\boldsymbol{P}_j'\boldsymbol{P}_j\boldsymbol{V}_y$$

идемпотентна. Таким образом, равенство $Q^* = \sum_{j=2}^{t} Q_j$ представляет собой разложение того же типа, что и обсуждавшееся в **35.7**. Поэтому из доказанного сейчас свойства идемпотентности следует, что Q_j $(j = 2, \ldots, t)$ имеют χ^2-распределения с n_j ст. св., причем, поскольку, согласно (36.24), параметры нецентральности равны $(\mathbf{1}\theta)'\lambda_j^{-1}\boldsymbol{P}_j'\boldsymbol{P}_j(\mathbf{1}\theta) = 0$, то соответствующие распределения центральны. Поскольку полученный выше результат о $\boldsymbol{P}_1\boldsymbol{y}$ эквивалентен тому, что квадратичная форма $Q_1 = \boldsymbol{y}'(\lambda_1^*)^{-1}\boldsymbol{P}_1'\boldsymbol{P}_1\boldsymbol{y}$ имеет нецентральное χ^2-распределение с одной ст. св. и параметром нецентральности $n\theta^2$, то окончательно получаем, что сумма рангов t квадратичных форм Q_j равна n, т. е. рангу их суммы Q из (36.26). В соответствии с **35.7** получаем, что квадратичные формы не зависят друг от друга.

Дисперсионный анализ в модели II

36.13 Отметим теперь важное различие между моделью I и моделью II. В последней СК $\boldsymbol{y}'\boldsymbol{y}$ не может быть представлена в виде суммы квадратичных форм, которые *сами* независимы и

имеют (нецентральные или центральные) χ^2-распределения, поскольку, учитывая зависимость между наблюдениями, распределение самого $\boldsymbol{y}'\boldsymbol{y}$ не является таковым. Вместо $\boldsymbol{y}'\boldsymbol{y}$ соответствующим свойством обладает (36.22). Однако **36.11—12** показывают, что t квадратичных форм Q_j имеют соответствующие распределения. Матрицы этих форм равны

$$\boldsymbol{P}_1'\boldsymbol{P}_1/\lambda_1^*, \qquad \boldsymbol{P}_j'\boldsymbol{P}_j/\lambda_j, \qquad j = 2, \ldots, t,$$

и поскольку $\sum_{j=1}^{t} \boldsymbol{P}_j'\boldsymbol{P}_j = \boldsymbol{P}'\boldsymbol{P} = \boldsymbol{I}$, то мы видим, что $\boldsymbol{y}'\boldsymbol{y}$ можно представить в виде суммы t квадратичных форм таких, что после деления на соответствующие различные характеристические числа матрицы $\boldsymbol{V}_y$ они становятся независимыми, имеющими χ^2-распределения случайными величинами. Более того, число ст. св. равно просто кратности соответствующего характеристического значения, причем все формы, за исключением Q_1, имеют центральное распределение.

36.14 Поскольку

$$\mathsf{M}(Q_j) = n_j, \qquad j \geqslant 2,$$

то сразу получаем

$$\mathsf{M}(\boldsymbol{y}'\boldsymbol{P}_j'\boldsymbol{P}_j\boldsymbol{y}/n_j) = \lambda_j, \qquad j \geqslant 2. \tag{36.27}$$

Таким образом, характеристические числа могут быть найдены как математические ожидания соответствующих средних квадратов (СрК) в таблице ДА. Поскольку

$$\mathsf{M}(\boldsymbol{P}_1\boldsymbol{y}) = n^{1/2}\theta,$$

то, как и понятно, θ оценивается с помощью среднего значения от всех наблюдений. Нам нужно оценить оставшиеся $k+1$ параметров, являющихся дисперсиями. Если λ_j представляют собой $k+1$ различных функций этих параметров, то (36.27) можно разрешить относительно этих параметров для получения соответствующих оценок.

В простом случае, когда λ_j являются *линейными* функциями параметров, (36.27) особенно просто разрешается для получения несмещенных оценок величин σ_j^2.

Грейбилл и Халтквист (1961) показали, что ДА в определенном выше расширенном смысле существует с характеристическими числами, являющимися произвольными различными функциями параметров тогда и только тогда, когда выполняется условие коммутативности (36.21), а матрица $\boldsymbol{W} = \mathsf{M}(\boldsymbol{y}\boldsymbol{y}')$ имеет $s = k + 2$ (согласно **36.10** — минимально возможное число) различных характеристических чисел. При предположении многомерной нормальности множество достаточных статистик является тогда полным (см. **36.11**) и, согласно **17.35** т. 1, оценки

являются единственными несмещенными МД-оценками соответствующих математических ожиданий. Грейбилл и Халтквист (1961) показали, что СрК в (36.27) остается несмещенной квадратичной МД-оценкой своего математического ожидания и при более слабых предположениях, чем многомерная нормальность.

36.15 Теперь мы собираемся связать изучение модели II с моделью I ДА, которая исследовалась в главе 35. Любая таблица ДА для модели I соответствует разложению СК $y'y$ на компоненты, являющиеся СК (одна из которых соответствует генеральному среднему). При этом сумма чисел ст. св. в таблице (рангов квадратичных форм) всегда равна n, где n — полное число наблюдений. Если выполняется (36.21), то (см. **36.13**) в модели II справедливо то же самое разложение, поскольку ранги остаются аддитивными. Но теперь в основе лежит (36.22), разложенное в соответствии с (36.26). При этом отношение каждой СК (за исключением генерального среднего) к математическому ожиданию соответствующего СрК имеет χ^2-распределение, которое теперь всегда центрально.

Таким образом, каждой модели I ДА (сбалансированной, чтобы выполнялось (36.21)), соответствует сбалансированная модель II ДА.

Пример 36.3 Сбалансированная однофакторная классификация

В соответствии с **36.15** для сбалансированной однофакторной классификации, рассмотренной в примерах 36.1—2, таблица (35.24) ДА модели I остается пригодной и для модели II. Нам нужны теперь математические ожидания СрК из таблицы, за исключением генерального среднего (которое рассматривается в примере 36.5).

Совершенно очевидно, что для остаточной СК, которую мы теперь обозначим S_3, справедливо тождество

$$S_3 = \sum_i \sum_q (y_{iq} - y_{i.})^2 \equiv \sum_i \sum_q (\varepsilon_{iq} - \varepsilon_{i.})^2,$$

а, значит, она имеет то же самое распределение, что и в модели I, поскольку не зависит от θ и u_i. Таким образом, математическое ожидание соответствующего СрК равно σ_ε^2, которое мы обозначим $\lambda_3 = \sigma_\varepsilon^2$. Так же как и в (35.21), СК между групп обозначим

$$S_2 = \sum_i n_i (y_{i.} - y_{..})^2 \equiv n_i \sum_i \{(u_i + \varepsilon_{i.}) - (u_. + \varepsilon_{..})\}^2;$$

здесь n_i равны между собой, поскольку теперь мы имеем n наблюдений, разбитых на p групп *) равного размера, и $n_i = n/p$.

*) В примере 35.1 мы писали k, а теперь пишем p, поскольку в данной главе k имеет совсем другой смысл.

Поскольку величины u и величины ε независимы, а $\varepsilon_{i.}$ является средним значением n_i членов, соответствующих ошибкам, то

$$\mathbf{D}(u_i + \varepsilon_{i.}) = \sigma_1^2 + (\sigma_\varepsilon^2/n_i).$$

Поскольку $\mathbf{M}(u_i + \varepsilon_{i.}) = 0$, то отсюда непосредственно следует, что случайная величина

$$\sum_{i=1}^{p} \{(u_i + \varepsilon_{i.}) - (u_. + \varepsilon_{..})\}^2 / (\sigma_1^2 + (\sigma_\varepsilon^2/n_i))$$

имеет χ^2-распределение с $p-1$ ст. св., поскольку она является приведенной суммой квадратов относительно выборочного среднего. Таким образом, математическое ожидание величины S_2 (СК между групп) равно $(p-1)\,n_i(\sigma_1^2 + \sigma_\varepsilon^2/n_i)$, а значит, для СрК имеем $\mathbf{M}\{S_2/(p-1)\} = \sigma_\varepsilon^2 + n_i\sigma_1^2$, что мы обозначим $\lambda_2 = \sigma_\varepsilon^2 + n_i\sigma_1^2$.

В результате мы получили две независимые, имеющие χ^2-распределение случайные величины

$$S_2/\lambda_2 = \sum_i n_i (y_{i.} - y_{..})^2 / (\sigma_\varepsilon^2 + n_i\sigma_1^2),$$

$$S_3/\lambda_3 = \sum_i \sum_q (y_{iq} - y_{i.})^2 / \sigma_\varepsilon^2,$$

отношение которых после деления на число ст. св. ($(p-1)$ и $(n-p)$ соответственно) имеет F-распределение. Если $\sigma_1^2 = 0$, то знаменатели совпадают. Поэтому для проверки гипотезы $\sigma_1^2 = 0$, которая в модели II соответствует отсутствию разницы между группами, можно использовать ту же самую F-статистику (35.22), что и в примере 35.1.

Таким образом, в двух разных моделях для проверки одной и той же гипотезы можно использовать одну и ту же статистику. Но следует обратить внимание на два момента. Во-первых, мы ничего не знаем о том, является ли это *оптимальным* критерием для модели II. В модели I оптимальность следовала из теории общей линейной гипотезы. Во-вторых, хотя в обеих моделях статистика критерия одна и та же, ее распределение совпадает *только* при справедливости гипотезы об отсутствии разницы между группами. Функции мощности по необходимости отличаются в двух моделях, ибо альтернативные гипотезы совершенно разные (см. упражнение 36.1).

Из выражений для $\mathbf{M}(S_2)$ и $\mathbf{M}(S_3)$ следует, что

$$\mathbf{M}\left\{\left(\frac{S_2}{p-1} - \frac{S_3}{n-p}\right)\Big/ n_i\right\} = \sigma_1^2.$$

Отсюда имеем несмещенную оценку для σ_1^2, которая, очевидно, может принимать отрицательные значения. Будучи функцией

полного множества достаточных статистик ($\bar{y}$, S_2, S_3) (см. **36.11**), эта оценка является несмещенной МД-оценкой. Если провести усечение в нуле в том случае, когда НК-оценка величины λ_2 меньше, чем НК-оценка величины λ_3 (как в упражнении 36.5), то получится смещение, но среднеквадратичная ошибка становится меньше. Возможны и дальнейшие улучшения (см. Уэнг (1967)).

Пример 36.4. Сбалансированная двухфакторная перекрестная классификация

В соответствии с **36.15** таблица (35.65) и упражнение 35.1 для сбалансированной двухфакторной перекрестной классификации модели I ДА сохраняется и в случае модели II, за исключением последнего столбца, который относится специально к модели I. В модели (36.7) имеем теперь $k=3$. Для удобства обозначим три оцениваемые дисперсии через σ_R^2, σ_C^2 и σ_{RC}^2, подчеркивая тем самым, что они соответствуют дисперсиям строк (row), столбцов (column) и взаимодействий (между строчками и столбцами) для величин u. Как и в примере 36.2, легко видеть, что условие коммутативности (36.21) выполняется. Это просто следует из учета симметрии.

Оставляя пока в стороне генеральное среднее, нужно получить оценки математических ожиданий четырех СрК. Как и в примере 36.3, рассмотрение модели, записанной теперь в понятных обозначениях

$$y_{ijp}=\theta+u_{i*}+u_{*j}+u_{ij}+\varepsilon_{ijp}, \tag{36.28}$$

$$(i=1,\ldots,r;\ j=1,\ldots,c;\ p=1,\ldots,m)$$

показывает, что остаточная СК, обозначаемая теперь $S_5=\sum_i\sum_j\sum_p(y_{ijp}-y_{ij.})^2$, тождественно совпадает с $\sum_i\sum_j\sum_p(\varepsilon_{ijp}-\varepsilon_{ij.})^2$ и имеет то же самое распределение, что и в модели I. Так что соответствующий СрК имеет математическое ожидание σ_ε^2, свойство, очевидно, справедливое вообще для сбалансированной модели II.

СК строк записывается теперь в виде

$$S_2=cm\sum_i(y_{i..}-y_{...})^2\equiv$$

$$\equiv cm\sum\{(u_{i*}+u_{i.}+\varepsilon_{i..})-(u_{.*}+u_{..}+\varepsilon_{...})\}^2$$

и, как в примере 36.3, $\mathbf{D}(u_{i*}+u_{i.}+\varepsilon_{i..})=\sigma_R^2+\sigma_{RC}^2/c+\sigma_\varepsilon^2/(cm)$.

Отсюда математическое ожидание СрК строк равно

$$\mathbf{M}\{S_2/(r-1)\}=\sigma_\varepsilon^2+m\sigma_{RC}^2+cm\sigma_R^2. \tag{36.29}$$

Меняя символы строк и столбцов, получим аналогично для СрК столбцов

$$\mathbf{M}\{S_3/(c-1)\}=\sigma_\varepsilon^2+m\sigma_{RC}^2+rm\sigma_C^2. \tag{36.30}$$

СК взаимодействий записывается в виде

$$S_4=m\sum_i\sum_j(y_{ij.}-y_{i..}-y_{.j.}+y_{...})^2\equiv$$

$$\equiv m\sum\sum\{(u_{ij}+\varepsilon_{ij.})-(u_{i.}+\varepsilon_{i..})-(u_{.j}+\varepsilon_{.j.})+(u_{..}+\varepsilon_{...})\}^2. \tag{36.31}$$

Выражение в фигурных скобках в правой части (36.31) имеет нулевое математическое ожидание, поэтому дисперсия совпадает с математическим ожиданием квадрата этого выражения. Далее, как нетрудно проверить, вычисляя необходимые четыре дисперсии и шесть ковариаций, имеем

$$\mathbf{D}(u_{ij}-u_{i.}-u_{.j}+u_{..})=\sigma_{RC}^2\left(1-\frac{1}{c}\right)\left(1-\frac{1}{r}\right).$$

Аналогично

$$\mathbf{D}(\varepsilon_{ij.}-\varepsilon_{i..}-\varepsilon_{.j.}+\varepsilon_{...})=\frac{\sigma_\varepsilon^2}{m}\left(1-\frac{1}{c}\right)\left(1-\frac{1}{r}\right),$$

и отсюда математическое ожидание СК равно

$$\mathbf{M}(S_4)=mrc\left(\sigma_{RC}^2+\frac{\sigma_\varepsilon^2}{m}\right)\left(1-\frac{1}{c}\right)\left(1-\frac{1}{r}\right),$$

Деля на число ст. св. $(r-1)$ $(c-1)$ и сокращая, получим для СрК

$$\mathbf{M}\{S_4/(r-1)(c-1)\}=\sigma_\varepsilon^2+m\sigma_{RC}^2. \tag{36.32}$$

Таким образом, вместо последнего столбца в таблице (35.65) мы получили столбец математических ожиданий СрК:

$$\left.\begin{array}{ll} & \mathbf{M}(\text{СрК}) \\ \text{Строчек:} & \lambda_2=\sigma_\varepsilon^2+m\sigma_{RC}^2+cm\sigma_R^2, \\ \text{Столбцов:} & \lambda_3=\sigma_\varepsilon^2+m\sigma_{RC}^2+rm\sigma_C^2, \\ \text{Взаимодействий:} & \lambda_4=\sigma_\varepsilon^2+m\sigma_{RC}^2, \\ \text{Остаточный:} & \lambda_5=\sigma_\varepsilon^2. \end{array}\right\} \tag{36.33}$$

Разрешая (36.33), сразу получаем несмещенные оценки дисперсий, являющихся параметрами модели. За исключением $\hat{\sigma}_\varepsilon^2$, любая из этих оценок может принимать и отрицательные значения, являясь разностью соответствующих оценок величин λ. Ситуация аналогична концу примера 36.3, и сделанные там замечания применимы к изучаемому случаю.

В (36.33) указывается, какие функции параметров нужно взять в качестве делителей, чтобы из СК в таблице ДА получить величины, имеющие центральное χ^2-распределение. Из их рассмотрения следует, что удачное совпадение критериев для модели I и модели II, которое имело место при однофакторной классификации в примере 36.3, теперь не сохраняется. Если мы хотим проверить гипотезу $\sigma_R^2 = 0$ (или $\sigma_C^2 = 0$), то математическое ожидание СрК строк (или столбцов) будет совпадать с соответствующей величиной для взаимодействий, а не для остаточной СК. Это приводит к критериям, основанным на S_2/S_4 и S_3/S_4. Но для проверки гипотезы $\sigma_{RC}^2 = 0$ можно, так же как в (35.65) для модели I, использовать критерий отношения взаимодействий и остатков, основанный на S_4/S_5. Таким образом, F-критерии в таблице ДА могут отличаться выбором разных СрК в знаменателе статистики.

Проверка гипотез в модели II ДА

36.16 Замечания в примерах 36.3—4 обращают наше внимание на тот факт, что до сих пор мы не дали теоретического обоснования использования F-критериев в модели II ДА. Даже если эти критерии совпадают с критериями модели I, их характеристики различны и нельзя гарантировать оптимальность. В любом случае видно, что могут понадобиться новые статистики критериев. Таким образом, мы приходим к рассмотрению теории критериев модели II.

Помня, что мы находимся в ситуации ДА, где $t-1$ квадратичных форм из (36.26) являются все суммами квадратов, можно переписать многомерное нормальное распределение наблюдений в виде

$$dG \propto \exp\left\{-\frac{1}{2}\left[\frac{n(\bar{y}-\theta)^2}{\lambda_1^*} + \sum_{j=2}^{t}\frac{S_j}{\lambda_j}\right]\right\}. \qquad (36.34)$$

Заметим теперь, что λ_1^*, определяемое как $\lambda_1 - n\theta^2$, не зависит от θ, поскольку, согласно **36.12**, $\lambda_1^* = \boldsymbol{P}_1\boldsymbol{V}_y\boldsymbol{P}_1'$, т. е. является суммой всех элементов матрицы $\boldsymbol{V}_y$, деленной на n. А элементы $\boldsymbol{V}_y$ являются дисперсиями и ковариациями, а значит, не зависят от θ. В тех случаях, которые будут рассматриваться,

$t = k+2$, т. е. числу оцениваемых параметров. Таким образом, у нас есть $k+2$ характеристических числа, которые являются функциями (в нашем случае всегда линейными) от $k+1$ параметров σ_j^2. Лишний параметр можно устранить, выразив λ_1^* в терминах остальных характеристических чисел, так что (36.34) примет вид

$$dG \propto \exp\left\{-\frac{1}{2}\left[\frac{n\bar{y}^2 - 2n\bar{y}\theta}{p(\lambda_2, \ldots, \lambda_t)} + \sum_{j=2}^{k+2}\frac{S_j}{\lambda_j}\right]\right\}. \qquad (36.35)$$

Гаучи (1959) показал, что совокупность распределений

$$f(t, \tau) \propto \exp\left\{\sum_{j=1}^{r} t_j\tau_j + t_1^2 g(\tau_2, \ldots, \tau_r)\right\} \qquad (36.36)$$

полна (результат, который не содержится в (23.19) т. 2). Сразу понятно, что (36.35) является частным случаем (36.36) с

$$t_j = S_j, \quad \tau_j = -1/(2\lambda_j), \quad j \geqslant 2,$$
$$t_1 = \bar{y}, \quad \tau_1 = n\theta/p(\lambda_2, \ldots, \lambda_t)$$

и

$$g(\tau_2, \ldots, \tau_r) = -\frac{1}{2}n/p(\lambda_2, \ldots, \lambda_t).$$

Таким образом, при заданной параметризации система (36.35) полна.

36.17 Теперь можно использовать результаты главы 23. Мы не в состоянии методами **23.27—36** получить РНМН-критерий, поскольку (36.36) более общее, чем (23.73). Однако, непосредственно применяя методы **23.20**, можно указать аналогичные РНМ-критерии.

На примере (36.33) видно, что в применениях к сбалансированной модели II ДА характеристические числа являются линейными формами параметров. Как указывает (36.33) или в простейшем случае пример 36.3, гипотеза, которую мы хотим проверять, о том, что какое-то конкретное σ_j^2 и только оно одно равняется нулю, эквивалентна гипотезе о совпадении двух конкретных характеристических чисел.

Заметим сначала, что гипотеза H_0: $\lambda_q = \lambda_p$; $q, p > 1$ оставляет нам множество из $(k+1)$ полных достаточных статистик

$$T = \{\bar{y}, S_2, \ldots, S_{p-1}, S_{p+1}, \ldots, S_{q-1}, S_{q+1}, \ldots, S_{k+2}, (S_p + S_q)\}.$$

Предположим теперь, что $p(\lambda)$ из (36.35) является функцией от λ_p или от λ_q, но не от обоих аргументов одновременно. Следуя **23.20**, видим, что каждая подобная область для H_0 состоит

из доли α каждого контура постоянства T. Зафиксируем теперь T и запишем

$$-\frac{1}{2}\left(\frac{S_p}{\lambda_p}+\frac{S_q}{\lambda_q}\right)=-\frac{1}{2}\left\{\frac{S_p+S_q}{\lambda_q}+\frac{S_p(\lambda_q-\lambda_p)}{\lambda_p\lambda_q}\right\}. \quad (36.37)$$

При фиксированном T применение леммы Неймана — Пирсона (22.6) к (36.35), в которое подставлено (36.37), показывает, что равномерно наиболее мощная критическая область размера α для проверки H_0 против H_1: $\lambda_p > \lambda_q$ задается соотношением

$$-\frac{1}{2}S_p(\lambda_q-\lambda_p) > c_\alpha(T),$$

так что РНМ-критическая область при любом фиксированном значении T задается малыми значениями S_p вне зависимости от значений параметров. Поскольку при фиксированном T значение S_p+S_q также фиксируется, то эта критическая область соответствует большим значениям отношения S_q/S_p. Наконец, легко видеть, что если H_0 справедлива, то распределение этого отношения не зависит от параметров и, в соответствии с упражнением 23.7, не зависит от полного семейства достаточных статистик T. Таким образом, критерий, основанный на больших значениях S_q/S_p, является безусловно подобным РНМ-критерию.

Таким образом, мы установили, что критерий, подобный РНМ, равенства математических ожиданий двух средних квадратов в таблице ДА для сбалансированной модели II, против альтернативы того, что одно из них больше другого, является F-критерием. Его статистика представляет из себя отношение (потенциально) большего СрК к меньшему. Большие значения соответствуют отклонению проверяемой гипотезы. По существу, впервые этот результат получен Гербахом (1959). В упражнении 36.4 читателю предлагается показать, что этот критерий является также РНМН. Однако в общем случае он не является критерием ОП (см. упражнения 36.5—6).

Пример 36.5

В примере 36.3 математические ожидания СрК были:

$$\begin{aligned} &\text{Групп:} && \lambda_2=\sigma_\varepsilon^2+n_i\sigma_1^2,\\ &\text{Остаточный:} && \lambda_3=\sigma_\varepsilon^2. \end{aligned} \quad (36.38)$$

При условии $\sigma_\varepsilon^2>0$, что мы всегда предполагаем, гипотеза $\lambda_2=\lambda_3$ совпадает с $H_0:\sigma_1^2=0$, а $H_1:\lambda_2>\lambda_3$ с $H_1:\sigma_1^2>0$. Таким образом, в соответствии с **36.17**, критерий, приведенный в примере 36.3, подобен РНМ. Оказывается, что в этом случае $\lambda_1^*=\lambda_2$, но это не влияет на упомянутый выше критерий (см. упражнение 36.2—3). Однако отсюда следует, что для проверки гипотезы о генеральном среднем в модели II сбалансированной однофакторной классификации, соответствующий СрК нужно сравнивать с СрК групп, а не с остаточным СрК, как это было для модели I в примере 35.1. (Это непосредственно следует из того факта, что p групповых средних являются независимыми нормальными величинами с математическим ожиданием θ и одинаковой дисперсией. Поэтому для проверки значения θ можно использовать t^2-распределение Стьюдента с $p-1$ ст. св.) Для модели II этот критерий о генеральном среднем можно применять вне зависимости от того, равняется σ_1^2 нулю или нет. Для модели I в примере 35.1 критерий был применим только при условии, что между группами нет разницы. Таким образом, модели отличаются даже в этом простом случае.

Пример 36.6

В примере 36.4 из (36.33) сразу следует, что если $\sigma_\varepsilon^2>0$, то $\sigma_R^2=0$ эквивалентно $\lambda_2=\lambda_4$; $\sigma_C^2=0$ эквивалентно $\lambda_3=\lambda_4$, а $\sigma_{RC}^2=0$ эквивалентно $\lambda_4=\lambda_5$. Таким образом, в соответствии с **36.17**, критерии для проверки этих гипотез, приведенные в конце примера 36.4, являются подобными РНМ.

Если мы сначала проверим и *примем* гипотезу $\sigma_{RC}^2=0$ ($\lambda_4=\lambda_5$), то соблазнительно проверять равенство $\sigma_R^2=0$ (или $\sigma_C^2=0$), сравнивая S_2 (или S_3) с объединением СК (S_4+S_5). Очевидно, что увеличение числа ст. св. в знаменателе дисперсионного отношения приводит к возрастанию мощности критерия. Но поскольку решение о возможности объединения S_4 и S_5 зависит от результата применения первоначального критерия, то оно может оказаться неверным при $\lambda_4\neq\lambda_5$. В результате подсчет размера полной процедуры проверки становится трудным. Сложная для вычислений теория и рекомендации для таких процедур объединения содержатся в работах Бозивича, Банкрофта и Хартли (1956) и Шривастава и Бозивича (1962).

Если заранее *известно*, что $\sigma_{RC}^2=0$, то приведенные критерии для σ_R^2 и σ_C^2 с объединением S_4 и S_5 справедливы и совпадают с критериями для столбцов и строк в модели I (пример 35.2), когда взаимодействия предполагаются равными нулю.

Поскольку СрК взаимодействий является знаменателем критериев об эффектах строк и эффектах столбцов, эти критерии можно применять даже в случае, когда все частоты в ячейках равны единице $m=1$. В модели I (см. пример 35.3) это было не так, если только заранее нельзя было предположить, что все

взаимодействия равны нулю. В рассматриваемом случае только критерий для проверки равенства $\sigma^2_{RC}=0$ становится неприменимым, когда все $n_{ij}=1$ и остаточная СК тождественно равна нулю.

Общие сбалансированные перекрестные классификации

36.18 Пример вычисления математических ожиданий СрК, задаваемых (36.33) для двухфакторной перекрестной классификации, служит образцом для сбалансированных перекрестных классификаций более высоких порядков. Для трехфакторной перекрестной классификации совершенно аналогично получаем, что математические ожидания СрК задаются следующими соотношениями, с очевидным обобщением обозначений:

$$\begin{array}{lll}
 & & \mathbf{M}(\text{СрК}) \\
\text{Строчек: } R & & \lambda_2=\sigma^2_\varepsilon+m\sigma^2_{RCL}+lm\sigma^2_{RC}+cm\sigma^2_{RL}+clm\sigma^2_R, \\
\text{Столбцов: } C & & \lambda_3=\sigma^2_\varepsilon+m\sigma^2_{RCL}+lm\sigma^2_{RC}+rm\sigma^2_{CL}+rlm\sigma^2_C, \\
\text{Слоев: } L & & \lambda_4=\sigma^2_\varepsilon+m\sigma^2_{RCL}+cm\sigma^2_{RL}+rm\sigma^2_{CL}+rcm\sigma^2_L, \\
\text{Взаимодействий первого порядка:} & (R\times C) & \lambda_5=\sigma^2_\varepsilon+m\sigma^2_{RCL}+lm\sigma^2_{RC}, \\
 & (R\times L) & \lambda_6=\sigma^2_\varepsilon+m\sigma^2_{RCL}+cm\sigma^2_{RL}, \\
 & (C\times L) & \lambda_7=\sigma^2_\varepsilon+m\sigma^2_{RCL}+rm\sigma^2_{CL}, \\
\text{Взаимодействий второго порядка:} & (R\times C\times L) & \lambda_8=\sigma^2_\varepsilon+m\sigma^2_{RCL}, \\
\text{Остаточный:} & & \lambda_9=\sigma^2_\varepsilon.
\end{array} \tag{36.39}$$

Это соответствует следующей модели, обобщающей (36.28):

$$y_{ijkp}=\theta+u_{i**}+u_{*j*}+u_{**k}+u_{ij*}+u_{i*k}+u_{*jk}+u_{ijk}+\varepsilon_{ijkp} \tag{36.40}$$

$$(i=1,\ldots,r;\ j=1,\ldots,c;\ k=1,\ldots,l;\ p=1,\ldots,m)$$

с $\mathbf{D}(u_{i**})=\sigma^2_R,\ \ldots,\ \mathbf{D}(u_{ij*})=\sigma^2_{RC},\ \ldots,\ \mathbf{D}(u_{ijk})=\sigma^2_{RCL}$.

Теперь понятно правило образования (36.39) (а также (36.33) и (36.38)). Математическое ожидание любого СрК равно σ^2_ε плюс умноженная на m линейная функция дисперсий. Эта линейная функция содержит те и только те из участвующих в модели дисперсий, индексы которых содержат все буквы, определяющие соответствующий СрК. Коэффициент при каждой такой дисперсии равен произведению максимально возможных значений индексов из (36.40), причем берется произведение тех букв, которые не входят в индекс соответствующей дисперсии. Если в индекс входят все буквы, то коэффициент равен единице. Так, для примера, рассматривая математическое ожидание СрК для $(R\times L)$-взаимодействий, получаем, что единственными дисперсиями, индекс которых содержит как R, так и L, являются σ^2_{RL} и σ^2_{RCL}. Индекс у σ^2_{RL} не содержит только C, которому в (36.40) соответствует j с максимально возможным значением c. Индекс у σ^2_{RCL} содержит все буквы, и соответствующий коэффициент равен единице. В результате мы получаем, что $(c\sigma^2_{RL}+\sigma^2_{RCL})$ нужно умножить на m и сложить с σ^2_ε, что совпадает со значением из (36.39). Более общее правило вычисления математических ожиданий СрК, как частный случай содержащее приведенную сбалансированную модель II, формулируется в работе Корнфилда и Тьюки (1956) (см. также Шеффе (1959)).

36.19 В (36.39) обнаруживается новое свойство трехфакторной перекрестной классификации, которое сохраняется для случая всех более высоких порядков. В примерах 36.5—6 было видно, что каждая из интересующих нас гипотез (о равенстве некоторой дисперсии нулю) в однофакторном и двухфакторном случае была эквивалентна гипотезе о равенстве двух математических ожиданий СрК, обозначаемых λ_j. Из (36.39) видно, что это свойство сохраняется только тогда, когда рассматриваются σ^2_{RCL}, σ^2_{RC}, σ^2_{RL} и σ^2_{CL}. Они равны нулю соответственно тогда и только тогда, когда $\lambda_8=\lambda_9$, $\lambda_5=\lambda_8$, $\lambda_6=\lambda_8$ или $\lambda_7=\lambda_8$. Таким образом, основываясь на **36.17**, взаимодействия первого и второго порядков, можно исследовать с помощью F-критерия, подобного РНМ. Однако ситуация отлична для других дисперсий σ^2_R, σ^2_C и σ^2_L.

Рассмотрим для примера σ^2_R, входящее в λ_2. Гипотезу H_0: $\sigma^2_R=0$ нельзя выразить в терминах равенства двух λ_j. Однако нетрудно видеть, что $\lambda_2+\lambda_8\geqslant\lambda_5+\lambda_6$ и что H_0 совпадает с

$$H_0:\ \lambda_2+\lambda_8=\lambda_5+\lambda_6. \tag{36.41}$$

Теория из **36.17** теперь неприменима, и, насколько нам известно, до сих пор не проведено исследование об оптимальном выборе критерия проверки (36.41). Однако, основываясь на результатах упражнения 36.7, можно указать приближенный критерий, причем, как понятно, аналогичное приближение возможно во всех тех случаях, когда исходную гипотезу о равенстве нулю дисперсии можно выразить в эквивалентной форме как гипотезу о равенстве нулю линейной комбинации λ_j (см. упражнение 36.8).

Иерархические классификации

36.20 Общая теория **36.13—15** и **36.16—17** применима не только к перекрестным, но и к сбалансированным иерархическим классификациям. В сбалансированном случае для получения дополнительного столбца с математическими ожиданиями СрК, необходимого в таблице ДА, можно сразу использовать тот факт, что иерархическую классификацию можно рассматривать как неполную перекрестную классификацию.

Пример 36.7 Сбалансированная двухфакторная иерархическая классификация

Рассмотрим модель:

$$y_{ijp} = \theta + u_i + u_{ij} + \varepsilon_{ijp} \qquad (36.42)$$
$$(i = 1, \ldots, k;\ j = 1, \ldots, l;\ p = 1, \ldots, m),$$

соответствующую k группам первого уровня классификации с одним и тем же числом l подгрупп в каждой из них и числом наблюдений m в каждой из kl подгрупп. Таблица (35.88) ДА из примера 35.5 сохраняется и здесь с очевидным изменением обозначений *). Нужно только привести еще математические ожидания СрК (за исключением генерального среднего). Заметим теперь, что (36.42) является вырожденным случаем (36.28), в котором нужно положить $\sigma_C^2 \equiv 0$ и заменить u_{i*} на u_i. Заменим также σ_R^2 и σ_{RC}^2 на σ_1^2 и σ_2^2. Учитывая эти изменения, получаем, что λ_3 и λ_4 из (36.33) совпадают, причем полная кратность этого характеристического числа равна $(l-1)+(k-1)(l-1) = k(l-1)$, т. е. числу ст. св. для подгрупп. Из (36.33) теперь получаем

$$\begin{array}{ll} & \mathbf{M}\,(\text{СрК}) \\ \text{Групп:} & \sigma_\varepsilon^2 + m\sigma_2^2 + lm\sigma_1^2, \\ \text{Подгрупп:} & \sigma_\varepsilon^2 + m\sigma_2^2, \\ \text{Остаточный:} & \sigma_\varepsilon^2. \end{array} \qquad (36.43)$$

Из **36.17** следует, что для проверки $\sigma_1^2 = 0$ или $\sigma_2^2 = 0$ применим критерий, подобный РНМ.

*) Заметим, в частности, что l из примера 35.5 соответствует здесь kl, полному числу подгрупп.

Пример 36.8 Сбалансированная трехфакторная иерархическая классификация

Видоизменяя (36.40) точно так же, как и в предыдущем примере, получим, что (36.39) примет такой вид (читателю предлагается проделать это самостоятельно в упражнении 36.9):

$$\begin{array}{ll} & \mathbf{M}\,(\text{СрК}) \\ \text{Групп:} & \sigma_\varepsilon^2 + m\sigma_3^2 + l_3 m\sigma_2^2 + l_2 l_3 m\sigma_1^2, \\ \text{Подгрупп:} & \sigma_\varepsilon^2 + m\sigma_3^2 + l_3 m\sigma_2^2, \\ \text{Подподгрупп:} & \sigma_\varepsilon^2 + m\sigma_3^2, \\ \text{Остаточный:} & \sigma_\varepsilon^2. \end{array} \qquad (36.44)$$

Здесь имеется l_1 групп, каждая из которых содержит l_2 подгрупп. Каждая подгруппа разбита на l_3 подподгрупп, содержащих m наблюдений. Так что число наблюдений равно $n = l_1 l_2 l_3 m$.

При этом мы не столкнемся с трудностями при получении критериев методом **36.17**. При изучении перекрестной классификации высоких порядков трудности в **36.19** возникли из-за кратности взаимодействий, чего нет в рассматриваемом случае.

Шеффе (1959) приводит обобщение трехфакторной иерархической классификации на случай с возможно неравными частотами.

Мощность критериев, доверительные интервалы и отрицательные значения оценок

36.21 В соответствии с **36.17** критерий, подобный РНМ, для проверки гипотезы о равенстве нулю какой-либо дисперсии эквивалентен проверке H_0: $\varphi = 0$ в $\lambda_q = \lambda_p + \varphi$ против H_1: $\varphi > 0$. Отношение $S_q/(\lambda_p + \varphi)$ к S_p/λ_p всегда имеет центральное F-распределение, так что имеем

$$F = \frac{S_q}{S_p}\left(1 + \frac{\varphi}{\lambda_p}\right)^{-1}. \qquad (36.45)$$

Отсюда немедленно получается мощность критерия проверки H_0, основанного на статистике S_q/S_p. В упражнении 36.1 рассматривается простейший случай.

36.22 В то время, как в модели I мы пришли к множественным сравнениям для параметров модели, которые были средними значениями, естественный следующий шаг в модели II состоит в построении доверительных интервалов для параметров, которые теперь являются дисперсиями.

Из (36.45) немедленно получаем доверительные интервалы для параметра φ/λ_p. Простейший случай рассматривается

в упражнении 36.11. Как показывает даже этот сымый простой случай, доверительные интервалы могут содержать отрицательные значения (и даже целиком содержаться в отрицательной области). Это соответствует возможности для точечных оценок дисперсий, полученных в примерах 36.3—4, принимать отрицательные значения. Для практических целей отрицательные значения оценок неотрицательных величин неприемлемы, и поэтому эти значения обычно заменяются нулем. И хотя таким образом нарушается несмещенность оценки, это может привести к уменьшению среднеквадратичной ошибки (см. пример 36.3). Аналогично отрицательная часть доверительного интервала обычно заменяется нулем.

В. Томпсон (1962) приводит алгоритм для получения неотрицательных оценок дисперсий. Этот алгоритм дает интуитивно приемлемый результат для однофакторной и двухфакторной перекрестной классификации, однако слишком сложен даже в трехфакторном случае.

Балмер (1957) (см. также Шеффе (1959)) получает приближенные доверительные интервалы для самого φ из (36.45), а не уже упомянутые менее полезные интервалы для φ/λ_p (см. упражнение 36.15).

Во всех случаях не возникает трудностей при построении доверительных интервалов для дисперсий ошибок σ_ε^2, исходя из распределения остаточной СК. Эти интервалы всегда лежат в положительной области.

Несбалансированный случай в модели II

36.23 Начиная с примера 36.2, мы ограничились изучением сбалансированного случая (с равными частотами). Чтобы показать трудности, возникающие в несбалансированном случае модели II, и пояснить, откуда они возникают, вернемся к однофакторной классификации, модель которой приведена в примере 36.1.

36.24 Мы видим, что

$$\operatorname{cov}(y_{iq}, y_{rs}) = \begin{cases} 0 & \text{при} \quad i \neq r, \\ \sigma_1^2 & \text{при} \quad i = r, \end{cases}$$

так что из предположения многомерной нормальности следует, что наблюдения, не принадлежащие одной и той же из p групп, независимы. Ковариация наблюдений, принадлежащих одной и той же группе, равна σ_1^2. Все наблюдения имеют нулевые математические ожидания и дисперсии $\sigma_1^2 + \sigma_\varepsilon^2$. (Это простейший случай общей формулы (36.11).)

Чтобы записать показатель экспоненты в формуле плотности многомерного нормального распределения, нам нужно обратить матрицу $\boldsymbol{V}_y$, которая, как мы видели, состоит из нулей, за исключением p квадратных подматриц, расположенных вдоль главной диагонали и имеющих размеры $n_1, \ldots, n_p$ (числа наблюдений в каждой из p групп). Каждая подматрица имеет вид

$$\boldsymbol{V}_{n_i} = \begin{pmatrix} \sigma_1^2 + \sigma_\varepsilon^2 & \sigma_1^2 & \ldots & \sigma_1^2 \\ \sigma_1^2 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \sigma_1^2 \\ \sigma_1^2 & \ldots & \sigma_1^2 & \sigma_1^2 + \sigma_\varepsilon^2 \end{pmatrix}.$$

Как нетрудно проверить простым умножением, обратная к $\boldsymbol{V}_{n_i}$ матрица имеет вид

$$\boldsymbol{V}_{n_i}^{-1} = \left\{\sigma_\varepsilon^2\left(\sigma_\varepsilon^2 + n_i\sigma_1^2\right)\right\}^{-1} \begin{pmatrix} \sigma_\varepsilon^2 + (n_i - 1)\sigma_1^2 & -\sigma_1^2 & \ldots & -\sigma_1^2 \\ -\sigma_1^2 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & -\sigma_1^2 \\ -\sigma_1^2 & \ldots & -\sigma_1^2 & \sigma_\varepsilon^2 + (n_i - 1)\sigma_1^2 \end{pmatrix},$$

а значит, $\boldsymbol{V}_y^{-1}$ состоит из $\boldsymbol{V}_{n_i}^{-1}$ вдоль главной диагонали, а остальные элементы равны нулю.

36.25 Из **36.24** следует, что логарифм ФП для однофакторной классификации равен

$$\log L = c(\sigma_1^2, \sigma_\varepsilon^2) - \frac{1}{2\sigma_\varepsilon^2}\left\{\sum_{i=1}^{p} \frac{\sigma_\varepsilon^2 + (n_i - 1)\sigma_1^2}{\sigma_\varepsilon^2 + n_i\sigma_1^2} \sum_{q=1}^{n_i} (y_{iq} - \theta)^2 - \right.$$

$$\left. - \sum_{i=1}^{p} \frac{\sigma_1^2}{\sigma_\varepsilon^2 + n_i\sigma_1^2} \sum\sum_{q \neq l} (y_{iq} - \theta)(y_{il} - \theta)\right\} =$$

$$= c(\sigma_1^2, \sigma_\varepsilon^2) - \frac{1}{2\sigma_\varepsilon^2}\left\{\sum_i \sum_q (y_{iq} - \theta)^2 - \sum_i \frac{\sigma_1^2}{\sigma_\varepsilon^2 + n_i\sigma_1^2}\left[\sum_q (y_{iq} - \theta)\right]^2\right\} =$$

$$= c(\sigma_1^2, \sigma_\varepsilon^2) - \frac{1}{2\sigma_\varepsilon^2}\left\{\sum_i \sum_q (y_{iq} - y_{..})^2 + n(y_{..} - \theta)^2 - \right.$$

$$\left. - \sum_i \frac{\sigma_1^2}{\sigma_\varepsilon^2 + n_i\sigma_1^2}[n_i(y_{i.} - \theta)]^2\right\}. \qquad (36.46)$$

Поскольку $y_{..} = \sum_{i=1}^{p} n_i y_{i.}/n$, то из (36.46) понятно, что $p+1$ статистик $\sum_i \sum_q (y_{iq} - y_{..})^2$, $y_{i.}$, $i = 1, \ldots, p$, достаточны для трех параметров задачи. Если все n_i равны, то самую последнюю сумму в правой части (36.46) можно переписать в виде

$$-\frac{n_i\sigma_1^2}{\sigma_\varepsilon^2 + n_i\sigma_1^2} \sum_i n_i (y_{i.} - \theta)^2 \equiv$$

$$\equiv -\frac{n_i\sigma_1^2}{\sigma_\varepsilon^2 + n_i\sigma_1^2} \left\{ \sum_i n_i (y_{i.} - y_{..})^2 + n(y_{..} - \theta)^2 \right\}, \qquad (36.47)$$

и, подставляя (36.47) в (36.46), убеждаемся, что в соответствии с общим результатом **36.11** для сбалансированного случая получается минимальное достаточное множество из трех статистик $\left\{ \sum_i \sum_q (y_{iq} - y_{..})^2,\ y_{..},\ \sum_i n_i (y_{i.} - y_{..})^2 \right\}$. Естественно, что они, по существу, и являются величинами, входящими в таблицу ДА, рассмотренную в примере 36.3.

Однако если не все n_i совпадают, то (36.47) перестает быть справедливым, и минимальная достаточная статистика имеет больше трех компонент (см. упражнение 36.12).

36.26 Отсутствие сбалансированности существенно влияет на ДА. СК между группами $S_2 = \sum_{i=1}^{p} n_i (y_{i.} - y_{..})^2$ не пропорциональна теперь величине, имеющей χ^2-распределение, поскольку теперь она является взвешенной суммой квадратов относительно среднего для нормальных величин с нулевыми математическими ожиданиями, но не равными дисперсиями. Однако распределение остаточной СК S_3 не меняется, так что, как и ранее,

$$\mathbf{M}\{S_3/(n-p)\} = \sigma_\varepsilon^2. \qquad (36.48)$$

С помощью таблицы ДА все еще можно оценить σ_1^2, но теперь это не является больше единственной оптимальной процедурой, как это было для сбалансированного случая в примере 36.3. В **36.25** мы видели, что S_3 и p групповых средних всегда образуют достаточную статистику. Рассмотрим функцию групповых средних

$$S(m_i) = \sum_{i=1}^{p} m_i y_{i.}^2 - \left(\sum_{i=1}^{p} m_i\right) y_{..}^2,$$

где m_i — константы. Поскольку из примера 36.3 следует

$$\mathbf{D}(y_{i.}) = \sigma_1^2 + \sigma_\varepsilon^2/n_i, \qquad (36.49)$$

мы убеждаемся, что

$$\mathbf{M}(S(m_i)) = \sigma_\varepsilon^2 \sum_i m_i \left(\frac{1}{n_i} - \frac{1}{n}\right) + \sigma_1^2 \left(1 - \frac{\sum_i n_i^2}{n^2}\right) \sum_i m_i. \qquad (36.50)$$

Соотношение (36.50), наряду с (36.48), можно использовать для получения несмещенной оценки σ_1^2 при любых m_i. Таким образом, получается множество несмещенных оценок значения σ_1^2 (за исключением сбалансированного случая (см. упражнение 36.13)). «Естественный» выбор $m_i = n_i$, который сводит $S(m_i)$ к S_2, является удобным, но не оптимальным.

Это отсутствие единственности указывает, что (как и вообще в случае, когда размерность вектора достаточных статистик превышает число параметров) в несбалансированном случае отсутствует полнота множества достаточных статистик.

Спитволл (1967) показал, что критерий проверки $\sigma_1^2 = 0$, основанный на S_2/S_3, против *больших* альтернативных значений σ_1^2, почти эквивалентен наиболее мощному инвариантному критерию, точная форма которого зависит от значения рассматриваемой альтернативы. Он показал также, что аналогичное соответствие имеет место для доверительных интервалов Вальда (1940) для $\sigma_1^2/\sigma_\varepsilon^2$, (см. упражнение 36.16), которые Вальд (1941, 1947) обобщил на более сложные несбалансированные модели.

Для альтернативных значений σ_1^2, близких к нулю, Мустафа (1967) показал, что критерий, использующий статистику $\sum_{i=1}^{p} n_i^2 (y_{i.} - y_{..})^2$, является значительно более мощным при проверке гипотезы $\sigma_1^2 = 0$, чем критерий, основанный на S_2/S_3.

36.27 Тьюки (1956—7) рассматривал задачу оптимального оценивания для несбалансированной однофакторной классификации. Он получил довольно сложный результат, который тем не менее нужно принять во внимание. Сёрл (1958) и Лоу (1964) исследовали несбалансированную двухфакторную перекрестную классификацию, когда снова СК в таблице ДА не пропорциональны величинам, имеющим χ^2-распределение. Хендерсон (1953) рассмотрел несколько методов несмещенного оценивания в общем несбалансированном случае. Харвилл (1967а) исследовал два из них, а также получил необходимые и достаточные условия, для того чтобы было возможно несмещенное оценивание.

Обобщение моделей ДА: обсуждение

36.28 Как исследуемая методом НК общая линейная модель I из главы 35, так и модель II настоящей главы, являются экстремальными случаями в том смысле, что *все* элементы вектора $\boldsymbol{\theta}$ в (19.8) являются константами (параметрами), в то время как *все* элементы $\boldsymbol{u}$ из (36.1) являются случайными вели-

чинами, некоррелированными между собой и с вектором ошибок ε. На практике может встретиться ситуация, когда уместно рассмотреть смесь двух моделей, т. е. когда

$$y = 1\theta_0 + X_1\theta + X_2 u + \varepsilon, \qquad (36.51)$$

где генеральное среднее обозначено теперь θ_0, чтобы отличить от вектора параметров θ, элементы которого константы.

36.29 Если с самого начала ограничиться обсуждением ситуации ДА (что мы не делали вплоть до главы 35 для модели I и вплоть до **36.13** для модели II), то легко понять, как может возникнуть «смешанная» модель вида (36.51). Рассмотрим для примера двухфакторную перекрестную классификацию.

Предположим, что проводится эксперимент для исследования прочности на разрыв пяти различных типов бумаги и что каждый тип бумаги рассматривается при трех различных уровнях содержания влаги. Результаты располагаются в таблицу с пятью строчками и тремя столбцами. Пять типов бумаги взяты в эксперименте потому, что они представляют самостоятельный интерес — мы хотим знать свойства этих типов бумаги. Поэтому естественно интерпретировать их прочность на разрыв в качестве некоторых искомых постоянных (параметров), измерение которых подвержено обычным экспериментальным ошибкам.

Ситуация может быть иной при классификации по столбцам. Три уровня содержания влаги будут возможно выбираться как удобные уровни для представления «высокого», «среднего» и «низкого» содержания влаги, не давая полной информации о точном значении уровня. В этом смысле три рассматриваемых уровня будут выбираться из совокупности потенциально возможных уровней содержания влаги некоторым (не обязательно вероятностным) образом. Таким образом, эффект столбцов будет иметь в некотором смысле распределение совершенно другой природы, чем ошибки экспериментов. Следовательно, при изучении классификации по столбцам более приемлемой идеализацией эксперимента является модель II (без предположения нормальности), хотя и она ни в коей мере не является адекватной. В результате в первом приближении приходим к представлению эксперимента с помощью модели типа (36.51), где θ соответствует эффектам строк, а u — эффектам столбцов.

36.30 В эксперименте, описанном в **36.29**, можно рассмотреть взаимодействия строк и столбцов, если допустить возможность того, что у пяти типов бумаги различная структура зависимости прочности на разрыв от уровня содержания влаги. Поскольку сами эффекты столбцов рассматриваются как случайные величины, кажется логически необходимым считать их взаимодействия со строками тоже случайными величинами, а не константами. Этого легко добиться, допустив, что u из (36.51) имеет компоненты, соответствующие взаимодействиям. Однако тогда немедленно появляется необходимость обратить внимание на момент, возникающий в самой модели II. Если допустить, что эффекты столбцов и взаимодействия строк и столбцов являются случайными величинами, то как можно оправдать предположение о том, что они не коррелированы, особенно если мы вспомним, что при дополнительном предположении нормальности это влечет независимость?

На основе реальных фактов вряд ли можно допустить независимость случайных эффектов и взаимодействий, и модель II, оставаясь интересной с математической точки зрения, вряд ли может служить удовлетворительным описанием многих практических ситуаций, в которых заведомо присутствуют взаимодействия.

36.31 Тот факт, что модель II ДА отождествляется с, вообще говоря, неприемлемым предположением о *независимости взаимодействий*, как мы его будем называть в дальнейшем, наталкивает на необходимость изучения других, более общих моделей ДА, не связанных с этим предположением. Мы увидим, что модель со *связанными взаимодействиями*, т. е. взаимодействиями, коррелированными с соответствующими главными эффектами, ведет на самом деле к процедуре ДА, отличной как от процедуры модели I, так и от процедуры модели II. Изложение полученных в последнее время результатов с историческими замечаниями приведено в работе Плэкетта (1960) и Шеффе (1956b).

Рассмотрим теперь модели ДА, разработанные в работах Корнфилда и Тьюки (1956) (следуя ранним неопубликованным работам этих авторов), Шеффе (1956a) и Уилка и Кемпсорна (1955—6). Мы ограничимся изложением двухфакторной перекрестной классификации, в которой проявляются все важные характерные черты моделей.

Общая модель

36.32 Предположим, что у нас имеется множество (дискретное или непрерывное) возможных уровней классификации по строкам, из которого r уровней выбираются для участия в эксперименте. Обозначим это множество P_R. Аналогично предположим, что имеется множество P_C возможных уровней классификации по столбцам, из которых выбираются c уровней. Процесс выбора для строк предполагается не зависящим от процесса выбора для столбцов. Если выбран i-й уровень строк и j-й уровень столбцов, то для этой комбинации проводится n_{ij} наблюдений (опытов). Как и ранее, p-е наблюдение в i-й строке

и j-м столбце обозначим y_{ijp} и положим

$$y_{ijp} = \mu_{ijp} + \varepsilon_{ijp}, \tag{36.52}$$

где $\mathsf{M}(\varepsilon_{ijp}) = 0$. На время оставим в стороне детальное рассмотрение вопроса о том, каким образом подбираются n_{ij} экспериментальных объектов (experimental units) для выбранной комбинации i-й строки и j-го столбца (см. **36.36** и **36.39** ниже). Обозначим через N_{ij} число экспериментальных объектов, которые в принципе (в силу структуры эксперимента) могут быть подобраны для этой комбинации, и обозначим μ_{ij*} среднее величин μ_{ijp} по всем возможным N_{ij} значениям индекса p. Далее μ_{i**} и μ_{*j*} — средние значения μ_{ij*} по всем уровням из P_C и всем уровням из P_R соответственно, а μ_{***} — среднее значение μ_{ij*}, взятое как по P_R, так и по P_C. Никакого предположения нормальности или гомоскедастичности, т. е. однородности условных дисперсий, не делается.

36.33 По аналогии с (35.31) генеральное среднее, эффекты столбцов, эффекты строк и взаимодействия строк и столбцов определяются теперь соответственно следующим образом:

$$\left.\begin{aligned}
\theta_{**} &= \mu_{***},\\
\theta_{i*} &= \mu_{i**} - \mu_{***},\\
\theta_{*j} &= \mu_{*j*} - \mu_{***},\\
\theta_{ij} &= \mu_{ij*} - \mu_{i**} - \mu_{*j*} + \mu_{***} =\\
&= \mu_{ij*} - \theta_{i*} - \theta_{*j} - \theta_{**}.
\end{aligned}\right\} \tag{36.53}$$

Хотя наблюдения будут проводиться только для r значений индекса i и для c значений индекса j, определения величин θ даются в терминах μ, которые являются функциями от *всех* априори рассматриваемых уровней, а не только от тех уровней, которые участвуют в проведении эксперимента. Далее в последней строке (36.53) θ_{ij} представляют собой *связанные взаимодействия*, относящиеся к эффектам популяций θ_{i*} и θ_{*j}.

Из (36.53) определим теперь компоненты дисперсии. Если в P_R содержится R элементов, а в P_C — C элементов, то положим

$$\left.\begin{aligned}
\sigma_R^2 &= \sum_{i=1}^{R} \theta_{i*}^2/(R-1),\\
\sigma_C^2 &= \sum_{j=1}^{C} \theta_{*j}^2/(C-1),\\
\sigma_{RC}^2 &= \sum_{i=1}^{R}\sum_{j=1}^{C} \theta_{ij}^2/\{(R-1)(C-1)\}.
\end{aligned}\right\} \tag{36.54}$$

При этом допускается, что R или C или R и C могут стремиться к бесконечности, что соответствует непрерывному случаю. Остаточная дисперсия определяется соотношением

$$\sigma_\varepsilon^2 = \frac{1}{RC}\sum_{i=1}^{R}\sum_{j=1}^{C}(N_{ij}-1)^{-1}\sum_{p=1}^{N_{ij}}(\mu_{ijp}-\mu_{ij*})^2. \tag{36.55}$$

36.34 Рассмотрим теперь различные частные случаи общей модели.

Случай 1

В P_R содержится ровно $R = r$ элементов, в P_C — ровно $C = c$; никакого выбора уровней строк и столбцов не проводится; величины N_{ij} стремятся к бесконечности для всех i и j.

Этот случай является основой для модели I ДА, изученной в главе 35. При этом (36.53) совпадает с (35.31).

Случай 2

Как P_R, так и P_C являются непрерывными множествами, так что R и C стремятся к бесконечности, наряду с N_{ij} для всех i и j.

Этот случай является основой для модели II ДА, рассмотренной в начале этой главы, при этом никаких предположений о нормальности или гомоскедастичности не делается. Из последнего равенства в (36.53) имеем

$$\mu_{ij*} = \theta_{**} + \theta_{i*} + \theta_{*j} + \theta_{ij}, \tag{36.56}$$

где все члены в правой части, за исключением θ_{**}, являются случайными величинами. Теперь взаимодействия θ_{ij} сделаны независимыми от θ_{i*} и θ_{*j} в силу бесконечности множества, из которого они выбираются. На самом деле (36.56) эквивалентно, с очевидным изменением обозначений, соотношению (36.28) (взятому для $y_{ij.}$, а не для отдельных наблюдений y_{ijp}).

Случай 3

Наблюдаемые r уровней строк выбираются случайным образом без возвращения из R уровней в P_R, а c наблюдаемых уровней столбцов независимо от выбора строк выбираются аналогично из C уровней множества P_C.

В этой модели, как и в модели II, обе классификации имеют случайный эффект, но со связанными взаимодействиями. При R, C и N_{ij}, стремящимися к бесконечности, получаем случай 2.

Следует заметить, что даже при $r = R$, $c = C$ и $N_{ij} \to \infty$, случай 3 не совпадает с приведенным выше случаем 1, поскольку в случае 1 не было никакого процесса выбора, а здесь процесс выбора определяет порядок, в котором строки и столбцы снабжаются индексами 1, 2, ..., r и 1, 2, ..., c соответственно. Этот процесс выбора приводит к рандомизации строк и (независимо) столбцов в таблице. Описанный в этом абзаце случай мы обозначим 3 (1).

Смешанные модели

36.35 *Случай 4*

Предположим, что в описанном выше случае 3 $R = r$, так что возможны только перестановки строк, которые в остальном фиксированы, но что, как и в случае 2, C и N_{ij} стремятся к бесконечности.

Получается *смешанная модель* с фиксированными, не считая перестановок, уровнями строк, и уровнями столбцов, выбираемыми из бесконечного множества. Вектор-столбец $\{\mu_{ij*}\}$, имеющий r компонент, соответствующих тому, что индекс i принимает значения 1, ..., r, является случайным вектором из-за выбора столбцов. Для различных j величины $\{\mu_{ij*}\}$ предполагаются в совокупности независимыми с одним и тем же многомерным распределением. Однако эта модель не является столь общей, как это может показаться с первого взгляда, ибо из процесса перестановки строк следует, что все дисперсии элементов $\{\mu_{ij*}\}$ равны и аналогично равны все $r(r-1)/2$ их ковариаций. Понятно, что это условие полной симметрии не всегда желательно для практических целей. Ниже мы рассмотрим дальнейшее обобщение, принадлежащее Шеффе (1956а), которое будем называть случаем 5.

Случай 5

Как и в случае 1, $R = r$, выбор не проводится, $C \to \infty$ и, как и в случае 4, $\{\mu_{ij*}\}$ являются одинаково распределенными случайными векторами. Однако дисперсионная матрица r их компонент больше не является полностью симметричной, что необходимо следовало из процесса перестановки строк в случае 4. Таким образом, получается более общая модель, чем в случае 4.

Имхоф (1960) обобщил модель случая 5 на сбалансированную трехфакторную классификацию при двух случайных (рандомизированных) и одной фиксированной классификациях переменных.

36.36 Корнфилд и Тьюки (1956) дали детальное изложение случая 3 из **36.34** (при этом случаи 2 и 4 являются его частными подслучаями). Они рассматривали не только двухфакторную перекрестную классификацию, которой мы здесь ограничились, но и общую сбалансированную классификацию. Во всех случаях они предполагали, что для каждой выбранной комбинации строки и столбца необходимые n_{ij} наблюдений получаются в результате случайного выбора без возвращения из разных генеральных совокупностей, состоящих из N_{ij} возможных экспериментальных объектов, и что эти rc различных групп наблюдений выбраны независимо от обсуждавшегося выше выбора уровней строк и столбцов. В случае сбалансированной двухфакторной перекрестной классификации, когда все $n_{ij} = m$, а также все $N_{ij} = N$, их результат о математических ожиданиях СрК в таблице ДА приведен в первых двух столбцах (36.57). Оставшиеся три столбца являются частным случаем этого результата.

Математические ожидания СрК в таблице ДА

	Случай 3	Частные подслучаи случая 3		
		Случай 3 (1)	Случай 2	Случай 4
Строчек	$\sigma_\varepsilon^2\left(1-\frac{m}{N}\right)+ m\sigma_{RC}^2\left(1-\frac{c}{C}\right)+ cm\sigma_R^2$	$\sigma_\varepsilon^2+cm\sigma_R^2$	$\sigma_\varepsilon^2+m\sigma_{RC}^2+ cm\sigma_R^2$	$\sigma_\varepsilon^2+m\sigma_{RC}^2+ cm\sigma_R^2$
Столбцов	$\sigma_\varepsilon^2\left(1-\frac{m}{N}\right)+ m\sigma_{RC}^2\left(1-\frac{r}{R}\right)+ rm\sigma_C^2$	$\sigma_\varepsilon^2+rm\sigma_C^2$	$\sigma_\varepsilon^2+m\sigma_{RC}^2+ rm\sigma_C^2$	$\sigma_\varepsilon^2+rm\sigma_C^2$
Взаимодействий	$\sigma_\varepsilon^2\left(1-\frac{m}{N}\right)+ m\sigma_{RC}^2$	$\sigma_\varepsilon^2+m\sigma_{RC}^2$	$\sigma_\varepsilon^2+m\sigma_{RC}^2$	$\sigma_\varepsilon^2+m\sigma_{RC}^2$
Остаточный	σ_ε^2	σ_ε^2	σ_ε^2	σ_ε^2

(36.57)

При этом, как и утверждалось при обсуждении случая 2, в **36.34**, столбец, соответствующий случаю 2, совпадает с (36.33) из примера 36.4, т. е. с результатом для модели II.

С другой стороны, несмотря на различие, которое мы указали в **36.34**, между случаем 3(1) и случаем 1, столбец в (36.57), соответствующий случаю 3(1), совпадает с результатом примера 35.2 в (35.65) для модели I. Это нетрудно проверить, вспоминая, что математическое ожидание случайной величины, имеющей нецентральное χ^2-распределение, равно сумме параметра нецентральности и числа ст. св., и используя определение (36.54) с $R = r$ и $C = c$. Таким образом, рандомизация не влияет на математические ожидания СрК, поскольку соответствующие СК симметричны относительно уровней строк и уровней столбцов.

36.37 Столбец в (36.57), соответствующий случаю 4, является для нас новым, и мы видим, что математическое ожидание СрК для строк (которые фиксированы с точностью до перестановок) совпадают с соответствующим значением в модели II (где эффекты строк рандомизированы). В то же время математическое ожидание СрК для столбцов (которые теперь рандомизированы) совпадает с соответствующим значением в модели I (где эффекты столбцов фиксированы). В действительности математическое ожидание СрК для строк определяется процессом выбора при классификации *по столбцам* и наоборот. СрК для строк связан с различиями между уровнями строк, все множество значений которых наблюдаемо (но только для имеющейся выборки уровней столбцов). Тем самым обусловливается эффект взаимодействия уровней строк и столбцов и соответственно σ^2_{RC} появляется в выражении для математического ожидания СрК строк. Наблюдаемая же выборка уровней-столбцов связана с каждым возможным уровнем строк, так что СрК столбцов не зависит от взаимодействий уровней строк и столбцов рассматриваемого множества и σ^2_{RC} не входит в математическое ожидание соответствующего СрК. Аналогичное явление имеет место и при более сложных классификациях. Выбор уровней оставшихся классификаций, а не рассматриваемой, определяет структуру математического ожидания СрК, относящегося к рассматриваемой классификации.

36.38 Математические ожидания СрК в случае 5 такие же, как и в (36.57) для случая 4. Это объясняется тем, что, так же как и в рассмотренном случае 3(1) в конце **36.36**, соответствующие СК являются симметричными функциями строк. Однако если предположить, что $\{\mu_{ij*}\}$ имеют многомерное нормальное распределение, то возникает важное различие между случаями 4 и 5. Оказывается (подробное изложение см. Шеффе (1956а, 1959)), что смешанная модель случая 5 перестает быть упрощением модели II, в которой (как мы показали в **36.15—17**) отношение каждой СК в таблице ДА к математическому ожиданию соответствующего СрК имеет центральное χ^2-распределение. В результате в модели II мы могли получать F-критерии, подобные РНМ, сравнивая каждый СрК с другим, имеющим при заданной гипотезе то же самое математическое ожидание. Как подсказывает последний столбец в (36.57), для случая 5 остается верным, что (при $m > 1$) гипотезу $\sigma^2_{RC} = 0$ можно проверять, используя F-критерий, основанный на частном от деления СрК взаимодействий на остаточный СрК, и гипотезу $\sigma^2_C = 0$ можно проверять аналогично. Но статистика, равная отношению СрК строк к остаточному СрК, вообще говоря, не имеет F-распределения, несмотря на то, что при $\sigma^2_R = 0$ ее числитель и знаменатель являются независимыми случайными величинами с одним и тем же математическим ожиданием. В качестве альтернативы Шеффе (1956а, 1959) дает статистику критерия, имеющую T^2-распределение Хотеллинга (это распределение подробно рассматривается в **41.15—17**). Однако в случае 4 статистика, равная отношению СрК строк к остаточному СрК, снова имеет распределение дисперсионного отношения (см. Имхоф (1962), который указал, что эффект от применения несоответствующего критерия дисперсионного отношения в случае 5 может оказаться существенным, и предложил другой F-критерий). Аналогичное исследование для обобщения случая 5, соответствующего анализу группы экспериментов, было проведено Калински (1966).

С. Рой и Кобб (1960) рассмотрели смешанную модель (без взаимодействий) с нормальными ошибками и одним или большим числом случайных эффектов, имеющих распределение, отличное от нормального. Хартли и Рао (1967) дают процедуру МП для общей несбалансированной смешанной модели с нормальными ошибками.

Упомянем кратко, что некоторые последовательные процедуры ДА изучались Д. Коксом (1952), Н. Джонсоном (1953—4) и Б. Гхошем (1964, 1967). В их статьях содержатся другие ссылки по этому вопросу.

Размещение экспериментальных объектов: рандомизация

36.39 Большое увеличение числа моделей ДА уже привело нас к существенным усложнениям. Но мы еще не приступили к рассмотрению важного источника, приводящего к разбросу в экспериментальных данных.

В **36.32** не обсуждался вопрос о размещении экспериментальных объектов между различными используемыми комбинациями строк и столбцов. В **36.36** предполагалось, что в общей модели, обозначенной там как случай 3, n_{ij} объектов, подобранных для выбранной комбинации строки и столбца, принадлежат совокупности, соответствующей этой комбинации и состоящей из N_{ij} объектов, так что всего имеется rc совокупностей

экспериментальных единиц. Эта ситуация, когда совокупности объектов вообще не пересекаются, является крайней.

Другая крайняя ситуация состоит в том, что все используемые экспериментальные объекты (например, в сбалансированной двухфакторной классификации их rmc) выбираются случайным образом без возвращения из одной и той же совокупности, состоящей из N объектов. Здесь $N_{ij} = N$ для всех i, j и, можно сказать, что имеется, так сказать, полное пересечение. Такой метод размещения носит название *полной рандомизации*, и любой эксперимент, в котором используется этот метод, называется *вполне рандомизированным экспериментом*.

Имеется также очевидная промежуточная ситуация частичного пересечения, когда совокупности экспериментальных объектов соответствуют группам комбинаций столбцов и строк, причем число таких совокупностей больше единицы и меньше rc. В этом случае говорят просто о *рандомизации* (в отличие от «полной рандомизации»). Например, для каждой строчки (или для каждого столбца) в эксперименте может быть своя совокупность экспериментальных объектов. В этом случае говорят о *случайных блоках*, к обсуждению которых мы вернемся в главе 38.

36.40 Возможно, что читателю не совсем понятно, почему термин «рандомизированный» не употребляется только для случая, когда подбор производится из разных совокупностей, хотя даже в этом случае каждая из совокупностей выбирается случайным образом. Как и в большинстве аналогичных случаев, употребление соответствующих терминов можно понять, только рассмотрев историю предмета. Самые первые формулировки ДА неявно предполагали, что мы имеем дело со случаем разных совокупностей. В самом деле, (36.57) показывает, что если каждая совокупность велика, то их объем оказывает малое влияние на таблицу ДА для случая 3. С другой стороны, вопросы эффективности планирования экспериментов (к которым мы еще вернемся в главе 38) впервые привели к рассмотрению случайных блоков и других методов размещения экспериментальных объектов, причем во всех этих методах явно и по существу присутствует процедура рандомизации.

36.41 Очевидно, что рандомизацию экспериментальных объектов нужно принять во внимание при анализе данных, но это приводит к существенным усложнениям и требует введения новых определений.

Мы хотим допустить возможность того, что μ_{ijp}, определенное в (36.52), как среднее значение для p-го наблюдения в (i, j)-й комбинации строки и столбца само может зависеть не только от (i, j), но и от характеристик конкретного экспериментального объекта, над которым проводится p-е наблюдение.

Рассмотрим любую группу комбинаций строк и столбцов, такую, что этой группе соответствует одна и та же совокупность экспериментальных объектов (см. обсуждение в **36.39**). Предположим, что эта группа состоит из m элементов, а соответствующая ей совокупность — из M экспериментальных объектов, причем $M \geqslant m$. Приведем теперь формальную двухфакторную классификацию для элементов группы и объектов. При этом в правой части приводимого ниже тождества индексы i и j объединяются скобкой, чтобы подчеркнуть, что (i, j) рассматривается теперь как элемент отдельной классификации. Итак,

$$\mu_{ijp} \equiv \mu_{(*)*} + \{\mu_{(ij)*} - \mu_{(*)*}\} + \{\mu_{(*)p} - \mu_{(*)*}\} + \\ + \{\mu_{(ij)p} - \mu_{(ij)*} - \mu_{(*)p} + \mu_{(*)*}\} \equiv \quad (36.58)$$

$$\equiv \mu_{(ij)*} + \{\mu_{(*)p} - \mu_{(*)*}\} + \{\mu_{(ij)p} - \mu_{(ij)*} - \mu_{(*)p} + \mu_{(*)*}\}, \quad (36.59)$$

где, как и ранее, звездочка означает усреднение.

Соотношение (36.58) разбивает μ_{ijp} на «генеральное среднее», два «главных эффекта» и «взаимодействия». Если оба члена в фигурных скобках в (36.59) равны нулю, то подбор экспериментальных объектов для соответствующей комбинации строки и столбца не влияет на μ_{ijp}, поскольку в этом случае μ_{ijp} равно просто среднему значению по всем объектам. Первый член в фигурных скобках в (36.59) называется *ошибкой объекта* (unit error), поскольку он относится к экспериментальному объекту. Второй член в фигурных скобках мы будем называть *ошибкой взаимодействий* (interactiv error), поскольку он относится к экспериментальному объекту и к комбинации строки и столбца. (Обычно этот член называют взаимодействием объекта и обработки.)

36.42 Если в изучаемой схеме двухфакторной классификации, изложенной в **36.32—3**, рассмотреть разложение μ_{ijp} на три компоненты в соответствии с (36.59), то модель станет еще более сложной. При этом очевидно, что ошибка взаимодействий в (36.59) снабжается индексами i, j и p так же, как ε_{ijp} из (36.52) (которое теперь мы называем *ошибкой измерения*, чтобы отличить от ошибок объекта и взаимодействий, определенных выше). В связи с этим естественно ожидать, что могут возникнуть трудности при оценке различных компонент модели.

Стоит отметить, что только ошибка измерения является ошибкой в обычном смысле этого слова, возникая из-за неточности измерений или наблюдений. Ошибки объектов и взаимодействий возникают из-за размещения экспериментальных объектов между комбинациями строчек и столбцов.

36.43 Уилк и Кемпсорн (1955—6) рассматривали одно-, двух- и трехфакторные классификации для случая 3 из **36.34**,

включая несбалансированный случай, когда проводится полная рандомизация экспериментальных объектов. Для случая пропорциональных частот всегда возможен ортогональный ДА (см. **35.21—2**). В общем (непропорциональном) случае применяется неортогональный ДА, использующий невзвешенные средние уровней (см. **35.31**). При этом, как и следовало ожидать, возникают трудности, упомянутые в конце **36.42**, так что оценить можно только некоторые функции параметров. Более того, как указал Плэкетт (1960), если сами уровни столбцов и строк подлежат выбору, то не совсем естественно *неравные* частоты n_{ij} рассматривать как фиксированные. Добавление еще одного процесса выбора для определения n_{ij} (этот процесс может быть коррелирован с наблюденными значениями игреков) ведет к дальнейшему усложнению анализа. Харвилл (1967b) делает допущение, что частоты коррелированы с эффектами в несбалансированной однофакторной классификации из модели II и исследует результирующее смещение у оценки σ_1^2 (см. **36.23—7**).

36.44 Мы отложим дальнейшее обсуждение рандомизированных моделей вплоть до главы 38, где будут понятны причины, приводящие к их рассмотрению. Для целей же настоящего изложения влияние этих моделей на процедуру ДА рассмотрено достаточно полно, и мы закончим эту главу обсуждением того, как применять изложенные в ней соображения.

Выбор модели ДА

36.45 Изобилие рассмотренных моделей ДА сталкивает прикладного статистика с проблемой, которая в менее острой форме возникает при использовании статистических методов. Очевидно, для того чтобы можно было выбрать модель, хорошо соответствующую реальной ситуации, нужно тщательно проанализировать все известные факты, относящиеся к происхождению наблюдений. И если таких фактов мало, то может понадобиться много предположений. В этом отношении статистик находится в ситуации, обычной для почти всех прикладных областей знаний.

Стоит указать, что обсуждавшееся в **36.32—8** множество моделей ДА различается предположениями о методах выбора уровней анализируемых факторов, а не предположениями о реальной природе этих факторов или лежащих в их основе величин. Если данные возникают из планируемого эксперимента, то предположения касаются поведения экспериментатора, а не имеющегося у него материала. С другой стороны, сложности, указанные в **36.39—43**, связаны, по существу, с природой объектов, над которыми проводится эксперимент. Перед экспериментатором всегда должен стоять вопрос, достаточно ли различаются исследуемые объекты, чтобы стоило вносить дополнительные сложности в анализ. В биологии, сельском хозяйстве и социальных исследованиях часто так и поступают, а в промышленном экспериментировании и физике часто об этом не заботятся. Мы стоим на той точке зрения, что даже если наблюдения являются результатом продуманного эксперимента, имеется достаточный простор для интуиции при высказывании соответствующих суждений. A fortiori, когда наблюдения не являются результатом планируемого эксперимента, правильность выбранного анализа зависит от интуиции статистика.

36.46 Теперь понятно, что ДА, так же как и другие статистические методы, не является «мельницей, автоматически перемалывающей результаты», вне зависимости от умения статистика заранее рассчитать и позаботиться об этом. Скорее это набор точных инструментов, которые можно использовать в соответствии с их назначением. При их использовании нужно как умение, так и упорная работа. Не нужно использовать сложную технику (а так часто поступают) для доказательства фактов, которые почти очевидны с первого взгляда. Статистик никогда не должен упускать из виду необходимость тщательного исследования имеющегося материала. Равным образом, часто используются несоответствующие модели. ДА не имеет монополии в неправильном использовании статистики, однако множество рассмотренных моделей делает его особо уязвимым к ошибкам со стороны слишком прямолинейных энтузиастов.

36.47 В следующей главе, последней из трех, относящихся к технике ДА, мы исследуем сначала задачу преобразования данных так, чтобы можно было применять ДА. Это ведет к вопросу об устойчивости ДА и к методам, свободным от распределений. Наконец, мы рассмотрим там трудности, возникающие из-за неполноты данных.

УПРАЖНЕНИЯ

36.1 В примере 36.3 показать, что функция мощности критерия размера α для проверки гипотезы H_0: $\sigma_1^2 = 0$ против H_1: $\sigma_1^2 > 0$ равна

$$P\left(\sigma_1^2\right) = 1 - G\left\{\left(1 + n_i \frac{\sigma_1^2}{\sigma_\varepsilon^2}\right)^{-1} G_\alpha\right\},$$

где G — функция распределения центрального F-распределения с соответствующим числом степеней свободы, а G_α — значение, которое соответствующая случайная величина превышает с вероятностью α, так что $G\{G_\alpha\} = 1 - \alpha$. Показать, что $P\left(\sigma_1^2\right)$ является монотонной функцией своего аргумента такой, что критерий не смещен и что он состоятелен, когда размеры групп n_i стремятся к бесконечности, но не в том случае, когда лишь число групп p стремится к бесконечности.

36.2 В **36.17** показать, что, поскольку $\bar{y}$ является компонентой T, член $\exp\left\{-\frac{1}{2}n\bar{y}^2/p(\lambda)\right\}$ в (36.35) не влияет на критерий, подобный РНМ при проверке H_0 против H_1, даже если $p(\lambda)$ пропорционально λ_p или λ_q.

36.3 В (36.26) показать, что сумма всех n характеристических чисел $\lambda_1^* + \sum_{j=2}^{t} \lambda_i n_j$ равна сумме дисперсий n наблюдений. Показать отсюда, что в примере 36.5 $\lambda_1^* = \lambda_2$.

36.4 Показать, что в **36.17** односторонний F-критерий, подобный РНМ, является несмещенным, а значит, является РНМН-критерием размера α.

36.5 Для сбалансированной однофакторной классификации в примерах 36.3 и 36.5 показать, что МП-оценки величин λ_2 и λ_3 имеют вид

$$\hat{\lambda}_2 = S_2/p, \qquad \hat{\lambda}_3 = S_3/\{p(n_i - 1)\} \quad \text{при} \quad \hat{\lambda}_2 \geqslant \hat{\lambda}_3$$

и равны

$$\hat{\hat{\lambda}}_2 = \hat{\hat{\lambda}}_3 = \{p\hat{\lambda}_2 + p(n_i - 1)\hat{\lambda}_3\}/n \quad \text{при} \quad \hat{\lambda}_2 < \hat{\lambda}_3.$$

Показать отсюда, что статистика l критерия ОП для проверки гипотезы $H_0: \lambda_2 = \lambda_3$ имеет вид

$$l^2 = \begin{cases} \dfrac{\hat{\lambda}_2^p \hat{\lambda}_3^{p(n_i-1)}}{\hat{\hat{\lambda}}_2^n} & \text{при} \quad \hat{\lambda}_2 \geqslant \hat{\lambda}_3 \\ 1 & \text{при} \quad \hat{\lambda}_2 < \hat{\lambda}_3, \end{cases}$$

так что $l = 1$, если только статистика F-критерия меньше чем $p/(p-1)$, а значит, критерий ОП не эквивалентен F-критерию. Показать далее, что при $F \geqslant p/(p-1)$ функция l монотонно убывает с ростом F. Показать, что при $\alpha < 0{,}25$ критическое значение $F_\alpha\{p-1, p(n_i-1)\}$ всегда больше чем $p/(p-1)$, так что для всех практических целей критерий ОП эквивалентен F-критерию, подобному РНМ.

(Гербах, 1959.)

36.6 Для сбалансированной двухфакторной перекрестной классификации в примерах 36.4 и 36.6 показать, что критерий ОП для проверки H_0: $\sigma_R^2 = 0$ является функцией от S_2, S_3 и S_4, а F-критерий, подобный РНМ, является функцией лишь от S_2/S_4, а значит, эти критерии не эквивалентны.

(Гербах, 1959.)

36.7 Показать, что если S_j/λ_j — независимые случайные величины, имеющие χ^2-распределение с ν_j ст. св., и $\lambda = \sum_j c_j\lambda_j$, то вторые и первые моменты у величины $S = \sum_j (c_j S_j/\nu_j)$ такие же, как если бы величина $\nu S/\lambda$ имела χ^2-распределение с числом ст. св.

$$\nu = \lambda^2 \Big/ \sum_j (c_j^2\lambda_j^2/\nu_j),$$

оценкой которого служит

$$\nu^* = S^2 \Big/ \sum_j (c_j^2 S_j^2/\nu_j).$$

Показать отсюда, что в **36.19** *приближенный* F-критерий проверки (36.41) может быть основан на статистике

$$(\text{CpK})_R/\{(\text{CpK})_{RC} + (\text{CpK})_{RL} - (\text{CpK})_{RCL}\},$$

имеющей, если справедливо (36.41), распределение дисперсионного отношения с $(r-1)$ и ν^* ст. св. Здесь все $c_j^2 = 1$ и $(\text{CpK})_j = S_j/\nu_j$.

(Сэттерсвэйт (1941) и Бокс (1954) проверили приближение численно, когда c_j положительны.)

36.8 В **36.18—19** показать, что при общей r-факторной сбалансированной перекрестной классификации только гипотезы о дисперсиях с r и $r-1$ индексами можно проверять критериями, подобными РНМ из **36.17**, но что для всех других дисперсий можно строить приближенные критерии, основанные на упражнении 36.7.

36.9 В примере 36.8 проверить (36.39), а также (36.44).

36.10 В примере 36.3 показать, что дисперсия несмещенной оценки величины σ_1^2 равна

$$\frac{2}{p-1}\left\{\sigma_1^2\left(\sigma_1^2 + \frac{2\sigma_\varepsilon^2}{n_i}\right) + \frac{(n-1)}{(n-p)\,n_i^2}\,\sigma_\varepsilon^4\right\},$$

а значит, оценка состоятельна при $p \to \infty$, но не является таковой, если только n_i стремятся к бесконечности (см. упражнение 36.1, где для состоятельности *критерия* нужно было, чтобы $n_i \to \infty$).

(Тьюки (1956—7) дал общее выражение для дисперсий и ковариаций величин, являющихся оценками дисперсий.)

36.11 В примере 36.3 получить доверительные интервалы для σ_ε^2 и $\sigma_1^2/\sigma_\varepsilon^2$ из распределений S_3 и S_2/S_3 соответственно. Показать, что последние интервалы могут частично или полностью лежать в отрицательной области.

36.12 В **36.25** показать, что минимальная достаточная статистика для трех параметров имеет $s = p + 1 - \sum_{r=3}^{p}(r-2)\,d_r$ компонент, где d_r — число r-кратных общих значений среди p значений n_i. Показать также, что $s = 1 + d_1 + 2\sum_{r=2}^{p} d_r$.

36.13 Доказать (36.50). Показать, что тогда и только тогда, когда все n_i равны, из (36.50) с использованием (36.43) получается единственная оценка величины σ_1^2.

36.14 Пусть z_{g_j} — независимые случайные величины, имеющие χ^2-распределение с $2g_j$ ст. св., где g_j — положительные целые числа. Разлагая х. ф. величины $\sum_{j=1}^{r} c_j z_{g_j}$ на элементарные дроби, показать, что

$$\mathsf{P}\left\{\sum_{j=1}^{r} c_j z_{g_j} > x\right\} = \sum_{j=1}^{r}\sum_{s=1}^{g_j} \alpha_{js}\mathsf{P}\{z_s > x/c_j\}.$$

Обозначая $1-2itc=y$ в х. ф., показать, что константы в правой части равны

$$\alpha_{k,\, g_k-h}=f_k^{(h)}(0)/h!,$$

где

$$f_k(y)=\prod_{\substack{j=1\\ j\neq k}}^{r}\{1+(y-1)\lambda_j/\lambda_k\}^{-g_k}.$$

(Бокс, 1954.)

36.15 Пусть $S_1/(\varphi+\lambda)$ и S_2/λ — независимые величины, имеющие χ^2-распределение с f_1 и f_2 ст. св. соответственно, $M_i=S_i/f_i$, $F=M_1/M_2$ (статистика дисперсионного отношения), а $F_{\alpha,\, f_2}$ является 100α-процентным значением функции распределения F с f_1, f_2 ст. св. Показать, что если для монотонно возрастающей функции $g(F)$ имеет место

$$\mathbf{P}\{M_2 g(F)\leqslant\varphi\}\doteqdot\alpha,$$

то она должна удовлетворять условиям

$$g(F_{\alpha,\, f_2})=0,$$

$$g(F)\sim F/F_{\alpha,\,\infty}\quad\text{при}\quad F\to\infty.$$

Показать, что эти условия выполняются для

$$g_1(F)=(F-F_{\alpha,\, f_2})/F_{\alpha,\,\infty}$$

и для

$$g_2(F)=(F/F_{\alpha,\,\infty})-1+(F_{\alpha,\, f_2}/F)\{1-(F_{\alpha,\, f_2}/F_{\alpha,\,\infty})\}.$$

(Балмер (1957) показал, что g_1 является плохим, а g_2 — очень хорошим приближением, — см. также Шеффе, 1959.)

36.16 В 36.23—6 положим

$$\tau_i=\left(\frac{\sigma_1^2}{\sigma_2^2}+\frac{1}{n_i}\right)^{-1}.$$

Показать, что $100(1-\alpha)$-процентный доверительный интервал для σ_1^2/σ_2^2 имеет вид

$$G_{1-\alpha/2}<\frac{n-p}{p-1}\sum_{i=1}^{p}\tau_i\left(y_{i\cdot}-\frac{\sum_i\tau_i y_{i\cdot}}{\sum_i\tau_i}\right)^2\Bigg|\,S_3<G_{\alpha/2},$$

где G_α определено в упражнении 36.1.

(Вальд, 1940; см. также Спитволл, 1967.)

ГЛАВА 37

ПРЕДПОЛОЖЕНИЯ ДИСПЕРСИОННОГО АНАЛИЗА

37.1 Когда решено, что изучаемой ситуации соответствует конкретная модель, должен быть решен еще один важный вопрос. Хотя из соображений удобства или из технических возможностей следует, что наблюдения нужно проводить над переменной y, остается решить, какую функцию от y нужно использовать для целей анализа. Априори не ясно, почему измеряемая величина, а не какая-нибудь функция от нее, лучше удовлетворяет предположениям модели.

Вместе с тем могут быть существенные практические соображения для рассмотрения конкретной функции от y, скажем, $g(y)$ (которая может оказаться просто равной y). Например, $g(y)$ может быть связано со стоимостью или доходом от изучаемого процесса. Но отсюда следует только, что для практических целей окончательные выводы анализа должны быть выражены в терминах $g(y)$. При этом предположение о том, что именно $g(y)$, а не какая-либо другая функция от y, лучше соответствует модели, может оказаться неоправданным.

37.2 Ставя задачу несколько более формально, можно сказать, что множество наблюдений над y равным образом является множеством «наблюдений» над любой данной функцией $g(y)$. Вопрос состоит в том, какие «наблюдения» $g(y)$ использовать. Очевидно, что нужно попытаться найти функцию, которая как можно лучше удовлетворяет предположениям модели. Впервые вопрос поиска преобразований такого типа изучали Бокс и Д. Кокс (1964). Их исследования применимы вообще к линейной модели с нормальными ошибками, частным случаем которой является модель I ДА, изложенная в главе 35. Читатель заметит, что и у нас в силу общности задачи первые вводные параграфы не ограничиваются контекстом ДА.

Преобразования к нормальной линейной модели

37.3 Следуя Боксу и Д. Коксу (1964), предположим, что наблюдается зависимая переменная y и множество регрессоров $x_1, \ldots, x_k$, и что мы хотим применить линейную модель с нормальными ошибками. Однако мы не собираемся некритично

Обозначая $1-2itc=y$ в х. ф., показать, что константы в правой части равны

$$\alpha_{k,\, g_k-h}=f_k^{(h)}(0)/h!,$$

где

$$f_k(y)=\prod_{\substack{j=1\\ j\neq k}}^{r}\{1+(y-1)\lambda_j/\lambda_k\}^{-g_k}.$$

(Бокс, 1954.)

36.15 Пусть $S_1/(\varphi+\lambda)$ и S_2/λ — независимые величины, имеющие χ^2-распределение с f_1 и f_2 ст. св. соответственно, $M_i=S_i/f_i$, $F=M_1/M_2$ (статистика дисперсионного отношения), а $F_{\alpha,\, f_2}$ является 100α-процентным значением функции распределения F с f_1, f_2 ст. св. Показать, что если для монотонно возрастающей функции $g(F)$ имеет место

$$\mathbf{P}\{M_2 g(F)\leqslant\varphi\}\doteqdot\alpha,$$

то она должна удовлетворять условиям

$$g(F_{\alpha,\, f_2})=0,$$

$$g(F)\sim F/F_{\alpha,\,\infty}\quad\text{при}\quad F\to\infty.$$

Показать, что эти условия выполняются для

$$g_1(F)=(F-F_{\alpha,\, f_2})/F_{\alpha,\,\infty}$$

и для

$$g_2(F)=(F/F_{\alpha,\,\infty})-1+(F_{\alpha,\, f_2}/F)\{1-(F_{\alpha,\, f_2}/F_{\alpha,\,\infty})\}.$$

(Балмер (1957) показал, что g_1 является плохим, а g_2 — очень хорошим приближением, — см. также Шеффе, 1959.)

36.16 В 36.23—6 положим

$$\tau_i=\left(\frac{\sigma_1^2}{\sigma_2^2}+\frac{1}{n_i}\right)^{-1}.$$

Показать, что $100(1-\alpha)$-процентный доверительный интервал для σ_1^2/σ_2^2 имеет вид

$$G_{1-\alpha/2}<\frac{n-p}{p-1}\sum_{i=1}^{p}\tau_i\left(y_{i\cdot}-\frac{\sum_i\tau_i y_{i\cdot}}{\sum_i\tau_i}\right)^2\Bigg|\,S_3<G_{\alpha/2},$$

где G_α определено в упражнении 36.1.

(Вальд, 1940; см. также Спитволл, 1967.)

ГЛАВА 37

ПРЕДПОЛОЖЕНИЯ ДИСПЕРСИОННОГО АНАЛИЗА

37.1 Когда решено, что изучаемой ситуации соответствует конкретная модель, должен быть решен еще один важный вопрос. Хотя из соображений удобства или из технических возможностей следует, что наблюдения нужно проводить над переменной y, остается решить, какую функцию от y нужно использовать для целей анализа. Априори не ясно, почему измеряемая величина, а не какая-нибудь функция от нее, лучше удовлетворяет предположениям модели.

Вместе с тем могут быть существенные практические соображения для рассмотрения конкретной функции от y, скажем, $g(y)$ (которая может оказаться просто равной y). Например, $g(y)$ может быть связано со стоимостью или доходом от изучаемого процесса. Но отсюда следует только, что для практических целей окончательные выводы анализа должны быть выражены в терминах $g(y)$. При этом предположение о том, что именно $g(y)$, а не какая-либо другая функция от y, лучше соответствует модели, может оказаться неоправданным.

37.2 Ставя задачу несколько более формально, можно сказать, что множество наблюдений над y равным образом является множеством «наблюдений» над любой данной функцией $g(y)$. Вопрос состоит в том, какие «наблюдения» $g(y)$ использовать. Очевидно, что нужно попытаться найти функцию, которая как можно лучше удовлетворяет предположениям модели. Впервые вопрос поиска преобразований такого типа изучали Бокс и Д. Кокс (1964). Их исследования применимы вообще к линейной модели с нормальными ошибками, частным случаем которой является модель I ДА, изложенная в главе 35. Читатель заметит, что и у нас в силу общности задачи первые вводные параграфы не ограничиваются контекстом ДА.

Преобразования к нормальной линейной модели

37.3 Следуя Боксу и Д. Коксу (1964), предположим, что наблюдается зависимая переменная y и множество регрессоров $x_1, \ldots, x_k$, и что мы хотим применить линейную модель с нормальными ошибками. Однако мы не собираемся некритично

предположить, что справедливо

$$y = X\theta + \varepsilon,$$

а хотим найти преобразования как над y, так и над каждым из x такие, чтобы выполнялось

$$y_\lambda = X_\mu \theta + \varepsilon, \tag{37.1}$$

где компоненты вектора ε — независимые нормальные величины с нулевым средним и одинаковой дисперсией σ^2. В (37.1) $\lambda = (\lambda_1, \lambda_2, \ldots)$ указывает на преобразование над y внутри выбранного параметрического семейства преобразований и аналогично $\mu = (\mu_1, \ldots, \mu_k)$ указывает на преобразования (разные) регрессоров. Тем самым мы обобщаем наши рассуждения из вводной части главы, где говорилось только о преобразованиях над y.

37.4 В соответствии с (37.1) логарифм ФП равен

$$\log L_{\lambda,\mu}(y \mid \theta, \sigma^2) =$$

$$= -\frac{1}{2} n \log (2\pi\sigma^2) - \frac{1}{2\sigma^2}(y_\lambda - X_\mu\theta)'(y_\lambda - X_\mu\theta) + \log J_\lambda, \tag{37.2}$$

где J_λ — якобиан обратного преобразования от y_λ (нормально распределенная величина в (37.1)) к реально наблюдаемой величине y. Максимизируя теперь ФП (37.2) относительно θ и σ^2 при фиксированных λ и μ, получим, как и в **24.28** т. 2, что средний член равен константе. Таким образом, если пренебречь константой, то для условного максимума при фиксированных λ и μ имеем

$$\log L_{\lambda,\mu}(y \mid \hat{\theta}, \hat{\sigma}^2) = -\frac{1}{2} n \log \hat{\sigma}^2_{\lambda,\mu} + \log J_\lambda, \tag{37.3}$$

где, как и в **24.28**, $n\hat{\sigma}^2_{\lambda,\mu} = y_\lambda' T y_\lambda$ является остаточной СК.

Теперь, исходя из условного максимума (37.3), нужно вычислить абсолютный максимум по всем возможным значениям λ и μ. Даже с применением электронных вычислительных машин это чрезвычайно сложная задача. Исключение составляют лишь случаи, когда преобразования определяются одним или двумя индексами, например когда

(а) преобразуется только зависимая переменная y, причем λ имеет одну или две компоненты;

(б) *одно и то же* преобразование применяется ко всем регрессорам (или к подмножеству), при этом μ имеет одну или две компоненты;

(в) как и в (а) λ имеет одну компоненту, а (б) выполняется с одной компонентой у μ.

В случаях (б) и (в), вообще говоря, необходимо численное построение линий уровня (37.2) для всех λ и μ. Далее мы ограничимся случаем (а), где преобразуется только зависимая переменная. В задачах ДА, когда изучаются классифицированные данные, регрессоры являются величинами, принимающими лишь значения 0 или 1 (см. **35.9—10**), а значит, указанное ограничение несущественно. При более общем изучении регрессии мы вынуждены будем выбирать нужный вид регрессоров до того, как будем рассматривать преобразования зависимой переменной. Бокс и Тидвелл (1962) рассматривают преобразования регрессоров к более простому виду (см. упражнение 37.9). Такие преобразования, конечно, не влияют на нормальность или гомоскедастичность ошибок.

37.5 Возвращаясь, таким образом, к задачам, о которых говорилось в начале обсуждения, мы рассмотрим преобразования одного только y. На практике самыми полезными являются степенные и логарифмические преобразования y с возможным сдвигом на константу. В соответствии с этим мы рассмотрим семейство преобразований

$$y_\lambda = \begin{cases} (y + \lambda_2)^{\lambda_1}, & \lambda_1 \neq 0, \\ \log (y + \lambda_2), & \lambda_1 = 0. \end{cases} \tag{37.4}$$

Для того чтобы при $\lambda_1 = 0$ сохранить непрерывность, мы перепишем это в эквивалентной форме

$$y_\lambda = \begin{cases} \{(y + \lambda_2)^{\lambda_1} - 1\}/\lambda_1, & \lambda_1 \neq 0, \\ \log (y + \lambda_2), & \lambda_1 = 0. \end{cases} \tag{37.5}$$

Тьюки (1957b) изучает и описывает структурные свойства семейства (37.4) при $\lambda_1 \leqslant 1$. Долби (1963) рассматривает свойства дифференциального уравнения, которому удовлетворяет y_λ:

$$\{(y_\lambda'/y_\lambda'')'\} = (\lambda_1 - 1)^{-1}.$$

Хили и Тейлор (1962) дают таблицы для преобразований с дробными степенями, когда $\lambda_2 = 0$, а λ_1 кратно 0,2.

37.6 Теперь в (37.3) получаем

$$\log J_\lambda = (\lambda_1 - 1) \sum_{i=1}^{n} \log (y_i + \lambda_2). \tag{37.6}$$

Для численного определения абсолютного максимума теперь можно построить значения (37.3) для выбранного множества значений (λ_1, λ_2). При этом для каждой такой пары (λ_1, λ_2) должен быть проведен ДА для вычисления остаточной СК в (37.3). В простейшем случае, когда $\lambda_2 = 0$, этого можно не

делать, приравнивая к нулю первую производную по λ_1 от выражения (37.3). С использованием (37.5—6) это приводит к

$$0 = \frac{\partial \log L_{\lambda_1}(y \mid \hat{\theta}, \hat{\sigma}^2)}{\partial \lambda_1} = -n \frac{y'_{\lambda_1} T u_{\lambda_1}}{y'_{\lambda_1} T y_{\lambda_1}} + n\lambda_1^{-1} + \sum_{i=1}^{n} \log y_i, \qquad (37.7)$$

где компоненты вектора $\boldsymbol{u}$ равны $\{\lambda_1^{-1} y_i^{\lambda_1} \log y_i\}$.

Критерии ОП для вложенных гипотез

37.7 Бокс и Д. Кокс (1964) приводят несколько интересных численных примеров применения описанного метода нахождения преобразований, а также байесовского метода анализа, который они исследуют. Кроме того, они рассматривают разложение максимизируемой ФП на три компоненты, соответствующие нормальности, гомоскедастичности и структуре математического ожидания y_λ. Их исследования применимы к более общей ситуации.

Рассмотрим множество ограничений $C_1, C_2, \ldots$, которые последовательно вводятся в математическую модель. Пусть $\hat{\lambda}_{(s)}$ — МП-оценка параметра λ, когда в модель введены $C_1, \ldots, C_s$. $\hat{\lambda}$ — без индекса означает МП-оценку без ограничений. Тогда для любого s

$$L(y \mid \hat{\lambda}_{(s)}) = L(y \mid \hat{\lambda}) \cdot \frac{L(y \mid \hat{\lambda}_{(1)})}{L(y \mid \hat{\lambda})} \cdot \frac{L(y \mid \hat{\lambda}_{(2)})}{L(y \mid \hat{\lambda}_{(1)})} \cdot \ldots \cdot \frac{L(y \mid \hat{\lambda}_{(s)})}{L(y \mid \hat{\lambda}_{(s-1)})} = \\ = L(y \mid \hat{\lambda}) \cdot l_1 \cdot \ldots \cdot l_s, \qquad (37.8)$$

где l_p — статистика критерия ОП для проверки множества ограничений $C_1, C_2, \ldots, C_{p-1}, C_p$ против множества $C_1, C_2, \ldots, C_{p-1}$ (см. **24.1**, т. 2). Каждая статистика l_p принимает значения между нулем и единицей и в условиях регулярности величина $-2\log l_p$ имеет асимптотически нецентральное χ^2-распределение с числом ст. св., равным числу независимых ограничений, которое C_p налагает на параметры (см. **27.7**). Если C_p справедливо, то получается центральное χ^2-распределение и, таким образом, $-2\log l_p$ можно использовать для проверки гипотезы о значениях добавляемого C_p к уже наложенным ограничениям $C_1, \ldots, C_{p-1}$. Следует заметить, что l_p, вообще говоря, не являются независимыми, хотя в частных случаях для некоторых гипотез они могут оказаться независимыми (см. упражнения 24.6 и 24.13 и более общий результат в упражнении 37.2). В упражнении 37.1 читателю предлагается применить разложение (37.8) к рассматриваемой задаче. Результат сразу следует из некоторых фактов, приведенных в главе 24.

Цели преобразования

37.8 Достоинство метода МП, обсуждавшегося в **37.3—7** состоит в том, что не требуется априорных знаний о соотношении между y и регрессорами, или о характере распределения ошибок непреобразованной величины y. Исходным являлось предположение о том, что в рассматриваемом семействе существует преобразование, для которого выполняются все условия линейной модели, включая гомоскедастичность и нормальность распределения ошибок. Конечно, на практике этого может и не быть, однако даже тогда метод МП нахождения преобразования может дать улучшение по сравнению с непосредственным использованием исходных y. Поразителен факт (иллюстрируемый численными примерами, приведенными Боксом и Коксом (1964)), что это преобразование МП очень близко к преобразованию, которое вытекает из нестатистического рассмотрения природы изучаемых величин. Естественно, что такое рассмотрение там, где оно только возможно, должно дополнять и служить основой для самого статистического анализа.

Крускал (1965) приводит основанный на вычислениях метод нахождения монотонного преобразования наблюдений, минимизирующего остаточную СК (взятую в нужном масштабе) от предполагаемой линейной модели. При этом не используется предположение нормальности и не требуется существования параметрического семейства, аналогичного (37.5). Он использует свой метод, чтобы заново проанализировать примеры Бокса и Д. Кокса и некоторые другие.

37.9 Другие подходы к нахождению преобразования исходных данных, позволяющего удовлетворить те или иные требования линейной модели, преследуют более узкие цели. Либо ищется преобразование, нормализующее ошибки, либо стабилизирующее дисперсии, либо преобразование, устраняющее взаимодействия, такое, чтобы эффекты стали аддитивными. При этом хотелось бы, чтобы преобразование, преследующее одну из этих целей, по крайней мере помогало в достижении других. Замечательно, что на самом деле это часто оказывается так, и мы кратко рассмотрим несколько важных примеров. Однако было бы сверхоптимистично ожидать, что так будет всегда. Легко привести примеры, когда цели аддитивности и гомоскедастичности несовместны. Так, если в двухфакторной перекрестной классификации ожидаемое среднее значение y аддитивно по эффектам строк и столбцов, а ошибки не имеют нормального распределения с дисперсией, являющейся функцией $\mathbf{M}(y)$, то любое преобразование, устраняющее неоднородность дисперсий (гетероскедастичность), будет нарушать строгую аддитивность.

Аналогичный эффект может проявиться при попытке устранить отсутствие нормальности.

Рассмотрим теперь по порядку различные типы таких преобразований.

Преобразования, стабилизирующие дисперсии

37.10 Пусть статистика t имеет среднее θ и дисперсию при фиксированном объеме выборки n

$$\mathsf{D}(t) = D_n^2(\theta). \tag{37.9}$$

Чтобы устранить эту зависимость дисперсии от параметра θ, мы хотим найти функцию $u(t)$ такую, чтобы ее дисперсия $\mathsf{D}(u)$ была константой, скажем, c. Однако, вообще говоря, это нельзя сделать точно, и ищется только такое $u(t)$, чтобы

$$\mathsf{D}\{u(t)\} = c\{1 + O(R^{-1})\}, \tag{37.10}$$

где R — известная константа, достаточно большая, чтобы можно было пренебречь R^{-1}. В частности, может быть $R = n$, где n — объем выборки. Предположим далее, что мы ограничиваемся значениями t, сосредоточенными в окрестности своего среднего значения θ. Тогда, используя **10.6—7**, из (10.14) получим приближение

$$\mathsf{D}\{u(t)\} = \left\{\left(\frac{du(t)}{dt}\right)^2\right\}_{t=\theta} \mathsf{D}(t). \tag{37.11}$$

Приравнивая (37.10) и (37.11), получим приближение первого порядка

$$\left\{\left(\frac{du(t)}{dt}\right)^2\right\}_{t=\theta} = \frac{c}{D_n^2(\theta)}. \tag{37.12}$$

Поскольку рассматривается окрестность θ, мы не будем в индексе писать $t = \theta$, а просто вместо t напишем θ. Таким образом,

$$\frac{du(\theta)}{d\theta} \propto \{D_n^2(\theta)\}^{-1/2}. \tag{37.13}$$

Мы заменили здесь константу c на единицу, поскольку ее значение определяется по нашему усмотрению, ибо умножение $u(t)$ на константу не влияет на нашу цель получить соотношение вида (37.10). Проинтегрируем теперь уравнение (37.13), опуская без потери общности аддитивную постоянную, возникающую при взятии неопределенного интеграла (это можно сделать, поскольку опять соотношение (37.10) сохраняется). В результате получим

$$u(t) \propto \left\{\int \frac{d\theta}{D_n(\theta)}\right\}_{\theta=t}. \tag{37.14}$$

37.11 Хотя (37.14) получено с помощью приближения, можно проверить его справедливость, если известно теоретическое распределение статистики t. Нужно подсчитать теоретическую дисперсию $u(t)$ и проверить стабильность при изменении θ. Может оказаться желательным модифицировать $u(t)$, чтобы улучшить стабильность. С другой стороны, если имеются только наблюдения величины t и нет никакой информации о ее распределении или о параметре θ этого распределения, то мы не можем даже точно вычислить $D_n^2(\theta)$. В таких случаях нужно подсчитать среднее значение и дисперсию величины t для отдельных групп наблюдений и построить график зависимости дисперсии от среднего для *оценки* соотношения (37.9), на котором основано соотношение (37.14). В этом случае аппроксимация более рискована, но тем не менее часто дает удовлетворительный для практики результат.

Пример 37.1

Если t имеет распределение Пуассона, то из (5.20) следует, что среднее и дисперсия равны, так что здесь (37.9) превращается в

$$D_n^2(\theta) = \mathsf{D}(t) = \theta,$$

и (37.14) дает

$$u(t) \propto \left\{\int \theta^{-1/2} d\theta\right\}_{\theta=t} \propto t^{1/2} \tag{37.15}$$

— простое преобразование: извлечение квадратного корня. Для первого приближения, в соответствии с (37.11),

$$\mathsf{D}(t^{1/2}) = \left\{\left(\frac{1}{2} t^{-1/2}\right)^2\right\}_{t=\theta} \mathsf{D}(t) = \frac{1}{4}, \tag{37.16}$$

что подтверждает, что в этом приближении дисперсия стабилизируется.

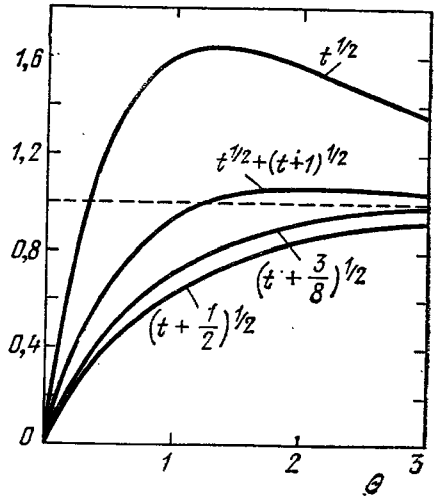

Рис. 37.1.

Бартлетт (1936) указывает, что в этом случае стабилизацию дисперсии можно улучшить, осуществив сдвиг t, перед извлечением квадратного корня. Если определить

$$u_c(t) = (t + c)^{1/2},$$

то Бартлетт предлагает использовать $c = 1/2$. В упражнении 37.15 показано, что лучшим является выбор $c = 3/8$. В нижеследующей таблице даны значения дисперсии $u_c(t)$ в долях от предельного значения дисперсии при $\theta \to \infty$ для $c = 0$, 1/2 и 3/8. Подсчет проведен Бартлеттом (1938) и Энскомбом (1948).

θ	Значения дисперсии $u_c(t)$ в долях от предельного значения дисперсии			θ	Значения дисперсии $u_c(t)$ в долях от предельного значения дисперсии		
	$c=0$	$c=1/2$	$c=3/8$		$c=0$	$c=1/2$	$c=3/8$
0	0	0	0	6,0	1,104	0,980	1,002
0,5	1,240	0,408		9,0	1,052	0,988	
1,0	1,608	0,640	0,717	10,0			1,001
2,0	1,560	0,856	0,924	12,0	1,036	0,992	
3,0	1,360	0,928	0,983	15,0	1,024	0,992	
4,0	1,224	0,960	0,999	20,0			1,000

При малых θ очевидна неадекватность простейшего преобразования с $c = 0$. При $\theta < 3$ то же самое сравнение проведено графически на рис. 37.1, который взят из статьи М. Фримэна и Тьюки (1950). Предложенное этими авторами стабилизирующее дисперсию преобразование $u' = t^{1/2} + (t+1)^{1/2}$ при $\theta \leqslant 2$ более стабильно, чем $u_{3/8}(t)$, а при $\theta > 2$ оба служат хорошим приближением. Для u' при $\theta \geqslant 1$ отклонение от постоянного значения не превышает 6% и u' кажется наилучшим выбором (см. упражнение 37.17). Мостеллер и Ютц (1961) приводят таблицу u' для $x = 1(1)50$.

Пример 37.2

Если t — выборочная дисперсия, а σ^2 — дисперсия генеральной совокупности, то, как неоднократно говорилось выше, для выборки из нормального распределения величина $z = nt/(2\sigma^2)$ имеет гамма-распределение с параметром $(n-1)/2$, т. е. распределение z задается соотношением

$$dF(z) = \frac{1}{\Gamma\left\{\frac{1}{2}(n-1)\right\}} e^{-z} z^{\frac{1}{2}(n-1)-1}\, dz$$

(см., например, (11.25) в других обозначениях). Таким образом, как среднее, так и дисперсия величины z равны $(n-1)/2$, а значит, для самого t имеем

$$\theta = \mathsf{M}(t) = \frac{2\sigma^2}{n} \cdot \frac{1}{2}(n-1) = \sigma^2(n-1)/n,$$

$$D_n^2(\theta) = \mathsf{D}(t) = \left(\frac{2\sigma^2}{n}\right)^2 \cdot \frac{1}{2}(n-1) = 2\sigma^4(n-1)/n^2,$$

так что в этом случае

$$D_n^2(\theta) = 2\theta^2/(n-1).$$

Из (37.14) имеем

$$u(t) \propto \left\{\int \theta^{-1}\, d\theta\right\}_{\theta=t} = \log t, \tag{37.17}$$

т. е. мы получим просто логарифмическое преобразование. Поскольку

$$\log t = \log z + \log(2\sigma^2/n),$$

то семиинварианты величин $\log t$ и $\log z$ совпадают, за исключением постоянной разницы в среднем значении. Легко видеть, что эти семиинварианты вообще не зависят от σ^2. Обозначая $p = (n-1)/2$, получим, что характеристическая функция $\log z$ равна

$$\varphi(w) = \frac{1}{\Gamma(p)} \int_0^\infty e^{-x} x^{p-1+iw}\, dx = \Gamma(p+iw)/\Gamma(p).$$

Если p — целое (т. е. n — нечетное), то отсюда получаем

$$\varphi(w) = \frac{(p-1+iw)(p-2+iw)\ldots(1+iw)\,\Gamma(1+iw)}{(p-1)(p-2)\ldots 1\,\Gamma(1)} =$$

$$= \Gamma(1+iw) \prod_{s=1}^{p-1} \left(1 + \frac{iw}{s}\right),$$

так что производящая функция семиинвариантов равна

$$\psi(w) = \log\Gamma(1+iw) + \sum_{s=1}^{p-1} \log\left(1 + \frac{iw}{s}\right). \tag{37.18}$$

Далее $\Gamma(1-iw)$ является х. ф. распределения крайних значений (14.66) с семиинвариантами

$$\begin{aligned} \varkappa_1 &= \gamma \quad \text{(постоянная Эйлера, 0,577 \ldots)}, \\ \varkappa_r &= (r-1)! \sum_{s=1}^{\infty} s^{-r}, \qquad r \geqslant 2, \end{aligned} \tag{37.19}$$

(см. упражнение 14.21). Отсюда и из (37.18) для семиинвариантов величины $\log z$ получаем

$$\lambda_r = \left[\frac{\partial^r \psi(w)}{\partial (iw)^r}\right]_{w=0} = (-1)^r \varkappa_r + \sum_{s=1}^{p-1} (-1)^{r-1} (r-1)!\, s^{-r}. \tag{37.20}$$

Подстановка (37.19) в (37.20) дает

$$\lambda_1 = \sum_{s=1}^{p-1} s^{-1} - \gamma,$$
$$\lambda_r = (-1)^r (r-1)! \sum_{s=p}^{\infty} s^{-r}, \quad r \geqslant 2. \tag{37.21}$$

Таким образом, асимптотически, когда p растет, принимая только целые значения,

$$\gamma_1 = \lambda_3/\lambda_2^{3/2} = -2\sum_p^{\infty} s^{-3} \Big/ \left(\sum_p^{\infty} s^{-2}\right)^{3/2} \sim -\frac{1}{p^2}\Big/\left(\frac{1}{p}\right)^{3/2} = -p^{-1/2},$$
$$\gamma_2 = \lambda_4/\lambda_2^2 = 6\sum_p^{\infty} s^{-4} \Big/ \left(\sum_p^{\infty} s^{-2}\right)^{2} \sim \frac{2}{p^3}\Big/\left(\frac{1}{p}\right)^{2} = 2p^{-1}, \tag{37.22}$$

что иллюстрирует скорость сходимости распределения величины $\log t$ к нормальному.

Бартлетт и Кендалл (1946) табулировали среднее, дисперсию, γ_1 и γ_2 вплоть до значения $n = 20$ ($p = 9{,}5$), при котором асимптотическое приближение (37.22) становится достаточно точным.

37.12 Если понадобится, процедуру стабилизации дисперсии из **37.10** можно повторить. Пусть установлено, что для дисперсии $u(t)$ имеет место

$$\mathsf{D}\{u(t)\} = c\left\{1 + \frac{p(\theta)}{n}\right\} + o(n^{-1}), \tag{37.23}$$

т. е. выполняется (37.10). Будем искать приближение второго порядка, т. е. переменную $v(u)$ такую, что

$$\mathsf{D}\{v(u)\} = d\{1 + O(n^{-2})\}. \tag{37.24}$$

Так же как в (37.11), имеем в соответствии с (37.23) приближенное равенство

$$\mathsf{D}\{v(u)\} = \left(\frac{dv(u)}{du}\right)_\theta^2 \mathsf{D}(u) = \left(\frac{dv(u)}{du}\right)_\theta^2 c\left\{1 + \frac{p(\theta)}{n}\right\}.$$

Используя (37.12), получим

$$\mathsf{D}\{v(u)\} = \left(\frac{dv}{du}\right)_\theta^2 \left(\frac{du}{dt}\right)_\theta^2 D_n^2(\theta)\left\{1 + \frac{p(\theta)}{n}\right\} =$$
$$= \left(\frac{dv}{dt}\right)_\theta^2 D_n^2(\theta)\left\{1 + \frac{p(\theta)}{n}\right\}. \tag{37.25}$$

Далее, так же как в (37.14),

$$v(t) \propto \left\{\int \left[D_n^2(\theta)\left\{1 + \frac{p(\theta)}{n}\right\}\right]^{-1/2} d\theta\right\}_{\theta=t}. \tag{37.26}$$

Мы уже сталкивались с примером такой процедуры при рассмотрении в (16.81) т. 1, улучшения Хотеллинга z-преобразования Фишера, стабилизирующего дисперсию (см. упражнения 16.18—19 и ниже пример 37.3).

Если есть необходимость, то процедуру стабилизации дисперсии можно итерировать и дальше.

В упражнениях 37.4—6 приведены другие примеры применения метода **37.10** нахождения преобразования, стабилизирующего дисперсию (в первом приближении) для биномиального и отрицательного биномиального распределений.

Нормализующие преобразования

37.13 В **6.25—6** мы уже рассматривали метод Корниша — Фишера получения нормализующих полиномиальных преобразований, а в **6.27—35** обсуждалась система функциональных преобразований Джонсона.

Кёртисс (1943) приводит подробное математическое рассмотрение предельной нормальности преобразований, в частности тех, которые обсуждаются в наших примерах и упражнениях. Мы приведем теперь несколько примеров того, что преобразования, предназначенные для достижения одной цели (здесь стабилизация дисперсии), как упоминалось в **37.9**, часто помогают в достижении другой (здесь нормализация). Кроме того, в упражнении 37.16 показывается, что в случае, изложенном в примере 37.1, имеет место тот же эффект. Однако последний из приводимых примеров показывает, что такая гармония целей получается только в том случае, если не требовать оптимальности в решении каждой задачи. Преобразования, стабилизирующие дисперсию, обычно приводят заодно и к нормализации, но при этом не получается *оптимальное* приближение.

Пример 37.3

В **16.33** мы рассматривали стабилизирующее дисперсию преобразование Фишера коэффициента корреляции r. Как следует из (16.74), для дисперсии величины r справедливо

$$D_n^2(\rho) = \frac{(1-\rho^2)^2}{n}\left(1 + \frac{11\rho^2}{2n}\right) + O(n^{-3}), \tag{37.27}$$

где ρ — корреляционный параметр совокупности. Применяя (37.14) к главному члену в (37.27), получим $z(r) = \frac{1}{2}\log\{(1+r)/(1-r)\}$, и, как показано в (16.77), для дисперсии z имеем

$$\mathsf{D}(z) = \frac{1}{n-1}\left\{1 + \frac{4-\rho^2}{2(n-1)}\right\} + (O(n^{-3}), \tag{37.28}$$

т. е. зависимость от ρ очень слабая, а значит, стабилизация дисперсии хорошая. В (16.78) показано, что коэффициент асимметрии γ_1 величины z имеет порядок $n^{-3/2}$ против $n^{-1/2}$ для r, а γ_2 для обеих величин имеет порядок n^{-1}.

Отсюда понятно, что стабилизация дисперсии симметризует и заодно нормализует.

Применяя далее (37.26), получим

$$z^* = z - (3z + r)/4n \tag{37.29}$$

с дальнейшей стабилизацией дисперсии вплоть до $(n-1)^{-1} + O(n^{-3})$.

Пример 37.4

Вернемся к примеру 37.2, где, согласно (37.22), для величины, получающейся в результате стабилизирующего дисперсию логарифмического преобразования, имели место асимптотические равенства

$$\gamma_1 = -p^{-1/2}, \quad \gamma_2 = 2p^{-1}. \tag{37.30}$$

Из (16.6) следует, что для непреобразованной величины при $p = \nu/2$ выполнялось

$$\gamma_1 = 2p^{-1/2}, \quad \gamma_2 = 6p^{-1}, \tag{37.31}$$

так что при стабилизации дисперсии заодно в два раза уменьшается асимметрия (с изменением знака) и в три раза эксцесс.

Пример 37.5

Предположим теперь, что в примере 37.2 известно σ^2, а p является параметром. Тогда $\mathsf{M}(z) = \mathsf{D}(z) = p$, и мы снова в той же ситуации, что и для распределения Пуассона из примера 37.1: соотношение (37.14) приводит к извлечению квадратного корня. Как показывают приведенные ниже значения, полученные Бартлеттом (1936),

p	0	0,5	1,0	2,0	3,0	4,0	9,0	15,0
$\mathsf{D}(p^{1/2})$	0	0,182	0,215	0,213	0,239	0,242	0,247	0,250

стабилизация дисперсии хорошая. Это эквивалентно рассмотренному в **16.5—6** приближению Фишера для χ^2-распределения. В соответствии с (16.8) для приближения Фишера имеем

$$\gamma_1 = \frac{1}{2}p^{-1}, \quad \gamma_2 = \frac{3}{16}p^{-2}, \tag{37.32}$$

т. е. определенное улучшение по сравнению с (37.31) для непреобразованной величины (а также лучше, чем (37.30) для логарифмического преобразования, которое не может быть обосновано в данном случае). Как и ранее, стабилизация дисперсии приводит к лучшему приближению нормальным распределением. Заметим, однако, что при нормализации Уилсона — Хилферти из **16.7** получаем, согласно (16.13), даже лучшее значение γ_1, порядка $p^{-3/2}$, но не такое хорошее γ_2, порядка p^{-1}. Как сказано в **16.8**, преобразование Уилсона — Хилферти дает лучшее приближение к нормальному распределению, хотя, как следует из (16.12), оно вообще не стабилизирует дисперсию. Таким образом, лучшее из рассмотренных нормализующих преобразований не стабилизирует дисперсию, а извлечение квадратного корня является наилучшим компромиссом.

37.14 Читатель видит, что наше обсуждение нормализации формулировалось, по существу, в терминах эксцесса и асимметрии. Математически очень большое допущение предполагать, что малым значениям γ_1 и γ_2 соответствует лучшее нормальное приближение. Тем не менее нет ни одного убедительного примера, в котором это предположение привело бы нас к неверному выбору лучшего из имеющихся нормальных приближений.

Блом (1954) разыскивает функциональное преобразование $u(t)$ такое, что последующее применение полиномиального преобразования Корниша — Фишера из **6.25—6** имеет минимальные эксцесс и асимметрию, а значит, является, по-видимому, «ближайшим» к симметричному и нормальному. Для $u(t)$ получается дифференциальное уравнение, которому удовлетворяют все преобразования, с которыми мы столкнулись при изучении пуассоновского, гамма, биномиального и отрицательного биномиального распределений.

Преобразования, ведущие к аддитивности

37.15 На практике может быть важно найти шкалу, в которой эффекты аддитивны (т. е. взаимодействия отсутствуют) или близки к этому. Однако по сравнению с нормализацией и стабилизацией дисперсии в этом направлении сделано относительно мало. Некоторые из предложенных общих процедур основаны на минимизации внутри класса преобразований значения статистики критерия, используемого для проверки гипотезы о том, что взаимодействия равны нулю. Например, в двухфакторной перекрестной классификации можно минимизировать S_3 (или S_3/S_R) из (35.65) для сбалансированного случая, или (35.69), когда в каждой клетке имеется по одному наблюдению. Следует знать, что статистика критерия используется здесь для проведения сложной процедуры оценки, однако до сих пор имеется только интуитивное обоснование этого метода. Иногда на такие аддитивные преобразования наталкивает анализ остатков, который мы рассмотрим в **37.18—20**.

37.16 В этой главе мы не будем рассматривать другие, более специальные преобразования. Мы уже рассматривали (см. **31.39** т. 2) преобразования наблюдений посредством их ранжирования к стандартным нормальным порядковым статистикам или *нормальным меткам* (*normal scores*). В **31.71** говорилось об их использовании для получения свободного от распределений критерия для однофакторной классификации в ситуации ДА. *Пробит* и *логит* преобразования квантилей, особенно к нормальному и логистическому распределениям уклонений, возникают в основном в биологических контекстах. Они обсуждаются в работе Финни (1952).

Устранение вызванного преобразованием смещения

37.17 Каковы бы ни были цели преобразования, после того, как проведен анализ, часто возникает задача представления. На практике оценки средних или разностей, несмещенные в преобразованной шкале, не останутся таковыми, если сделать обратное преобразование, чтобы результат можно было представить в «естественных» терминах (см., например, упражнение 37.15). Чтобы устранить смещение, вызванное преобразованием, нужно провести некоторого рода корректировку. Общие методы уменьшения смещения приводились в **17.10**. Теперь мы рассмотрим один точный метод устранения смещения.

Пусть u нормально распределена со средним μ и дисперсией σ^2, и пусть функции $(\hat{\mu}, S^2/\nu)$ от результатов наблюдения дают в совокупности достаточную статистику для этих параметров. При этом $\hat{\mu}$ имеет нормальное распределение со средним μ и дисперсией $\lambda^2\sigma^2$, а S^2/σ^2 не зависит от $\hat{\mu}$ и имеет χ^2-распределение с ν ст. св. На практике обычно $\lambda^2 = n^{-1}$, а $\nu = n - 1$, где n — объем выборки. Рассмотрим теперь функцию $t(u)$, которая является в нашей терминологии обратным преобразованием. Нейман и Скотт (1960) (см. также последующую статью Шметтерера (1960)) показали, что если $t(u)$ удовлетворяет дифференциальному уравнению второго порядка

$$t''(u) = A + Bt(u),$$

где A и B — постоянные, то единственной несмещенной МД оценкой среднего для непреобразованной величины $\theta = \mathbf{M}(t)$ является

$$\hat{\theta} = \begin{cases} t(\hat{\mu}) + A(1-\lambda^2)S^2/(2\nu), & B = 0, \\ \{t(\hat{\mu}\} + A/B\} \sum\limits_{r=0}^{\infty} \dfrac{\Gamma(\nu/2)}{r!\,\Gamma\left(\frac{1}{2}\nu + r\right)} \left\{\dfrac{B(1-\lambda^2)S^2}{4}\right\}^r - \dfrac{A}{B}, & B \neq 0, \end{cases}$$

при этом ряд сходится очень быстро и для получения достаточной точности требуется обычно лишь небольшое число членов.

Отсюда следует, что смещение грубой оценки $t(\hat{\mu})$, являющейся просто обратным преобразованием $\hat{\mu}$, равно

$$\mathbf{M}\{t(\hat{\mu}) - \theta\} = \begin{cases} -A(1-\lambda^2)\sigma^2/2, & B = 0, \\ \{\theta + (A/B)\}[\exp\{-B(1-\lambda^2)\sigma^2/2\} - 1], & B \neq 0, \end{cases}$$

т. е. абсолютная величина этого смещения является всегда монотонно убывающей функцией λ^2. Поскольку обычно $\lambda^2 = 1/n$, то смещение *возрастает* с ростом выборки.

Ниже приведена таблица для самых важных частных случаев.

Преобразование $u(t)$	Обратное преобразование $t(\hat{\mu})$	A	B	Смещение $\mathbf{M}\{t(\hat{\mu})\} - \theta$	Знак смещения при $\lambda < 1$
$(t+c)^{1/2}$	$\hat{\mu}^2 - c$	2	0	$-(1-\lambda^2)\sigma^2$	< 0
$\log(t+c)$	$\exp(\hat{\mu}) - c$	c	1	$(\theta + c)[\exp\{-(1-\lambda^2)\times\sigma^2/2\} - 1]$	$-\operatorname{sign}(\theta + c)$
$\arcsin(t^{1/2})$	$\sin^2(\hat{\mu})$	2	-4	$\left(\theta - \frac{1}{2}\right)\times[\exp\{2(1-\lambda^2)\sigma^2\} - 1]$	$\operatorname{sign}\left(\theta - \frac{1}{2}\right)$
$\operatorname{arsh}(t^{1/2})$	$\operatorname{sh}^2(\hat{\mu})$	2	4	$\left(\theta + \frac{1}{2}\right)\times[\exp\{-2(1-\lambda^2)\sigma^2\} - 1]$	$-\operatorname{sign}\left(\theta + \frac{1}{2}\right)$

Видно, что при извлечении квадратного корня при $\lambda \to 0$ ($n \to \infty$) смещение стремится к $-\sigma^2$. Этот результат получен непосредственно в упражнении 37.15, где, как и в (**37.16**), $\sigma^2 = 1/4$. Читатель может вспомнить, что результат о смещении для логарифмического преобразования содержался в упражнении 18.7 т. 2. Два оставшихся преобразования из таблицы рассматриваются в упражнениях 37.4—5.

Анализ остатков

37.18 Энскомб (1961) и Энскомб и Тьюки (1963) разработали методы, полезные при изучении отклонений от постулируемой линейной модели, а также подсказывающие степенные преобразования, уменьшающие эти отклонения.

Ограничимся моделью с генеральным средним (скажем, θ_0), которую можно записать в виде

$$\boldsymbol{y} = \mathbf{1}\theta_0 + \boldsymbol{W}\theta + \varepsilon, \qquad (37.33)$$

где $\mathbf{1}$ — вектор, все координаты которого равны единице и имеющий, как и $\boldsymbol{\varepsilon}$ и $\boldsymbol{y}$ порядок $(N \times 1)$, $\boldsymbol{W}$ — является $(N \times p)$-матрицей, а $\boldsymbol{\theta}$ — вектор порядка $(p \times 1)$. Генеральное среднее θ_0 вводится для того (см. упражнение 19.1), чтобы заменить в НК-оценках параметров $\boldsymbol{\theta}$ элементы y_i вектора $\boldsymbol{y}$ на уклонения $z_i = y_i - \bar{y}$, которые образуют вектор

$$\boldsymbol{z} = \boldsymbol{y} - \mathbf{1}\boldsymbol{y}'\mathbf{1}/n, \tag{37.34}$$

а также чтобы заменить элементы w_{ij} матрицы $\boldsymbol{W}$ на отклонения от средних по столбцам

$$x_{ij} = w_{ij} - \bar{w}_j,$$

которые образуют матрицу $\boldsymbol{X}$. Таким образом,

$$\boldsymbol{z}'\mathbf{1} = \boldsymbol{X}'\mathbf{1} = \mathbf{0}. \tag{37.35}$$

А значит, без потери общности с самого начала можно предполагать, что модель имеет вид

$$\boldsymbol{y} = \mathbf{1}\theta_0 + \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \tag{37.36}$$

где выполняется (37.35). Тогда НК-оценкой θ_0 является $\bar{y}$, а НК-оценкой $\boldsymbol{\theta}$ в соответствии с (19.12)

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{z}.$$

Положим $\boldsymbol{M} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ и обозначим вектор подогнанной модели через

$$\boldsymbol{f} = \boldsymbol{X}\hat{\boldsymbol{\theta}} = \boldsymbol{M}\boldsymbol{z},$$

а вектор остатков от подогнанной модели — через

$$\boldsymbol{r} = \boldsymbol{z} - \boldsymbol{f} = (\boldsymbol{I} - \boldsymbol{M})\boldsymbol{z}.$$

В соответствии с (37.35)

$$\boldsymbol{r}'\mathbf{1} = 0, \tag{37.37}$$

и так как $\boldsymbol{M}$ идемпотентна, то также

$$\boldsymbol{M}\boldsymbol{r} = \boldsymbol{M}(\boldsymbol{I} - \boldsymbol{M})\boldsymbol{z} = \mathbf{0}. \tag{37.38}$$

37.19 Предположим теперь, что после подгонки модели мы построили диаграмму разброса (см. пример 26.7), откладывая по оси абсцисс подогнанные значения, а по оси ординат соответствующие остатки. В соответствии с (37.37) среднее значение ординат равно нулю. Далее, поскольку $\boldsymbol{M}$ симметрична, то по (37.38) имеем $\boldsymbol{f}'\boldsymbol{r} = (\boldsymbol{M}\boldsymbol{z})'\boldsymbol{r} = \boldsymbol{z}'\boldsymbol{M}\boldsymbol{r} = 0$, так что подогнанные значения не коррелированы с остатками, т. е. линия регрессии на диаграмме разброса находится в правом верхнем координатном углу (см. **26.9**).

Помимо этих двух общих свойств диаграммы разброса, можно использовать и другие ее свойства для проверки того, насколько хорошо выполняются предположения подогнанной модели. На практике предположение гомоскедастичности можно грубо проверять в терминах разброса остатков для различных подмножеств подогнанных значений и тем же способом проверять предположения нормальности, особенно при рассмотрении асимметрии. Во всех случаях, если будет найдено, что предположения не выполняются, то можно сделать соответствующее преобразование методами, рассмотренными выше в этой главе.

Однако, по-видимому, самым интересным использованием диаграммы разброса является проверка предположения аддитивности в многофакторных экспериментах. Неаддитивность может проявиться в очевидной нелинейности (скажем, квадратичности) регрессии остатков от подогнанных значений.

37.20 Приведенные выше грубые визуальные методы можно переформулировать в более точных терминах. Энскомб (1961) предложил использовать статистики

$$t_{pq} = \boldsymbol{r}_p'\boldsymbol{f}_q,$$

где $\boldsymbol{r}_p$ — вектор p-х степеней остатков (например, определенный выше $\boldsymbol{r}$ совпадает с $\boldsymbol{r}_1$), а $\boldsymbol{f}_q$ — вектор q-х степеней подогнанных значений $(\boldsymbol{f} = \boldsymbol{f}_1)$. Статистики t_{30} и t_{40} являются очевидными аналогами мер асимметрии и эксцесса. Статистика t_{21} измеряет гетероскедастичность, поскольку, по существу, она дает значение ковариации квадрата остатков с подогнанными значениями. В соответствии с концом **37.19** статистика t_{12} является мерой неаддитивности. На самом деле она близко связана со статистикой S_I, используемой в (35.69) для проверки аддитивности в двухфакторной перекрестной классификации с одним наблюдением в клетке. Числитель S_I равен просто t_{12}^2.

37.21 Энскомб (1961) развил приближенную выборочную теорию для t_{pq} (нормированных подходящим образом). Эта теория ведет к приближениям степенными преобразованиями, необходимыми для устранения соответствующего отклонения от предположений модели. В соответствии с тем, что говорилось в **37.9**, нет гарантии, что все эти статистики приведут к *одному и тому же* степенному преобразованию, но есть надежда, что они будут мало отличаться. В этой связи интересно, что Бокс и Д. Кокс (1964) (см. также комментарии Энскомба к их статье) выразили решение метода МП из (37.7) для степенного преобразования приближенно в терминах t_{pq} с $p + q = 3$ и 4. По существу МП-оценка приводит к некоему методу усреднения различных степенных преобразований, получающихся из

измерений гетероскедастичности, отклонения от нормальности и неаддитивности. Однако отсюда вовсе не следует, что подход по методу МП автоматически приводит к нужному преобразованию. Наоборот, он может затруднить выбор компромиссного варианта.

Устойчивость процедур ДА

37.22 Рассмотрим сначала оценки параметров в задачах ДА. При изучении Модели I ДА теория НК-оценок и их оптимальных свойств (установленных в **19.4—9** т. 2) вообще не опирается на предположение нормальности ошибок. Таким образом, при отсутствии нормальности все оценки, а также выражения для их дисперсий остаются верными. В этом отношении НК-оценки свободны от распределения. Предположение нормальности в **24.27—37** потребовалось только для целей проверки гипотез и для интервального оценивания.

Далее, если даже основная модель (19.8) метода НК неверна в смысле предположений о некоррелированности и гомоскедастичности дисперсий, это не приводит к смещению НК-оценок (19.12), поскольку (19.13—14) выполняется, если только ошибки имеют нулевые средние. Однако свойства МД теряются в этом случае, поскольку тогда они имеют место для истинных НК-оценок (19.59). Таким образом, гетероскедастичность и корреляция ошибок только уменьшают эффективность, но не приводят к смещению. Легко видеть, что для модели II ДА и других моделей, рассмотренных в главе 36, математические ожидания средних квадратов не меняются при нарушении предположения нормальности, так что оценки компонент дисперсии остаются несмещенными и в том случае, когда различные случайные величины не имеют нормального распределения. Однако дисперсии этих оценок (например, в простейшем случае, разобранном в упражнении 36.10) существенно меняются при отсутствии нормальности. Это происходит потому, что соответствующие выражения больше не упрощаются благодаря специальным соотношениям между семиинвариантами нормального распределения.

37.23 При рассмотрении критериев (и соответствующих интервальных оценок) мы уже отметили в **31.2—9** т. 2 замечательное общее свойство влияния отсутствия нормальности на процедуры, полученные в предположении нормальности. Критерии о средних значениях являются устойчивыми, а критерии о дисперсиях — нет. Исходя из этого общего свойства, можно ожидать, что критерии и интервальные оценки в модели I, которые, по существу, имеют дело с математическими ожиданиями, будут относительно устойчивы при нарушении нормальности. В то же время для модели II и других моделей ДА, в которых рассматриваются дисперсии, устойчивости не будет. В **37.24—35** мы детально рассмотрим устойчивость к отсутствию нормальности, а здесь только отметим, что соответствующие утверждения существенно подтверждаются в некоторых детальных исследованиях, очень подробный обзор которых приведен в последней главе книги Шеффе (1959). Следует обратить внимание также на более ранний обзор Кокрэна (1947).

В этих исследованиях (например, Хорснелл (1953), Бокс (1954)) показано также, что влияние гетероскедастичности ошибок, вообще говоря, может быть велико, однако оно несущественно, если во всех клетках классификации частоты равны. (Мы сталкивались уже с этим явлением в простейшем виде в **21.24** (см. также **31.4**).) Практический вывод состоит в том, что при планировании наблюдений на основании одного только свойства устойчивости нужно там, где это возможно, использовать равные частоты. (Удачное совпадение, что после такого вывода расчеты становятся существенно проще.) Далее эта устойчивость по отношению к гетероскедастичности для сбалансированного случая позволяет нам и в случае неравных частот в клетках получить простые приближения для ДА для средних значений в клетках (см. упражнения 37.7—8).

Влияние стохастической зависимости ошибок может быть большим (Бокс (1954)). Это перекликается с общим утверждением из **36.39—43** о том, что при анализе следует принимать во внимание методы рандомизации размещения экспериментов (которые, очевидно, могут внести зависимость в ошибки). Мы вернемся к этим методам в главе 38.

Устойчивость к отсутствию нормальности в линейной модели

37.24 Интересный подход Бокса и Ватсона (1962), развивающий идеи более ранней работы Бокса и Андерсена (1955), проливает свет на причины разной степени устойчивости к отсутствию нормальности.

Вернемся к линейной модели с генеральным средним, определенной в **37.18**. Здесь СК, соответствующая подогнанной модели, равна

$$S_0 = \boldsymbol{f}'\boldsymbol{f} = \boldsymbol{z}'\boldsymbol{M}\boldsymbol{z},$$

а остаточная СК —

$$S_R = \boldsymbol{r}'\boldsymbol{r} = \boldsymbol{z}'(\boldsymbol{I} - \boldsymbol{M})\boldsymbol{z}.$$

Если ошибки имеют нормальное распределение, то критерий ОП для проверки гипотезы $\boldsymbol{\theta} = \boldsymbol{0}$ основывается на статистике

$$F = (S_0/p)/\{S_R/(N - p - 1)\},$$

имеющей распределение дисперсионного отношения с числом ст. св. $\nu_1 = p$, $\nu_2 = N - p - 1$. В эквивалентной форме критерий можно строить, исходя из

$$W = \frac{S_0}{S_0 + S_R} = \frac{z'Mz}{z'z}, \tag{37.39}$$

поскольку

$$W = \left(1 + \frac{\nu_2}{\nu_1 F}\right)^{-1} \tag{37.40}$$

является монотонно возрастающей функцией F. В нормальном случае, когда $\theta = 0$, величина $1/F$ имеет распределение дисперсионного отношения с $N - p - 1$, p ст. св., а W имеет β-распределение с параметрами $\{(N - p - 1)/2, p/2\}$, которое получается из $1/F$ преобразованием, описанным в **16.19**.

Рассмотрим теперь распределение величины W в общем (не обязательно нормальном) случае. Знаменатель W инвариантен относительно перестановок элементов вектора z. Рассмотрим поэтому сначала *перестановочное распределение* величины W (см. **31.16**). Если совместное распределение элементов вектора z симметрично по N своим аргументам, что выполняется, в частности, для независимых и одинаково распределенных ошибок, то любая перестановка элементов имеет вероятность $(N!)^{-1}$.

После того как мы получим среднее и дисперсию величины W для этого перестановочного распределения (обозначим их $\mathsf{M}_P(W)$ и $\mathsf{D}_P(W)$), мы сумеем получить безусловные среднее $\mathsf{M}(W)$ и дисперсию $\mathsf{D}(W)$, зная исходное распределение генеральной совокупности для z.

37.25 Поскольку $z'Mz$ — скаляр, то, используя, что след произведения не зависит от порядка сомножителей, получим

$$z'Mz = \operatorname{tr}(z'Mz) = \operatorname{tr}(Mzz').$$

Отсюда и из (37.39) имеем

$$z'z\mathsf{M}_P(W) = \mathsf{M}_P\{\operatorname{tr}(Mzz')\} = \operatorname{tr}\{M\mathsf{M}_P(zz')\}.$$

Далее

$$\mathsf{M}_P(zz') = \frac{1}{N(N-1)} z'z(NI - 11'), \tag{37.41}$$

поскольку

$$\mathsf{M}_P(z_j^2) = z'z/N, \qquad \mathsf{M}_P(z_j z_l) = -z'z/\{N(N-1)\}$$

для $j \neq l$. Используя (37.41), получаем

$$\mathsf{M}_P(W) = \frac{1}{N(N-1)} \operatorname{tr}\{M(NI - 11')\} = \frac{1}{N-1} \operatorname{tr}(M) = p/(N-1), \tag{37.42}$$

поскольку в соответствии с (37.35) $M11' = 0$, а из **19.9** $\operatorname{tr}(M) = p$.

Из (37.42) следует, что $\mathsf{M}_P(W)$ не зависит от X и от z. Таким образом, при любом распределении ошибок, скажем f, безусловное среднее величины W также равно

$$\mathsf{M}(W) = \underset{f}{\mathsf{M}}\{\mathsf{M}_P(W)\} = p/(N-1). \tag{37.43}$$

В частности, это справедливо в нормальном случае, а значит, математическое ожидание W вполне устойчиво к отклонениям от нормальности.

37.26. Для получения $\mathsf{D}_P(W)$ найдем сначала $\mathsf{M}_P(W^2)$. Обозначая через M_{rs} элементы матрицы M, получим

$$z'Mz = \sum z_r^2 M_{rr} + \sum z_r z_s M_{rs}, \tag{37.44}$$

где сумма, как и далее, берется по всем возможным не равным значениям индексов. Возводя (37.44) в квадрат и беря математическое ожидание по всем перестановкам элементов z, получим

$$\begin{aligned}\mathsf{M}_P\{(z'Mz)^2\} = {} & \mathsf{M}_P(z_r^4)\sum M_{rr}^2 + \mathsf{M}_P(z_r^3 z_s)\, 4\sum M_{rr}M_{rs} + \\ & + \mathsf{M}_P(z_r^2 z_s^2)\left(2\sum M_{rs}^2 + \sum M_{rr}M_{ss}\right) + \\ & + \mathsf{M}_P(z_r z_s^2 z_t)\left(4\sum M_{rs}M_{st} + 2\sum M_{rr}M_{st}\right) + \\ & + \mathsf{M}_P(z_r z_s z_t z_u)\sum M_{rs}M_{tu}.\end{aligned} \tag{37.45}$$

Как и в (12.8), обозначим s_r сумму r-х степеней z (так что $s_1 = 0$ по (37.35), а $z'z = s_2$). Воспользовавшись (12.9) т. 1, применим теперь для подсчета математических ожиданий в (37.45) соотношения между расширенными симметрическими функциями и суммами степеней. Согласно таблице 10 дополнения (вес 4), получим *)

$$\begin{aligned} N\mathsf{M}_P(z_r^4) &= s_4, \\ N^{(2)}\mathsf{M}_P(z_r^3 z_s) &= -s_4, \\ N^{(2)}\mathsf{M}_P(z_r^2 z_s^2) &= s_2^2 - s_4, \\ N^{(3)}\mathsf{M}_P(z_r z_s^2 z_t) &= 2s_4 - s_2^2, \\ N^{(4)}\mathsf{M}_P(z_r z_s z_t z_u) &= 3s_2^2 - 6s_4. \end{aligned} \tag{37.46}$$

*) С помощью $N^{(k)}$ автор обозначает число размещений из N элементов по k, т. е. $N^{(k)} = N(N-1)\ldots(N-k+1)$. (*Прим. ред*).

Более того, все суммы в (37.45), в которые входят M_{rs}, можно выразить в терминах $m = \sum_{r=1}^{N} M_{rr}^2$, используя идемпотентность M, значение p ее следа и тот факт, что, согласно (37.35), $M\mathbf{1} = 0$. В результате получим

$$\left.\begin{aligned} \sum M_{rr}M_{rs} &= -m, \\ \sum M_{rs}^2 &= p - m, \\ \sum M_{rr}M_{ss} &= p^2 - m, \\ \sum M_{rs}M_{st} &= 2m - p, \\ \sum M_{rr}M_{st} &= 2m - p^2, \\ \sum M_{rs}M_{tu} &= p^2 + 2p - 6m. \end{aligned}\right\} \qquad (37.47)$$

Подставляя (37.46—7) в (37.45) и обозначая

$$k_2 = s_2/(N-1), \qquad k_4 = \{N(N+1)s_4 - 3(N-1)s_2^2\}/(N-1)^{(3)}$$

(что совпадает, согласно (12.28), с k-статистиками игреков), получим, разделив на $(z'z)^2 = s_2^2$, вычитая $\{\mathbf{M}_P(W)\}^2$ и упрощая,

$$\mathbf{D}_P(W) = \frac{2p(N-p-1)}{(N+1)(N-1)^2} + \frac{k_4}{k_2^2}\frac{1}{(N-1)^2}\left\{m - \frac{p^2}{N} - \frac{2p(N-p-1)}{N(N+1)}\right\}. \qquad (37.48)$$

37.27 В отличие от (37.42), выражение (37.48) зависит от X через m и от z через отношение k_4/k_2^2. Поскольку мы нашли, что $\mathbf{M}_P(W) = \mathbf{M}(W)$, то

$$\mathbf{D}(W) = \mathbf{M}(W^2) - \{\mathbf{M}_P(W)\}^2,$$

в то время как

$$\mathbf{D}_P(W) = \mathbf{M}_P(W^2) - \{\mathbf{M}_P(W)\}^2,$$

так что, если f — распределение ошибок, то

$$\mathbf{D}(W) = \underset{f}{\mathbf{M}}\{\mathbf{D}_P(W)\}. \qquad (37.49)$$

Таким образом, безусловная дисперсия величины W зависит, по существу, от $\mathbf{M}(k_4/k_2^2)$. В нормальном случае

$$\mathbf{M}(k_4/k_2^2) = \mathbf{M}(k_4)/\mathbf{M}(k_2^2) = 0,$$

поскольку k_2 не зависит от k_4/k_2^2 (k_2 является полной достаточной статистикой, а k_4/k_2^2 свободно от распределения для скалярного параметра, а значит, применимо упражнение 23.7). Таким образом, первый член в правой части (37.48) является безусловной дисперсией в нормальной теории, и результат можно переписать в виде

$$\mathbf{D}_P(W) = \{\mathbf{D}(W)\}_{\text{норм}} \times \left[1 + C_y \frac{(N+1)}{2p(N-p-1)}\left\{m - \frac{p^2}{N} - \frac{2p(N-p-1)}{N(N+1)}\right\}\right], \qquad (37.50)$$

где $C_y = k_4/k_2^2$.

37.28 В упражнении 37.10 читателю предлагается показать, что m можно выразить в терминах отношений k-статистик $(k_4/k_2^2)_i$ и $\{k_{22}/(k_{20}k_{02})\}_{ij}$ для p регрессоров x. Используя этот результат, соотношения (37.50) можно переписать в виде

$$\mathbf{D}_P(W) = \{\mathbf{D}(W)\}_{\text{норм}}\left[1 + \frac{N-3}{2N(N-1)} C_y C_X\right], \qquad (37.51)$$

где

$$C_X = \frac{N(N^2-1)}{p(N-p-1)(N-3)}\left\{m - \frac{p^2}{N} - \frac{2p(N-p-1)}{N(N+1)}\right\} = \frac{(N-2)}{p(N-p-1)}\left\{\sum_{i=1}^{p}\left(\frac{k_4}{k_2^2}\right)_i + \sum_{\substack{i,\,j=1 \\ i\neq j}}^{p}\sum\left(\frac{k_{22}}{k_{20}k_{02}}\right)_{ij}\right\} \qquad (37.52)$$

и

$$\{\mathbf{D}(W)\}_{\text{норм}} = \frac{2p(N-p-1)}{(N+1)(N-1)^2}. \qquad (37.53)$$

Из (37.52) видно, что C_X является многомерным обобщением отношения k_4/k_2^2, характеризующего эксцесс*). При $p = 1$ C_X сводится к этому отношению. По тем же соображениям, что и для k_4/k_2^2, в **37.27**, в нормальном случае математическое ожидание C_X равно нулю.

37.29 Выражения для моментов перестановочного распределения (37.42) и (37.51) позволяют нам подогнать непрерывное распределение к дискретному перестановочному распределению величины W в духе **31.47**, выбрав бета-распределение с теми же средним и дисперсией. Как W, так и подогнанное распределение сосредоточены на интервале (0, 1). Кроме того, мы знаем (см. **37.24**), что бета-распределение является безусловным распределением W в нормальном случае. Это приводит к обоснованной надежде, что мы получим хорошую аппроксимацию и для общего перестановочного распределения.

*) О многомерных обобщениях характеристик асимметрии и эксцесса см. также статью Ю. Н. Тюрина «Проверка гипотезы о нормальности многомерной выборки большого объема» (журнал «Теория вероятностей и ее применения», т. XVIII, 3, 1973). *(Прим. ред.)*

Среднее и дисперсия бета-распределения с параметрами $\nu_2/2$, $\nu_1/2$ равны (согласно примеру 2.8)

$$\mu_1' = \frac{\nu_1}{\nu_1+\nu_2},$$

$$\mu_2 = \frac{2\nu_1\nu_2}{(\nu_1+\nu_2+2)(\nu_1+\nu_2)^2},$$

а значит,

$$\begin{aligned} \nu_2 &= \nu_1(1-\mu_1')/\mu_1', \\ \nu_1 &= 2\mu_1'(\mu_1'-\mu_1'^2-\mu_2)/\mu_2. \end{aligned} \tag{37.54}$$

Теперь μ_1' и μ_2 нужно приравнять соответственно $\mathsf{M}_P(W)$ и $\mathsf{D}_P(W)$.

Согласно (37.54) ν_2/ν_1 является функцией только μ_1', а $\mathsf{M}_P(W)$ постоянно и равно $p/(N-1)$. Отсюда следует, что для корректировки ν_2 (по сравнению с нормальным случаем) нужен тот же множитель, что и для ν_1. Подставляя в (37.54) вместо μ_1' и μ_2 значения $\mathsf{M}_P(W)$ и $\mathsf{D}_P(W)$ соответственно, получим

$$\nu_1 = p\left\{1+\frac{(N+1)c}{N-1-2c}\right\}^{-1}, \tag{37.55}$$

где c — корректирующий множитель из (37.51), т. е.

$$c = \frac{N-3}{2N(N-1)}C_yC_X. \tag{37.56}$$

Согласно сказанному выше для ν_2 также имеем

$$\nu_2 = (N-p-1)\left\{1+\frac{(N+1)c}{N-1-2c}\right\}^{-1}. \tag{37.57}$$

Если либо C_y, либо C_X равны нулю, то $c=0$, и в нашем приближении справедлива нормальная теория для $\mathsf{D}_P(W)$ при любом распределении ошибок.

37.30 Соотношения (37.49) и (37.51) показывают, что безусловная дисперсия $\mathsf{D}(W)$ совпадает с $\mathsf{D}_P(W)$, только C_yC_X надо заменить на $\mathsf{M}_f(C_yC_X)$. С этим изменением приближение **37.29** справедливо для безусловного распределения W так же, как для перестановочного распределения.

37.31 Поскольку параметры приближающего распределения, определяемые (37.55—7), зависят, по существу, от корректирующего множителя c, интересно определить границы его изменения. В упражнении 37.11 предлагается показать, что

$$p^2/N \leqslant m \leqslant p(N-1)/N, \tag{37.58}$$

а значит, из (37.52)

$$-2 \leqslant C_X(N-3)/(N-1) \leqslant N-1. \tag{37.59}$$

Отсюда видно, что если C_X достигает этих границ или близко к ним, то корректирующий множитель c в (37.56) имеет порядок N^{-1} вблизи нижней границы и порядок N^0 вблизи верхней. Это определяет, по крайней мере для больших выборок, можно ли пренебречь корректирующим множителем, т. е. будет ли распределение W устойчиво или нет. Поскольку в (37.52) мы видели, что отклонение C_X от нуля является мерой отклонения от нормальности иксов, то можно сказать, что нормальность регрессоров приводит к устойчивости при отключении от нормальности для W.

Устойчивость к отсутствию нормальности для однофакторного ДА

37.32 Применим теперь изложенные в **37.22—31** общие методы, справедливые для линейной модели, к частному случаю однофакторной классификации. Поскольку в (37.36) присутствуют $p+1$ параметров $(\theta_0, \boldsymbol{\theta})$, то мы предположим, что в классификации имеется $p+1$ группа. При этом модель с новыми параметрами из конца примера 35.1 теперь не будет особенной, поскольку отсутствует дополнительный параметр, приводящий к сингулярности.

Пусть в j-й группе имеется n_j наблюдений, $\sum_{j=1}^{p+1} n_j = N$. Каждый из p регрессоров должен просто указывать, принадлежит ли наблюдение заданной группе или нет. При этом принадлежность $(p+1)$-й группе следует из непринадлежности всем остальным. Обычно для этой цели мы должны были определить регрессоры как переменные, применяющие значения 0 или 1, но поскольку должно выполняться (37.36), то вместо этого мы положим

$$x_{ij} = \begin{cases} 1-\dfrac{n_j}{N}, & \text{если } y \text{ принадлежит } j\text{-й группе}, \\ -\dfrac{n_j}{N} & \text{в противном случае} \end{cases} \tag{37.60}$$

для $i=1, \ldots, N$ и $j=1, \ldots, p$. Тогда для каждого j, как и требовалось, $\sum_{i=1}^{N} x_{ij} = 0$.

Из примера 35.1 мы знаем, что в этом случае СК, соответствующая подогнанной модели, равна

$$\boldsymbol{z}'\boldsymbol{M}\boldsymbol{z} = \boldsymbol{y}'\boldsymbol{M}\boldsymbol{y} =$$

$$= \sum_{j=1}^{p+1} n_j(\bar{y}_j-\bar{y})^2 \equiv \sum_j \frac{\left(\sum_{i=1}^{n_j} y_{ij}\right)^2}{n_j} - \frac{\left(\sum_i\sum_j y_{ij}\right)^2}{N}, \tag{37.61}$$

где на время, вместо точки в индексе, указывающей на усреднение, мы написали черту. Рассматривая коэффициенты при y_{ij}^2 в (37.61), мы видим, что для каждого члена в j-й группе $M_{rr}=\frac{1}{n_j}-\frac{1}{N}$. А значит,

$$m=\sum_{r=1}^{N} M_{rr}^2=\sum_{j=1}^{p+1} n_j\left(\frac{1}{n_j}-\frac{1}{N}\right)^2=\sum_{j=1}^{p+1}\frac{1}{n_j}-\frac{2p+1}{N},$$

и (37.52) принимает вид

$$\frac{N-3}{N-1}C_X=\frac{N(N+1)}{p(N-p-1)}\left\{\sum_{j=1}^{p+1}\frac{1}{n_j}-\frac{(p+1)^2}{N}\right\}-2. \qquad (37.62)$$

Подставляя (37.62) в (37.51), получим $\mathbf{D}_P(W)$ для данного случая. (Соответствующее выражение впервые было получено Уэлчем (1937).) Очевидно, что значение C_X существенно зависит от того, как N наблюдений распределены между $p+1$ группой.

37.33 Если $n_j=N/(p+1)$, так что во всех группах равное число наблюдений, то (37.62) принимает вид

$$\frac{N-3}{N-1}C_X=-2, \qquad (37.63)$$

т. е. достигается нижняя граница в (37.59). Тогда (37.56) равно

$$c=-C_y/N,$$

и множитель при p и $(N-p-1)$ в (37.55) и (37.57) равен

$$\left(1+\frac{2C_y}{N(N-1)}\right)\Big/\left(1-\frac{C_y}{N}\right)\sim 1+\frac{C_y}{N},$$

т. е. им можно пренебречь для больших выборок.

Как следует из того факта, что значение $C_X=0$ удовлетворяет неравенствам (37.59), эта поправка, которой можно пренебречь, может быть сведена к нулю (см. упражнение 37.12), если должным образом выбрать неравные групповые частоты. Однако кажется неразумным получать это несущественное увеличение устойчивости к отсутствию нормальности, помня, что это приведет к потере устойчивости относительно гетероскедастичности, справедливой в сбалансированном случае (см. **37.21**).

37.34 В качестве противоположности для изученного в **37.33** случая равных частот, рассмотрим случай $n_1=n_2=\ldots=n_p=1$, $n_{p+1}=N-p$. Тогда (37.62) сводится к

$$\frac{N-3}{N-1}C_X=N-1-\frac{N+1}{N-p}, \qquad (37.64)$$

что очень близко к верхней границе в (37.59), если p/N мало. Тогда (37.56) принимает вид

$$c=\frac{C_y}{2N}\left(N-1-\frac{N+1}{N-p}\right)\sim\frac{C_y}{2}\left(1-\frac{1}{N}\right),$$

и множитель в (37.55) и (37.57) равен приблизительно $1+C_y/2$, указывая на крайнюю неустойчивость.

Но самое интересное свойство этого крайне противоположного случая проявляется после того, как мы подсчитаем СК (37.61), которая равна

$$\boldsymbol{y}'\boldsymbol{M}\boldsymbol{y}=\sum_{j=1}^{p}(y_{1j}-\bar{y})^2+(N-p)(\bar{y}_{p+1}-\bar{y})^2. \qquad (37.65)$$

Понятно, что когда N растет, то различием между $\bar{y}_{p+1}$ и $\bar{y}=\left\{\sum_{j=1}^{p} y_{1j}+(N-p)\,\bar{y}_{p+1}\right\}\Big/N$ можно пренебречь. Таким образом, с точностью до первого порядка по N (37.65) равно

$$\sum_{j=1}^{p}(\bar{y}_{1j}-\bar{y})^2\sim\sum_{j}(y_{1j}-y_.)^2+p(y_.-\bar{y}_{p+1})^2, \quad \text{где} \quad y_.=\sum_{j=1}^{p} y_{1j}/p.$$

F-критерий из **37.24** основывается на отношении этого выражения к

$$\sum_{j=1}^{p+1}\sum_{i=1}^{n_j}(y_{ij}-\bar{y})^2-\sum_{j=1}^{p}(y_{1j}-\bar{y})^2=$$
$$=\sum_{i=1}^{N-p}(y_{i,\,p+1}-\bar{y})^2\sim\sum_{i=1}^{N-p}(y_{i,\,p+1}-\bar{y}_{p+1})^2.$$

Таким образом,

$$F\sim\frac{\left\{\sum_{j=1}^{p}(y_{1j}-y_.)^2+p(y_.-\bar{y}_{p+1})^2\right\}/p}{\sum_{i=1}^{N-p}(y_{i,\,p+1}-\bar{y}_{p+1})^2/(N-p-1)}. \qquad (37.66)$$

За исключением члена, сравнивающего $y_.$ и $\bar{y}_{p+1}$, в числителе и соответствующего лишнего числа степеней свободы, выражение (37.66) является F-статистикой для проверки равенства дисперсий в двух нормальных генеральных совокупностях с размерами выборки p и $N-p$ соответственно (см. упражнение 23.14). В свете результатов этого параграфа легко понять крайнюю неустойчивость последнего критерия, о которой в общих чертах говорилось в **37.23** и более детально в **31.6—8**, где, по существу, тот же самый корректирующий множитель $(1+C_y/2)$ был получен для общих критериев о дисперсиях.

Устойчивость к отсутствию нормальности в сбалансированной классификации

37.35 Для сбалансированной однофакторной классификации в **37.33** соотношения (37.63) и (37.52) показывают, что

$$m = \sum_{r=1}^{N} M_{rr}^2 = p^2/N. \qquad (37.67)$$

Если выполняется (37.67), то в (37.59) достигается нижняя граница, и, так же как в **37.33**, получаем, что коррекцией по сравнению с нормальным случаем можно пренебречь. В частности, это справедливо, если

$$M_{rr} = p/N \qquad (37.68)$$

для всех r, т. е. если все диагональные элементы матрицы $\boldsymbol{M} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ равны между собой. Если линейная модель удовлетворяет (37.68), то говорят, что она *квадратично сбалансирована*. Из одного только соображения симметрии легко видеть, что (37.68) выполняется для любой перекрестной классификации с равными частотами в каждой клетке (ранее называвшейся просто «сбалансированной»), а также для иерархических классификаций с равными частотами на каждом уровне иерархии. Атикулла (1962), используя иной метод, получил это как асимптотический результат для безусловного распределения W. Его результат применим и к другим F-критериям, отличным от критерия проверки гипотезы $\boldsymbol{\theta} = \boldsymbol{0}$, в отличие от нашего изложения, ведущегося только для этого критерия.

При анализе ковариаций в сбалансированном однофакторном анализе, Атикулла (1964) показал, что в соответствии с выводами **37.31**, степень отклонения от нормальности сопутствующих переменных определяет устойчивость F-критерия к отсутствию нормальности ошибок. Он рассматривает также устойчивость по отношению к другим отклонениям от предположений.

Свободные от распределения методы ДА

37.36 Изучение устойчивости стандартных методов ДА естественно приводит нас к вопросу, можно ли указать вполне устойчивые методы анализа. Другими словами, существуют ли свободные от распределения методы для задач ДА.

В **31.70—4** мы уже видели, что при рассмотрении однофакторной классификации ответ положителен: существуют свободные от распределения критерии для проверки равенства параметров расположения k выборок из любых непрерывных генеральных совокупностей, имеющих одинаковое (с точностью, быть может, до параметров расположения) распределение. Критерий может основываться на самих рангах, используя статистику (31.150), или на нормальных метках $E(S, n)$ (см. **31.71** т. 2). Эти критерии вполне устойчивы для любого непрерывного распределения и имеют очень высокую асимптотическую относительную эффективность против нормальных альтернатив о параметре расположения (см. **31.71**). Другой критерий, против альтернатив об упорядочениях, был приведен в **31.72—4** (см. **35.66** и упражнение 35.15).

Куад (1967) приводит ранговый критерий для больших выборок при однофакторной классификации в ковариационном анализе.

37.37 Далее, рассмотренное в **37.24—31** перестановочное распределение W в общей линейной модели свободно от распределения также, как и рассмотренные в главе 31 критерии перестановок. Критерий для проверки $\boldsymbol{\theta} = \boldsymbol{0}$ можно рассматривать как критерий для проверки симметрии $\boldsymbol{z}$ по своим аргументам при любом рассматриваемом распределении ошибок. Однако, как мы видели в **37.32—5**, с этим критерием проверки гипотезы $\boldsymbol{\theta} = \boldsymbol{0}$ нельзя многого добиться в задачах ДА, за исключением однофакторной классификации. Теперь мы рассмотрим, насколько далеко можно использовать свободные от распределения методы в более сложных ситуациях ДА, если мы хотим проверять значения главных эффектов, взаимодействий и т. п.

Двухфакторная перекрестная классификация: критерий перестановок

37.38 Рассмотрим простейшую двухфакторную перекрестную классификацию с одним наблюдением в клетке. Пусть имеется r строк и c столбцов, так что всего $n = rc$ наблюдений, и пусть мы хотим проверить значения эффектов столбцов (или строк). В духе наших рассмотрений в **31.21** и **31.39**, самым естественным было бы заменить n наблюдений y_{ij} их рангами или каким-нибудь другим множеством условных чисел, таким, как нормальные метки, и применить к этим числам обычный критерий ДА. Трудно надеяться получить точные результаты в теории распределений для таких процедур, поскольку в этом случае нет множества равновероятных перестановок, исходя из которых можно было бы получить общие выводы.

37.39 Однако можно построить критерий перестановок для эффектов столбцов, из которого, как мы увидим, будут вытекать свободные от распределения критерии.

Обычная СК для проверки эффектов столбцов в ДА, скажем, $S_C = r\sum_{j=1}^{c}(y_{.j} - y_{..})^2$, инвариантна относительно добавления константы к каждому наблюдению в любой строке. Если считать среднее значение для каждой строки равным нулю, то

СК строк, скажем $S_R = c\sum_{i=1}^{r}(y_{i.} - y_{..})^2$ обратится в нуль, поскольку $y_{i.} = y_{..} = 0$. Но тогда S_C сводится к

$$S_C = r\sum_j y_{.j}^2. \tag{37.69}$$

Если теперь $S = \sum_i\sum_j (y_{ij} - y_{..})^2$, то стандартная F-статистика имеет вид

$$F = \frac{S_C/(c-1)}{(S - S_R - S_C)/\{(r-1)(c-1)\}},$$

с числом ст. св. $\nu_1 = c - 1$ и $\nu_2 = (r-1)(c-1)$. Так же, как и S_C, величина $S - S_R$ инвариантна относительно прибавления произвольных констант к строчкам, а значит, инвариантна и сама статистика F. Таким образом, без ограничения общности можно положить $S_R \equiv 0$, так что получим $F = (r-1)S_C/(S - S_C)$. Соответствующее бета-преобразование (см. (37.40)) равно

$$W = \left(1 + \frac{\nu_2}{\nu_1 F}\right)^{-1} = \frac{S_C}{S} =$$

$$= \frac{\sum_{j=1}^{c}\left(\sum_{i=1}^{r} y_{ij}\right)^2}{r\sum_i\sum_j y_{ij}^2} = \frac{1}{r}\left\{1 + \frac{2U}{(c-1)\sum_{i=1}^{r}(k_2)_i}\right\}, \tag{37.70}$$

где

$$U = \sum_{i=1}^{r}\sum_{\substack{l=1\\ i\neq l}}^{r}\sum_{j=1}^{c} y_{ij}y_{lj}$$

и $(k_p)_i$ есть p-я k-статистика значений игреков в i-й строчке $y_{i1}, y_{i2}, \ldots, y_{ic}$.

37.40 Моменты величины (37.70) можно найти теперь так же как мы это делали в (37.40) при предположении симметрии. Гипотеза состоит в том, что в классификации отсутствуют эффекты столбцов. В силу инвариантности F, а значит W, относительно прибавления констант к строчкам никаких предположений об эффектах строк делать не надо. Если гипотеза справедлива, то для каждой из r строк классификации, каждая из $c!$ перестановок величин y_{ij} $(j = 1, \ldots, c)$ равновероятна. Таким образом, при выполнении гипотезы имеется $(c!)^r$ равновероятных расположений y_{ij}. Случай $c = 2$ соответствует рассмотренному в **31.78** критерию двойной симметрии Фишера.

Питмэн (1938), в статье которого можно найти более детальное изложение, нашел первые четыре момента величины U, а значит, и W. Он показал, что

$$\mathbf{M}_P(W) = \frac{1}{c}. \tag{37.71}$$

Так же как и в (37.42—3), среднее перестановочного распределения (37.71) одно и то же при любых наблюдениях, а значит, совпадает с результатом нормальной теории. Выражение для дисперсии, так же как и в (37.52), более сложное и равно (см. упражнение 37.13)

$$\mathbf{D}_P(W) = \frac{2}{r^2(c-1)}\left[1 - \frac{\sum_i (k_2^2)_i}{\left\{\sum_i (k_2)_i\right\}^2}\right]. \tag{37.72}$$

Выражения для более высоких моментов еще сложнее.

Дисперсия в нормальной теории, как и в **37.29**, равна

$$\{\mathbf{D}(W)\}_{\text{норм}} = \frac{2\nu_1\nu_2}{(\nu_1+\nu_2+2)(\nu_1+\nu_2)^2} = \frac{2(r-1)}{r^2\{r(c-1)+2\}}, \tag{37.73}$$

и, так же как (37.49), это выражение является математическим ожиданием от (37.72) при условии нормальности. Из (37.72) и (37.73) можно получить, что в нормальном случае

$$\mathbf{M}\left[\frac{\sum_i (k_2^2)_i}{\left\{\sum_i (k_2)_i\right\}^2}\right] = \frac{c+1}{r(c-1)+2} = \frac{1 + \frac{2}{c-1}}{r + \frac{2}{c-1}}. \tag{37.74}$$

Число ст. св. для F-критерия можно выбрать, как и в **37.29**. В упражнении 37.4 содержится соответствующий пример. Питмэн (1938) показал, что если взять бета-распределение с теми же средним и дисперсией, то третий и четвертый моменты также хорошо согласуются.

37.41 Самым интересным (из интересующих нас сейчас) частным случаем W является случай, когда наблюдения y_{ij} заменены условными числами, такими, что критерий становится свободным от распределения. Вместо процедуры, о которой говорилось в **37.38**, мы заменим теперь y_{ij} *в каждой строчке отдельно* на множество условных чисел, например их рангами или нормальными метками. Если в каждой строчке использовать *одно и то же* множество, то непосредственно получим, что $(k_p)_i$ совпадают для всех значений i. Таким образом,

$$\frac{\sum_i (k_2^2)_i}{\left\{\sum_i (k_2)_i\right\}^2} = \frac{rk_2^2}{(rk_2)^2} = \frac{1}{r}. \tag{37.75}$$

Сравнивая (37.75) и (37.74), видим, что они мало отличаются, если r или c не слишком малы. Статистика свободного от распределения критерия, основанная на замене наблюденных значений на одно и то же множество условных чисел в каждой строке, имеет приблизительно то же самое распределение, что и для критерия в нормальной теории. В частности, это справедливо, если наблюдения заменить на ранги или нормальные метки.

Следует отметить, что если в критерии используются условные числа, то нет нужды полагать $y_{i.} = 0$, поскольку в любом случае $S_R \equiv 0$. Конечно, как и в упражнении 37.14, в этом случае $y_{..}$ (константу) нужно прибавить как к S_C, так и к S.

Ниже в 38.65 обсуждается АОЭ рангового критерия.

Более сложные классификации

37.42 Таким образом, при двухфакторной перекрестной классификации с одним наблюдением в клетке имеются свободные от распределения критерии ДА для проверки эффектов столбцов (или получающиеся транспонированием эффектов строк), не зависящие от наличия эффектов строк (столбцов). Ситуация является простой по той причине, что, как мы видели в **37.39**, полная СК (без учета генерального среднего) имеет три компоненты ($S_R, S_C, S - S_R - S_C$), из которых одну можно сделать равной нулю подходящим выбором начала координат. В задаче остается всего две случайные величины, и F является их отношением. При использовании условных чисел S тоже становится константой, так что бета-преобразование W является константой, умноженной на оставшуюся случайную величину S_C.

Если мы начнем рассматривать обобщения критерия перестановок, то эта простота исчезает. Даже в сбалансированной двухфакторной перекрестной классификации с более чем одним наблюдением в клетке полная СК имеет дополнительные компоненты (взаимодействия). Для трехфакторной перекрестной классификации с одним наблюдением в каждой клетке также появляются дополнительные компоненты. В обоих этих случаях (а, значит, и в более сложных ситуациях) трудно найти перестановочное распределение F и W статистик, и по-видимому, это является причиной того, что не развита теория этих критериев.

37.43 Альтернативный метод обобщения состоит в рассмотрении аналога одной из свободных от распределения статистик в более сложной ситуации. Например, в упражнении 37.14, где используются ранги, W просто пропорционально T, т. е., по существу, — дисперсии суммарных рангов столбцов. Для сбалансированной перекрестной двухфакторной классификации можно получить распределение этой дисперсии. По-видимому, его можно найти для случая неравных частот, а также для трехфакторной $r \times c \times l$ перекрестной классификации с $(c!)^{rl}$ равновероятными перестановками рангов столбцов внутри каждой строки и каждого слоя классификации. Насколько нам известно, ни одно из указанных обобщений не исследовалось. Для взаимодействий этим методом, по-видимому, нельзя получить свободные от распределения критерии.

Мэра и Саранджи (1967) исследовали асимптотическую теорию *выровненного* критерия для эффекта столбцов, предложенного Ходжесом и Леманом (1962). Вместо того чтобы ранжировать наблюдения отдельно в каждой строке, из наблюдений сначала вычитаются оценки эффектов строк, а затем полученные в разных строках величины объединяются и ранжируются в одной общей последовательности. При этом используется статистика, обобщающая (37.70) и в клетках $(r \times c)$-классификации могут быть не равные (но обязательно ненулевые) частоты. Этот критерий более эффективен в нормальном случае, особенно при малых значениях c.

Медианные критерии

37.44 Другой подход к построению свободных от распределения критериев ДА следует Брауну и Муду (1951). Принципы построения критериев, которые они использовали, состоят в следующем: (а) оценить все параметры, неучтенные гипотезой, с помощью медианной статистики, (б) проверить, будут ли остатки от оцененной медианами модели иметь половину знаков отрицательными и половину положительными.

Например, в однофакторной классификации с n_j наблюдениями в j-й группе, $\Sigma n_j = n$, при гипотезе о равенстве групповых средних только генеральное среднее μ не учитывается гипотезой. А значит, μ оценивается медианой $\tilde{y}$ из n (которое предполагается четным) наблюдений. Затем мы смотрим, как много наблюдений лежит в каждой группе выше и ниже $\tilde{y}$:

Группа		1	2	...	k		
Число наблюдений	$\geqslant \tilde{y}$	m_1	m_2	...	m_k	$n/2$	(37.76)
	$< \tilde{y}$	$n_1 - m_1$	$n_2 - m_2$	...	$n_k - m_k$	$n/2$	
Полное		n_1	n_2	...	n_k	n	

Теперь проверяется гипотеза о том, что все k групп имеют одну и ту же медиану, т. е. $\mathbf{M}(m_j) = n_j/2$ для каждого j. Сразу видно, что это биномиальный критерий однородности, рассмотренный в **33.55**. Статистика (33.122), которая в используемых теперь обозначениях имеет вид

$$X^2 = \sum_{j=1}^{k} \frac{(m_j - n_j/2)^2}{n_j/4}, \qquad (37.77)$$

имеет асимптотически χ^2-распределение с $k-1$ ст. св. Критерий (который можно точно сформулировать, используя метод **33.19**, случай 1) свободен от распределения. Что касается критерия знаков из **32.3**, то использование медиан сводит задачу к биномиальному случаю.

37.45 Более сложные ситуации ДА можно рассматривать с помощью того же общего метода. Как показывают упражнения 37.18—20, для двухфакторной перекрестной классификации имеется несколько критериев. Эффекты столбцов можно проверять для случая одного наблюдения в клетке или в более общей ситуации, но когда нет взаимодействий. Если имеются взаимодействия, то эффекты столбцов можно проверять против взаимодействий или проверять совместно эффекты столбцов и взаимодействия.

37.46 Эти медианные критерии привлекательны тем, что вычисления очень просты, а теория, по крайней мере для больших выборок, немедленно следует из $2 \times c$ таблиц контингенции. Однако не все задачи решаются с их использованием. Например, неизвестно критерия для проверки эффектов столбцов против остатков в общей сбалансированной перекрестной классификации. Далее, даже если имеется в распоряжении такого рода критерий, он не всегда свободен от распределения. Так, Браун и Муд (1951) показали, что медианный критерий для взаимодействий в сбалансированной двухфакторной классификации не является таковым. Далее, эффективность этих медианных критериев, вообще говоря, не выше, чем эффективность критериев, которые используют ранги или нормальные метки, если ошибки почти нормальны. В **32.6—7** было найдено, что критерий знаков имеет АОЭ, равную $2/\pi$ в нормальном случае, против $3/\pi$ для рангового критерия. В упражнении 31.12 показано, что медианный критерий случайности имеет АОЭ 0,78, в то время как для критерия из **31.38**, использующего ранги, имеем 0,98. Эндрюс (1954) показал, что рассмотренный в **37.44** критерий для однофакторной классификации имеет ту же самую АОЭ, равную $2/\pi$, что и критерий знаков, в то время как из **31.71** следует, что для конкурирующей статистики (31.150), основанной на рангах, АОЭ равна $3/\pi$.

Бхапкар (1963) приводит некоторые результаты об эффективности для медианных критериев при двухфакторной классификации.

37.47 Ограниченность применений и относительно низкая эффективность медианных критериев, полученных на основе принципов, изложенных в **37.44**, несколько разочаровывают. Интуитивно кажется, что, по-видимому, возможно указать общие процедуры ДА с высокой эффективностью, такой, которая, как мы видели в главе 31, характеризовала критерии, основанные на рангах. Лучшее приближение к такого рода процедурам было развито в серии работ Лемана (1963a, b, c, 1964) и Ходжеса и Лемана (1963) (см. также Хейланд (1965), Бикел (1965) и Адичи (1967a, b)). Эти процедуры только асимптотически свободны от распределения. Интересно, что и они основаны на методе медианных оценок, но иного рода, чем метод из **37.44**.

37.48 Пусть

$$x_{ip} = \mu_i + \varepsilon_{ip} \qquad (i = 1, \ldots, c;\ p = 1, \ldots, n_i)$$

— модель для множества из $n = \sum_{i=1}^{c} n_i$ наблюдений. ε_{ip} независимы, а в остальном мы предполагаем только, что они имеют одинаковое распределение. Обозначим $\theta_{ij} = \mu_i - \mu_j$ параметры, в терминах которых можно выразить все интересующие нас величины. Рассмотрим медианные оценки $\tilde{\theta}_{ij}$ величин θ_{ij}.

Пусть y_{ij} — медиана для $n_i n_j$ разностей $(x_{ip} - x_{jq})$, где $p = 1, \ldots, n_i$; $q = 1, \ldots, n_j$. Величины y_{ij}, очевидно, являются оценками для θ_{ij}, но они не обладают желаемым свойством транзитивности

$$\tilde{\theta}_{ij} + \tilde{\theta}_{jk} + \tilde{\theta}_{ki} = 0 \quad \text{для всех} \quad i, j, k. \qquad (37.78)$$

Усовершенствованной оценкой, удовлетворяющей (37.78), является

$$\tilde{\theta}_{ij} = y_{i.} - y_{j.}, \qquad (37.79)$$

где $y_{i.} = \frac{1}{c} \sum_{l=1}^{c} y_{il}$. Леман (1963a) показывает, используя числовые примеры, что $\tilde{\theta}_{ij}$ хорошо согласуется с обычными НК-оценками $\hat{\theta}_{ij}$.

Если $n \to \infty$ должным образом, то совместное распределение $\tilde{\theta}_{ij}$ сходится к многомерному нормальному. Эффективность оценок по сравнению со стандартными оценками ДА, основанными на средних значениях, такая же, как у критерия Вилкоксона по сравнению с t-критерием Стьюдента. Если f — общая

функция плотности величин ε_{ip}, то из (31.115) следует, что эффективность равна

$$12\sigma^2\left[\int_{-\infty}^{\infty}\{f(x)\}^2\,dx\right]^2 = k^2. \qquad (37.80)$$

Согласно **31.60—1** k^2 может равняться бесконечности, но для любого непрерывного распределения f не может быть меньше чем 0,864. В нормальном случае $k^2 = 3/\pi \doteqdot 0{,}95$. Итак, в общем случае $kn^{1/2}(\tilde{\theta}_{ij} - \theta_{ij})$ имеет то же самое предельное распределение, что и $n^{1/2}(\hat{\theta}_{ij} - \theta_{ij})$, где $\hat{\theta}_{ij}$ — стандартные НК-оценки ДА для θ_{ij}.

37.49 Из асимптотических свойств, сформулированных в последнем абзаце **37.48**, следует, что если мы сумеем дать состоятельную оценку k^2 из (37.80), то можно сформулировать аналоги всех обычных процедур ДА в терминах медианных оценок $\tilde{\theta}_{ij}$. Леман (1963c) дает две состоятельные оценки k^2 и приводит (1963b) доверительные интервалы для больших выборок для любого сравнения или даже множества сравнений параметров. Он же (см. также Ходжес и Леман (1962) и Доксум (1967)) распространяет эти результаты на ситуацию, когда имеется «мешающий фактор» у наблюдений (т. е. в терминологии главы 38 «блоки» внутри экспериментов), с равными числами наблюдений в каждой клетке одного и того же «блока». Двухфакторные классификации рассматриваются Пури и Сеном (1967b). Гринберг (1966) разработал теорию планирования неполных блоков, которая обсуждается ниже в главе 38 (см. также Пури и Сен (1967a)).

Бхучонгкул и Пури (1965) распространили асимптотическую теорию на класс оценок ограничений, включающий оценки, основанные на нормальных метках. П. Сен (1966) получил одновременные доверительные интервалы, аналогичные интервалам из **35.57—63**.

Потерянные наблюдения в общих линейных моделях

37.50 Преимущества сбалансированного устройства модели I ДА, а именно ортогональность, простота вычислений и лучшая устойчивость, приводят к тому, что при планировании анализа стараются воспользоваться этими преимуществами. Тем не менее сила обстоятельств приводит иногда к невольным отклонениям от планируемого равенства частот. Растения или животные могут погибнуть, люди могут оказать сопротивление при кооперировании, наконец, записи могут быть утеряны перед анализом. Если так произойдет, то мы всегда можем проводить анализ для полученных неравных частот соответствующими неортогональными методами, однако, как мы видели, эти методы часто сложны. Более того, случайные потери наблюдений редко бывают большими. Обычно теряется одно или несколько наблюдений. Поэтому стоит исследовать, можно ли сохранить исходную структуру ДА, скорректировав полученные данные с потерянными наблюдениями, а не просто отказываться совсем от этой структуры. Ниже приводятся рассуждения, справедливые для общей линейной модели, частным случаем которой является ДА.

37.51 Положим, что m из n планируемых наблюдений потеряны. Без ограничения общности будем считать, что они являются последними m компонентами вектора наблюдений $\boldsymbol{y}$. Обозначим их $u_1, \ldots, u_m$. Таким образом, можно положить $\boldsymbol{y} = \begin{pmatrix} \boldsymbol{z} \\ \boldsymbol{u} \end{pmatrix}$, где $((n-m)\times 1)$-вектор $\boldsymbol{z}$ состоит из действительно наблюденных значений вектора $\boldsymbol{y}$, а $(m \times 1)$-вектор $\boldsymbol{u}$ состоит из наблюдений, значения которых неизвестны. В результате мы сталкиваемся с необходимостью оценить дополнительное множество неизвестных в добавление к исходным параметрам модели. В этих обстоятельствах естественно оценивать значения $\boldsymbol{u}$ тем же самым методом НК, который мы использовали для исходных параметров.

37.52 Сумму квадратов остатков

$$S = (\boldsymbol{y} - \boldsymbol{X}\theta)'(\boldsymbol{y} - \boldsymbol{X}\theta)$$

нужно минимизировать теперь не только по переменным θ (как это делалось в **19.4**), но и по переменным $\boldsymbol{u}$. Если сначала минимизировать S по переменным θ, то получится исходное решение методом НК (т. е. решение, как если бы не было потерянных наблюдений), но как оценки, так и остаточная СК, полученные этим методом, являются теперь функциями $\boldsymbol{u}$. Обозначим их $\hat{\theta}(\boldsymbol{u})$ и $S_0(\boldsymbol{u})$. Процесс минимизации можно теперь завершить, взяв минимум $S_0(\boldsymbol{u})$ по $\boldsymbol{u}$.

Однако эта двухшаговая процедура минимизации, предложенная Иэйтсом (1933), вообще говоря, не является простейшей. Вместо этого минимизируем S сначала по $\boldsymbol{u}$. Разбивая $\boldsymbol{X}$ на $\begin{pmatrix} \boldsymbol{X}_z \\ \boldsymbol{X}_u \end{pmatrix}$ в соответствии с разбиением $\boldsymbol{y} = \begin{pmatrix} \boldsymbol{z} \\ \boldsymbol{u} \end{pmatrix}$, получим

$$S = (\boldsymbol{z} - \boldsymbol{X}_z\theta)'(\boldsymbol{z} - \boldsymbol{X}_z\theta) + (\boldsymbol{u} - \boldsymbol{X}_u\theta)'(\boldsymbol{u} - \boldsymbol{X}_u\theta). \qquad (37.81)$$

В правой части (37.81) оба члена неотрицательны, причем от $\boldsymbol{u}$ зависит только второй из них. Таким образом, минимум S в (37.81) достигается, если положить

$$\boldsymbol{u} = \boldsymbol{X}_u\theta. \qquad (37.82)$$

Тогда второй член в правой части (37.81) обращается в нуль, а минимальное значение S совпадает с первым членом, который нужно теперь минимизировать по θ.

Но результаты двух описанных выше двухшаговых методов минимизации должны совпасть. Таким образом, если мы получили $\hat{\theta}(u)$ первым методом, который дает

$$\hat{\theta}(u) = (X'X)^{-1} X'y = (X'X)^{-1}(X'_z z + X'_u u), \qquad (37.83)$$

то, используя это в сочетании с (37.82), получим

$$u = X_u \hat{\theta}(u). \qquad (37.84)$$

Из (37.84) следует, что каждое потерянное наблюдение следует приравнять к его оцениваемому среднему в исходном анализе методом НК.

37.53 (37.84) представляет собой систему линейных уравнений относительно u. Ее решение $\hat{u}$ нужно использовать в исходном анализе по методу НК. Непосредственное решение системы (37.84) получил Точер (1952).

Предположим сначала, что при исходном анализе методом НК мы заменили u на нулевой вектор $\mathbf{0}$. Тогда (37.83) принимает вид

$$\hat{\theta}(0) = (X'X)^{-1} X'_z z. \qquad (37.85)$$

Рассмотрим теперь

$$\hat{u} = \{I - X_u (X'X)^{-1} X'_u\}^{-1} X_u \hat{\theta}(0). \qquad (37.86)$$

Замечая из (37.83), что $\hat{\theta}(u) = \hat{\theta}(0) + (X'X)^{-1} X'_u u$, получим, что (37.86) сводится к

$$\{I - X_u (X'X)^{-1} X'_u\} \hat{u} = X_u \hat{\theta}(\hat{u}) - X_u (X'X)^{-1} X'_u \hat{u},$$

так что

$$\hat{u} = X_u \hat{\theta}(\hat{u}). \qquad (37.87)$$

А значит, $\hat{u}$, определяемое (37.86), является решением системы (37.84).

37.54 Таким образом, мы видим, что для того, чтобы оценить m потерянных наблюдений u так, чтобы можно было сохранить исходную форму расчета анализа по методу НК, нужно сделать следующее:

(а) провести исходный анализ с $u = \mathbf{0}$, чтобы получить $\hat{\theta}(\mathbf{0})$ в (37.85);

(б) вычислить $\hat{u}$ в (37.86);

(в) снова провести исходный анализ, подставляя $\hat{u}$ в y.

Следует отметить, что матрица в фигурных скобках, которую нужно обращать в (37.86), имеет порядок $(m \times m)$. Таким образом, если потеряно всего одно наблюдение, то матрица превращается в скаляр и приведенный выше пункт (б) тривиален.

Вторая из четырех статей Уилкинсона (1957—60) о потерянных наблюдениях дает подробное решение системы (37.84) для многих общих ситуаций ДА (см. также Биггерс (1959)).

37.55 Легко видеть, что оценки $\hat{\theta}(\hat{u})$, полученные с использованием (37.86) в исходном анализе по методу НК, совпадают с оценками, которые были бы получены с использованием только $n - m$ наблюденных значений (вообще говоря, при неортогональном анализе). В самом деле, из (37.83), используя (37.87), получим

$$X'X\hat{\theta}(\hat{u}) = X'_z z + X'_u \hat{u} = X'_z z + X'_u X_u \hat{\theta}(\hat{u}),$$

а значит,

$$X'_u z = (X'X - X'_u X_u)\, \hat{\theta}(\hat{u}) = X'_z X_z \hat{\theta}(\hat{u}). \qquad (37.88)$$

Но (37.88) в точности то же самое множество уравнений, которому удовлетворяет $\hat{\theta}$, если анализируется только z.

Из этого результата и из равенства нулю второго члена в правой части (37.81) сразу следует, что остаточная СК, полученная с использованием $\hat{u}$ в исходном анализе по методу НК совпадает с остаточной СК, когда анализируется только z. Однако число ст. св. остаточной СК, очевидно, должно уменьшиться, поскольку теперь у нас имеется только $n - m$ наблюдений. Если X и X_z имеют один и тот же ранг (т. е. если обе эти матрицы имеют полный ранг), то число ст. св. остаточной СК уменьшится на m, где m — число потерянных наблюдений. В более общей ситуации (см. упражнение 19.8) уменьшение будет равно разности между m и разностью рангов X и X_z.

37.56 Выражение для остаточной СК в таблице ДА остается неизменным. Однако все остальные СК окажутся неверными, если в исходном анализе по методу НК воспользоваться значением $\hat{u}$ из (37.86). Проще всего это понять из того факта (см. пример 35.4, 35.38 и пример 35.6), что все остальные СК в таблице ДА могут быть представлены в виде разности остаточных СК для двух линейных моделей, одна из которых является вариантом с дополнительными ограничениями другой. Очевидно, что правильные выражения для этих остаточных СК можно получить рассуждениями **37.50—5**, но используя значение $\hat{u}$ из (37.86) для соответствующей модели. При этом, разумеется, $\hat{u}$ будет, вообще говоря, отличаться от того, которое

получалось для рассматриваемой до сих пор полной модели. При этом, если взять $\hat{u}$ для полной модели, то значение каждой из остаточных СК будет завышено, поскольку это $\hat{u}$ не является нужным (минимизирующим). Таким образом, другие СК в таблице ДА нужно подправить на разность между поправочными уменьшениями соответствующих остаточных СК (или на одно поправочное уменьшение, если одна из этих остаточных СК является остаточной СК для полной модели). Соответственно нужно подправить и число степеней свободы на разность между двумя уменьшениями числа степеней свободы, значение которых рассматривалось в **37.55**. Однако эта разность часто оказывается равной нулю.

В третьей из своих четырех работ Уилкинсон (1957—60) дает явное описание метода получения уменьшающих поправок для других остаточных СК, а значит, и для других СК в таблице ДА. К счастью, как указал Иэйтс (1963), последние поправки, будучи разностью величин одного и того же знака, часто малы, так что приближенно можно использовать неисправленные значения СК.

37.57 Точер (1952) приводит аналогичные методы анализа для «испорченных» экспериментов другого типа, а именно, для экспериментов, в которых некоторые наблюдения перепутаны или некоторые наблюдения случайно взяты дважды. Плэкетт (1950) также обсуждает последнюю ситуацию.

Афифи и Елашов (1966) обобщают анализ на случай, когда не только среди y, но и среди элементов X имеются потерянные (см. упражнение 37.21).

УПРАЖНЕНИЯ

37.1 Имеется G групп наблюдений, и все наблюдения внутри данной группы имеют нормальное распределение с одинаковым средним и одинаковыми дисперсиями σ_g^2, так что выполняются предположения модели (37.1), за исключением условия гомоскедастичности между группами. Рассмотрим множества ограничений:

C_1: все σ_g^2 равны ($G-1$ ограничение),

C_2: r из k параметров θ равны нулю.

Используя переменные $z_\lambda = y_\lambda J^{-1/n}$ так, что (37.3) сводится к

$$L_\lambda(z \mid \hat{\theta}, \hat{\sigma}^2) = \{\hat{\sigma}_\lambda^2(z)\}^{-n/2},$$

показать, что (37.8) приводит к

$$L(z \mid \hat{\lambda}_{(2)}) = L(z \mid \hat{\lambda})\, l_1(z)\, l_2(z),$$

где l_1 — статистика критерия ОП, определенная в (24.40), а

$$l_2 = \left\{1 + \frac{r}{n-k} F\right\}^{-n/2},$$

где F — статистика критерия дисперсионного отношения, определенная в общем виде в (24.40), а для рассматриваемого конкретного случая в примере 24.8 (этот результат обобщает упражнение 24.6).

(Бокс и Кокс, 1964.)

37.2 Используя упражнение 23.7, показать, что если в (37.8) распределение l_p не зависит от некоторых параметров, для которых существует полная достаточная статистика t (вектор), а l_q является функцией только от t, то l_p и l_q стохастически независимы. Применить этот результат для получения свойства независимости в упражнениях 24.6 и 24.13. В упражнении 37.1 показать, что $l_1(z)$ и $l_2(z)$ независимы, если выполняется как C_1, так и C_2.

(См. Хогг, 1961.)

37.3 При подгонке ортогональными полиномами степени k в **28.16** уменьшение полной СК, связанное с членом степени r, согласно (28.72), т. 2, равно $Q_r = \hat{\alpha}_r^2 \sum_{j=1}^{n} \varphi_i^2(x_j)$. Показать, что отношения

$$z_r = Q_{k-r+1} \Big/ \sum_{s=k-r+2}^{k+1} Q_s, \qquad r = 1, \ldots, k,$$

где $Q_{k+1} = (n-k)s^2$ — остаточная СК, являются независимыми, если коэффициенты регрессии α_r все равны нулю:

(а) используя результат упражнения 37.2,

(б) используя результат упражнения 23.27.

(Из этого результата следует (см. Хогг (1961)), что можно использовать независимые критерии для проверки коэффициентов регрессии, начиная с высшей степени и опускаясь вниз, присоединяя СК, соответствующие коэффициентам, признанным нулевыми, к остаточной СК до тех пор, пока не найдется коэффициент, который признается не равным нулю. В этом случае процесс обрывается. Все критерии являются, конечно, $t^2(F)$-критериями, и составной критерий имеет размер $1-(1-\alpha)^k \sim k\alpha$, если на каждом шагу используются критерии размера α. Т. Андерсон (1962) показал при более слабых предположениях, что эта процедура максимизирует вероятность правильного определения ненулевого коэффициента.)

37.4 В **37.10** показать, что для биномиального распределения из **5.4**, соотношение (37.14) приводит к стабилизирующему дисперсию преобразованию $u\left(\frac{x}{n}\right) = \arcsin\left\{\left(\frac{x}{n}\right)^{1/2}\right\}$, где x/n — наблюденная доля «успехов».

(Энскомб (1948) показал, что если заменить x/n на $\left(x+\frac{3}{8}\right)\Big/\left(n+\frac{3}{4}\right)$, то получится лучшее стабилизирующее дисперсию преобразование. М. Фримэн и Тьюки (1950) предложили $\arcsin\left\{\left(\frac{x}{n+1}\right)^{1/2}\right\} + \arcsin\left\{\left(\frac{x+1}{n+1}\right)^{1/2}\right\}$ (см. пример 37.1). Эта функция табулирована Мостеллером и Ютцем (1961). См. также Лаубшер (1961).)

37.5 В **37.10** показать, что для отрицательного биномиального распределения из **5.15** соотношение (37.14) приводит к стабилизирующему дисперсию

преобразованию $u(t) = \operatorname{arsh}\left\{\left(\frac{x}{n}\right)^{1/2}\right\}$, где x/n — наблюденная доля «успехов».

(Энскомб (1948) показал, что если x/n заменить на $\left(x+\frac{3}{8}\right)\Big/\left(n-\frac{3}{4}\right)$, то получится лучшее стабилизирующее дисперсию преобразование. См. также Лаубшер (1961).)

37.6 В упражнении 37.4 показать, что другое преобразование $u\left(\frac{x}{n}\right) = \log\left\{\frac{x}{n}\Big/\left(1-\frac{x}{n}\right)\right\}$ стабилизирует дисперсию вблизи $p=\frac{1}{2}$. Показать, что это преобразование соответствует случаю, когда в (37.9) имеет место $D_n^2(\theta) = c\theta^2(1-\theta)^2$.

(см. Бартлетт, 1947а.)

37.7 Для перекрестной классификации с неравными частотами и клетках показать, что если клеточные выборочные средние рассматривать как отдельные наблюдения, то их усредненную дисперсию можно оценивать величиной s^2/H, где s^2 — остаточный СрК исходных наблюдений, а H — гармоническое среднее клеточных частот. Показать отсюда, как можно проводить приближенный ДА для клеточных средних.

(См. Шеффе, 1959.)

37.8 Применяя метод упражнения 37.7 к численным данным из примера 35.7, показать, что приближенный ДА для клеточных средних дает

	СК	Число ст. св.	СрК
Между полами	0,023	1	0,023
Между породами	0,020	7	0,003
Взаимодействия	0,006	7	0,001
Остаточная	0,049	15 517	0,0017(=0,0227×0,0759)

Сравнить значения F-отношений в этой таблице с точными значениями из упражнения 35.7.

37.9 Показать, что если линейная модель содержит члены $\theta_i x_i^{\mu_i}$, $\mu_i \neq 0$, то приближенно имеем

$$x_i^{\mu_i} = x_i + (\mu_i - 1)\, x_i \log x_i.$$

Показать отсюда, как можно оценить μ_i. Можно использовать итерации этого процесса.

(Бокс и Тидвелл, 1962.)

37.10 В 37.28 показать, что $\boldsymbol{M} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}$ инвариантна относительно любого неособого преобразования $\boldsymbol{W} = \boldsymbol{X}\boldsymbol{T}$. Рассматривая такое преобразование, что $\boldsymbol{W}'\boldsymbol{W}$ диагональна, показать, что $m = \sum_r M_{rr}^2$ удовлетворяет

$$m = \frac{(N-2)^{(2)}}{(N+1)^{(3)}}\left\{\sum_{i=1}^{p}\left(\frac{k_4}{k_2^2}\right)_i + \sum_{\substack{i,\,j=1\\ i\neq j}}^{p}\sum\left(\frac{k_{22}}{k_{20}k_{02}}\right)_{ij}\right\} + \frac{(N-1)\,p\,(p+2)}{N\,(N+1)}.$$

(Бокс и Ватсон, 1962.)

37.11 В 37.31, рассматривая дисперсию диагональных элементов M_{rr} матрицы $\boldsymbol{M}$, показать, что $m = \sum_r M_{rr}^2 \geqslant p^2/N$. Используя свойство инвариантности из упражнения 37.10, положить $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}$ и присоединить $N-p$ столбцов к матрице $\boldsymbol{X}$ (один из столбцов должен быть равен $N^{-1/2}1$), чтобы получить ортогональную матрицу. Показать отсюда, что $m \leqslant p(N-1)/N$. Исходя из вышесказанного, показать, что нижняя граница для m достижима, а верхняя нет.

(Бокс и Ватсон, 1962.)

37.12 В 37.32 показать, что если выбрать при однофакторной классификации $p+1$ групповые частоты n_j следующим образом:

$$\sum_{j=1}^{p+1}\frac{1}{n_j} = \frac{(N-1)}{N(N+1)}p^2 + \frac{4}{N+1}p + \frac{1}{N} \sim \frac{p^2+4p+1}{N+1},$$

то C_X в (37.52) обращается в нуль, и, значит, каково бы ни было распределение ошибок, никакая коррекция для $\mathrm{D}_p(W)$ при отклонении от нормальности не нужна. Если $p=1$, $n_1 = rN$, $n_2 = (1-r)N$, показать, что $C_X = 0$, если

$$r = \frac{1}{2}\left\{1 \pm \left(\frac{N-2}{3N}\right)^{1/2}\right\} \sim \frac{1}{2}(1 \pm 3^{-1/2}),$$

и что если $N = 12$, то оптимальными целыми значениями групповых частот являются 9 и 3.

(Бокс и Ватсон, 1962.)

37.13 Вывести (37.71—2), записав U, определенное после (37.70), в виде

$$U = \sum_{i\neq l}\sum R_{il},$$

где

$$R_{il} = \sum_{j=1}^{c} y_{ij}y_{lj},$$

и показав, что

$$\mathrm{M}_P(R_{il}) = 0, \quad \mathrm{M}_P(R_{il}^2) = (c-1)(k_2)_i(k_2)_l,$$

а значит,

$$\mathrm{M}_P(U) = 0, \quad \mathrm{M}_P(U^2) = (c-1)\sum_{i\neq l}\sum(k_2)_i(k_2)_l.$$

(Питмэн, 1938.)

37.14 В 37.41 показать, что если справедливо (37.75), то

$$\mathsf{D}_P(W) = \frac{2(r-1)}{r^3(c-1)}.$$

Методом 37.29 вывести отсюда, что число ст. св. приближенного F-критерия нужно подогнать к

$$\nu_1 = (c-1) - \frac{2}{r}, \qquad \nu_2 = (r-1)\nu_1.$$

Показать, что если в качестве условных чисел использовать в каждой строке ранги 1, 2, ..., c и обозначить затем R_j сумму рангов в j-м столбце, а $T = \sum_{j=1}^{c}\left\{R_j - \frac{r(c+1)}{2}\right\}^2$, то статистика W сведется к

$$W = \frac{12T}{r^2c(c^2-1)},$$

и что при $r \to \infty$

$$\nu_2 W/(1-W) \sim \frac{12T}{rc(c+1)}$$

имеет χ^2-распределение с $(c-1)$ ст .св.

(М. Фридмэн (1937); Кендалл и Бэбингтон Смит (1939); М. Фридмэн (1940) сравнивает F и χ^2 приближения.)

37.15 В примере 37.1 разложить

$$u_c(t) = (\theta + c)^{1/2}\left(1 + \frac{t-\theta}{\theta + c}\right)^{1/2}$$

в ряд и показать, что

$$\mathsf{M}(u_c) = (\theta + c)^{1/2} - \frac{1}{8}\theta^{-1/2} + \frac{24c-7}{128}\theta^{-3/2} + o(\theta^{-3/2}),$$

и

$$\mathsf{D}(u_c) = \frac{1}{4}\left\{1 + \frac{3-8c}{8\theta} + \frac{32c^2-52c+17}{32\theta^2} + o(\theta^{-2})\right\},$$

так что выбор $c = 3/8$ обращает в нуль коэффициент при θ^{-1} в выражении для дисперсии, которое принимает вид

$$\mathsf{D}(u_{3/8}) \sim \frac{1}{4}\left(1 + \frac{1}{16\theta^2}\right).$$

Показать отсюда, что

$$\{\mathsf{M}(u_c)\}^2 - c \sim \theta - \frac{1}{4} - \frac{3-8c}{32\theta},$$

так что, если для оценки θ использовать обратное к u_c преобразование, то понижающее смещение почти постоянно и равно 1/4.

(Результат с $c = 3/8$ восходит к А. Джонсону; см. Энскомб (1948).)

37.16 В упражнении 37.15 показать, что коэффициенты асимметрии и эксцесса величины $u_c(t)$ равны

$$\gamma_1 = -\frac{1}{2\theta^{1/2}}\left\{1 + \frac{25-48c}{16\theta}\right\} + o(\theta^{-3/2}),$$

$$\gamma_2 = \frac{1}{\theta}\left\{1 + \frac{945-1536c}{256\theta}\right\} + o(\theta^{-2}).$$

Сравнить с

$$\gamma_1 = \theta^{-1/2}, \qquad \gamma_2 = \theta^{-1}$$

для исходной пуассоновской величины t. Таким образом, при любом выборе c значение γ_1 уменьшается примерно вдвое (с изменением знака), а γ_2 с точностью до первого порядка не изменяется.

(Энскомб, 1948.)

37.17 Используя результат для $\mathsf{D}(u_c)$ из упражнения 37.15, показать, что преобразование

$$u'_\delta = \left(t + \frac{1}{2} + \delta\right)^{1/2} + \left(t + \frac{1}{2} - \delta\right)^{1/2}$$

имеет дисперсию

$$\mathsf{D}(u'_\delta) = 1 - \frac{1}{8\theta} + \frac{16\delta^2 - 1}{32\theta^2} + o(\theta^{-2}),$$

так что, если мы выберем $\delta = 1/2$, чтобы получить u' из примера 37.1, то

$$\mathsf{D}(u'_{1/2}) = 1 - \frac{1}{8\theta} + \frac{3}{32\theta^2} + o(\theta^{-2}).$$

37.18 Для двухфакторной перекрестной классификации с одним наблюдением в клетке (или с большим числом, если нет взаимодействий) показать, что медианный критерий для эффектов столбцов получается подсчетом числа m_j наблюдений в j-м столбце, превосходящих медиану строки, и построением $(2 \times c)$-таблицы, аналогичной (37.76), с использованием (37.77) как статистики критерия для больших выборок.

(Браун и Муд, 1951.)

37.19 Для общей сбалансированной двухфакторной перекрестной классификации показать, что эффекты столбцов можно проверять против взаимодействий, находя медиану в каждой клетке и применяя критерий упражнения 37.18 к этим медианам. Показать, что критерий остается свободным от распределения, если клеточные частоты отличаются для разных строк, но не внутри строки.

(Браун и Муд, 1951.)

37.20 В упражнении 37.19 показать, что эффекты столбцов и взаимодействий можно проверять совместно, подсчитывая число m_{ij} из n_{ij} наблюдений в (i, j)-й клетке, которые превосходят медиану i-й строки, и проверяя затем гипотезу о том, что $\mathsf{M}(m_{ij}) = n_{ij}/2$ для всех i, j. Показать, что это эквивалентно гипотезе о том (см. 33.60), что в каждой строчке трехфакторной $(r \times c \times 2)$-таблицы столбцы и слои независимы, что ведет для больших выборок к χ^2-критерию с $r(c-1)$ ст. св.

(Браун и Муд, 1951.)

37.21 Обобщить **37.51—4**, положив $y = \begin{Bmatrix} z \\ u \\ v \\ w \end{Bmatrix}$, где z содержит наблюденные значения y, не связанные с потерянными значениями x, u — потерянные значения y, не связанные с потерянными значениями x, v — наблюденные значения y, связанные с некоторыми из потерянных x, w — потерянные значения y, связанные с некоторыми из потерянных x. Пусть $\begin{Bmatrix} X_z \\ X_u \\ X_v \\ X_w \end{Bmatrix}$ — соответствующее разбиение X, а X_0 совпадает с X, если $X_v = X_w = 0$.

(а) Положить $u = w = 0$ и заменить X на X_0. Оценить θ, обозначив полученную оценку $\hat{\theta}(0)$.

(б) Оценить u в соответствии с (37.86), заменив X на X_0.

(в) Получить новую оценку для θ, положив $u = \hat{u}$, $w = 0$.

Показать, что в результате получается НК-оценка.

(Афифи и Елашов, 1966.)

ГЛАВА 38

ПЛАНИРОВАНИЕ ЭКСПЕРИМЕНТОВ

38.1 В большей части этой книги мы сталкивались с задачами, возникающими при анализе наблюдений, точнее с задачами оценки и проверки гипотез, возникающими в различных теоретических исследованиях. Понятно, что каждое исследование какого-то метода анализа является уроком на будущее. Так, например, если мы узнаем, что один метод оценки более эффективен, чем другой, то естественно, что в будущем мы будем использовать именно этот лучший метод. Однако отсюда не следует, что тем самым улучшится *метод проведения* наблюдений, ибо улучшится только метод *анализа* наблюденных значений. В этой и двух следующих главах будут рассмотрены вопросы планирования, под которыми мы понимаем влияние метода проведения или отбора наблюдений на анализ.

38.2 По существу, рассмотрение планирования для нас не ново. В примере 28.4 т. 2 при анализе задачи простой линейной регрессии мы получили, что, выбрав должным образом начало отсчета и экспериментальные значения регрессора, можно гарантировать ортогональность анализа и минимальность соответствующих дисперсий оценок. Это вопрос планирования, поскольку он связан с тем, как проводить наблюдения. В том же примере указывалось на опасность того, что это оптимальное решение может привести к невозможности проверки предположения о том, что регрессионная модель линейна по регрессорам. Если нет уверенности, что это не так, возможно, следует несколько «застраховаться», уклонившись от оптимального выбора регрессоров.

Далее из результатов изучения устойчивости в **37.21** следует, что в экспериментах следует брать равные частоты во всех клетках. Это снова вопрос планирования, поскольку ответ влияет на метод проведения наблюдений.

38.3 В этой главе будут рассмотрены вопросы о влиянии планирования на экспериментирование. При рассмотрении будут в основном использоваться линейные модели и техника ДА из глав 35—37. В главах 39—40 мы обратимся к задачам планирования в выборочных обследованиях. Различие между этими областями весьма четко. Можно сказать, что при обследованиях

37.21 Обобщить **37.51—4**, положив $y = \begin{Bmatrix} z \\ u \\ v \\ w \end{Bmatrix}$, где z содержит наблюденные значения y, не связанные с потерянными значениями x, u — потерянные значения y, не связанные с потерянными значениями x, v — наблюденные значения y, связанные с некоторыми из потерянных x, w — потерянные значения y, связанные с некоторыми из потерянных x. Пусть $\begin{Bmatrix} X_z \\ X_u \\ X_v \\ X_w \end{Bmatrix}$ — соответствующее разбиение X, а X_0 совпадает с X, если $X_v = X_w = 0$.

(а) Положить $u = w = 0$ и заменить X на X_0. Оценить θ, обозначив полученную оценку $\hat{\theta}(0)$.

(б) Оценить u в соответствии с (37.86), заменив X на X_0.

(в) Получить новую оценку для θ, положив $u = \hat{u}$, $w = 0$.

Показать, что в результате получается НК-оценка.

(Афифи и Елашов, 1966.)

ГЛАВА 38

ПЛАНИРОВАНИЕ ЭКСПЕРИМЕНТОВ

38.1 В большей части этой книги мы сталкивались с задачами, возникающими при анализе наблюдений, точнее с задачами оценки и проверки гипотез, возникающими в различных теоретических исследованиях. Понятно, что каждое исследование какого-то метода анализа является уроком на будущее. Так, например, если мы узнаем, что один метод оценки более эффективен, чем другой, то естественно, что в будущем мы будем использовать именно этот лучший метод. Однако отсюда не следует, что тем самым улучшится *метод проведения* наблюдений, ибо улучшится только метод *анализа* наблюденных значений. В этой и двух следующих главах будут рассмотрены вопросы планирования, под которыми мы понимаем влияние метода проведения или отбора наблюдений на анализ.

38.2 По существу, рассмотрение планирования для нас не ново. В примере 28.4 т. 2 при анализе задачи простой линейной регрессии мы получили, что, выбрав должным образом начало отсчета и экспериментальные значения регрессора, можно гарантировать ортогональность анализа и минимальность соответствующих дисперсий оценок. Это вопрос планирования, поскольку он связан с тем, как проводить наблюдения. В том же примере указывалось на опасность того, что это оптимальное решение может привести к невозможности проверки предположения о том, что регрессионная модель линейна по регрессорам. Если нет уверенности, что это не так, возможно, следует несколько «застраховаться», уклонившись от оптимального выбора регрессоров.

Далее из результатов изучения устойчивости в **37.21** следует, что в экспериментах следует брать равные частоты во всех клетках. Это снова вопрос планирования, поскольку ответ влияет на метод проведения наблюдений.

38.3 В этой главе будут рассмотрены вопросы о влиянии планирования на экспериментирование. При рассмотрении будут в основном использоваться линейные модели и техника ДА из глав 35—37. В главах 39—40 мы обратимся к задачам планирования в выборочных обследованиях. Различие между этими областями весьма четко. Можно сказать, что при обследованиях

проводятся наблюдения над выборкой из конечной популяции индивидуумов, в то время как в экспериментах проводятся наблюдения, которые в принципе порождены гипотетической бесконечной генеральной совокупностью точно так, как это было при бросании монеты (см. **1.29** и **9.4**, а также приведенный ниже пример 38.1). Конечно, можно иногда проводить эксперимент над членами выборки, полученной в результате обследования или даже считать выборочное обследование результатом (охватывающего) эксперимента. Тем не менее должно быть понятно существо разницы между этими двумя областями.

Кокрэн (1965) дает интересное общее обсуждение проблем, связанных с сущностью статистического вывода, возникающих при обследовании, а не при контролируемых экспериментах.

Принципы экспериментирования

38.4 При классическом изложении принципов экспериментирования подчеркивается важность варьирования в процессе эксперимента (предположительно) причинных факторов, чтобы изучить их влияние на рассматриваемую зависимую переменную. Однако несовершенство подобного подхода обнаруживается по меньшей мере в двух отношениях. Проанализируем эти недостатки, отправляясь от общей идеологии статистического исследования зависимостей, выраженной, например, Миллом (1843) в его пятом правиле экспериментальных исследований: «Если какое-то одно явление изменяется неким образом всякий раз, когда меняется другое явление, то первое является либо причиной, либо следствием второго, либо связано с ним некой опосредованной причинной зависимостью».

Первая неадекватность, с которой мы сталкиваемся, следуя классическим принципам экспериментирования, связана с тем, что мы изменяем только один из причинных факторов в каждом «единичном эксперименте».

В свете результатов главы 35 мы видим теперь, что без одновременного варьирования значений нескольких факторов нет надежды количественно определить взаимодействия между причинными факторами. Такой подход не только не дает нам возможности узнать связи между причинными факторами, но и может привести к неправильным выводам. Так предположим, что цель эксперимента состоит в определении комбинации ингредиентов A и B при производстве керамики, которая дает наименьшую хрупкость. Если мы определим дозу ингредиента A, которая приводит к наименьшей хрупкости, а потом соответствующую дозу ингредиента B, нет никакой гарантии, что, объединяя эти значения, мы получим желаемую оптимальную комбинацию. Взаимодействия между факторами, которые приводят к таким эффектам, можно изучать только при одновременном изменении факторов.

38.5 Вторая неадекватность, возникающая при классическом подходе, даже более существенна и снова иллюстрируется приведенной в **38.4** цитатой из Милла. Она состоит в опасности приписать одному или нескольким экспериментальным факторам влияние на зависимую переменную, которое на самом деле объясняется изменением некоторого причинного фактора, не учтенного в структуре эксперимента. Во время проведения эксперимента некоторый неизвестный экспериментатору причинный фактор может меняться таким образом, что конкретная комбинация экспериментальных факторов оказывается предпочтительной. Тогда будет создаваться впечатление, что эта комбинация высоко эффективна, хотя на самом деле к хорошему результату приводит неизвестный нам фактор.

При классическом рассмотрении этот вопрос не имеет решения. Он дает представление о том насколько сложной является задача. Мы *никогда* не можем быть вполне уверены в том, что все важные, или даже самые важные, причинные факторы учтены структурой эксперимента. Некоторые факторы могут быть просто неизвестны. Другие, хотя и известны, но могут необоснованно рассматриваться как не имеющие большого значения, и ими умышленно пренебрегают. При проведении эксперимента всегда нужно иметь гарантии от искажения выводов, обусловленного побочными внешними эффектами.

Рандомизация

38.6 Современное решение вопросов из **38.5** было впервые предложено Р. А. Фишером в двадцатых годах нашего столетия (см. его книгу — Фишер (1935)). На протяжении всех трех томов мы видели, как велик и разносторонен вклад Фишера в статистическую теорию. Тем не менее, возможно, не будет преувеличением сказать, что его пропаганда *рандомизации* при планировании экспериментов была самым важным и имеющим самое большое влияние результатом из его многочисленных достижений в статистике.

38.7 Принцип рандомизации формулируется очень просто: *Если при проведении эксперимента проводится размещение экспериментальных объектов по комбинациям факторов, то это должно делаться случайным образом с использованием равных вероятностей.* Таким образом, каждая комбинация факторов будет иметь одинаковый шанс быть примененной к каждому экспериментальному объекту, к которому она может быть применена.

Очевидно, что если проводится рандомизация в соответствии с этим принципом, то при конкретном размещении экспериментальных объектов, к которому мы приходим, все еще может оказаться в предпочтении конкретная комбинация факторов. Однако описанные в **38.5** трудности нас больше не беспокоят, поскольку мы включили процесс рандомизации в те рамки, внутри которых делаются наши выводы. Гипотетическая совокупность, про которую мы делаем выводы, включает все возможные случаи размещения экспериментальных объектов, к которым может привести рандомизация. Внутри этой совокупности при естественном процессе рандомизации не может возникнуть предпочтение для учтенных экспериментом факторов из-за влияния факторов, не учтенных экспериментом. Таким образом, наши выводы будут свободны от смещения.

Даже если взаимосвязь между зависимой переменной и некоторым неизвестным причинным фактором не будет узнана вплоть до окончания эксперимента, справедливость выводов не нарушится, если при проведении эксперимента влияние этого фактора было «рандомизировано».

38.8 Таким образом, вопросы из **38.5** разрешаются изменением основы, на которой делаются выводы. Это неизбежно приводит к изменению теоретических основ наших выводов, и мы далее коротко на этом остановимся. Но сначала проиллюстрируем вышесказанное на простом примере.

Пример 38.1

Проводится эксперимент для исследования зависимости времени реакции от содержания алкоголя в крови у мужчин водителей автомобилей. Принимающие участие в эксперименте водители должны принять заданную дозу алкоголя и после фиксированного промежутка времени пройти испытание на содержание алкоголя в крови и некоторое стандартное испытание на время реакции. Задача состоит в том, как «разместить» водителей между различными дозами алкоголя.

По существу, мы имеем дело с задачей регрессии, где время реакции — зависимая переменная (y), а содержание алкоголя в крови — регрессор (x). Следует заметить, что x не является строго контролируемой переменной. Строго проконтролировать можно только дозу алкоголя (z), а после этого мы просто наблюдаем значение x в каждом случае. Однако z и x связаны очень тесно, и разумно предположить, что каждому фиксированному значению z соответствует некоторое множество значений x. При этом, если выбрать значения z достаточно хорошо, то соответствующие множества значений x не будут пересекаться. Тогда задачу можно рассматривать как однофакторную классификацию, причем индексы классификации соответствуют значениям z.

На этом примере хорошо видны трудности, которые возникают при отсутствии рандомизированного размещения доз алкоголя между водителями. Предположим, например, что каждому водителю позволили самому выбрать для себя дозу алкоголя. По-видимому, те водители, которые обычно употребляют много алкоголя, выберут дозы бо́льшие, чем другие водители. Но нормальная привычка к выпивке влияет на устойчивость человека к воздействию алкоголя. Поэтому в соответствующем испытании на время реакции может оказаться замаскированной истинная разница между воздействием различных доз алкоголя.

Можно возразить, что, позволив водителям самим выбрать дозу алкоголя, мы получим более правильную картину того, что произойдет в практике вождения. Даже если бы это было верно, следует понимать, что соответствующий вывод не имеет никакого отношения к проводимому научному исследованию. Существо разницы между обследованием и экспериментом состоит в том, что при обследовании делается попытка описать существующую совокупность, в то время как эксперимент имеет дело с исследованием соотношений, которые не обязательно соответствуют какой-то конкретной совокупности. Об этом различии говорилось выше в **38.3**, и мы еще вернемся к обсуждению этого вопроса в **38.12**. В отношении же приведенного выше примера нужно понимать, что рандомизированное размещение доз алкоголя между водителями необходимо, чтобы предохранить от смещения выводы, которые будут сделаны из эксперимента.

38.9 Тот факт, что «внешнее» влияние, такое, как обычная привычка к выпивке, устраняется рандомизацией, никоим образом не мешает нам проводить анализ этого влияния после того, как закончен (рандомизированный) эксперимент. В случае примера 38.1, по-видимому, будет трудно точно установить обычную привычку к выпивке всех участников. Но рассмотрим влияние времени дня, в которое проводится испытание. Вне зависимости от того, «рандомизируется» оно или нет при проведении эксперимента, ничто не мешает нам провести в последующем регрессионный анализ влияния времени дня на время реакции. Если окажется, например, что время реакции становится выше при проведении испытаний в более позднее время, то это послужит поводом для дальнейших исследований и, возможно, приведет в дальнейшем к модификации процедуры эксперимента.

38.10 Наше утверждение о достоинстве рандомизации не следует понимать так, что все рандомизированные эксперименты не оставляют желать ничего лучшего. Рассмотрим, как

и в **38.9**, влияние времени дня на время реакции, и предположим, что все испытания проводились в 6 часов после полудня, т. е. в конце рабочего дня водителей, принимающих участие в эксперименте. Тогда фактор «время дня» постоянен на одном уровне и эксперимент уязвим к возможности того, что этот фактор взаимодействует с влиянием содержания алкоголя в крови на время реакции. Рандомизация по отношению к дозам алкоголя не помогает в этом случае, и здесь снова применима критика классической процедуры, приведенная в **38.5**. На самом деле рандомизация была неполной, поскольку мы пренебрегли временем дня как возможным причинным фактором. Рандомизация приносит пользу в выводах только по отношению к тем аспектам, к которым она применялась.

38.11 Понятно, что факторы, влияющие на зависимую переменную в любом эксперименте, делятся экспериментатором, явно или неявно, на следующие три класса:

(1) факторы, учтенные структурой эксперимента (доза алкоголя в примере 38.1);

(2) факторы, «рандомизированные» при эксперименте (обычная привычка к выпивке в примере 38.1);

(3) факторы, не учтенные структурой эксперимента и не рандомизированные (время дня в **38.10**).

Следует заметить, что классы (1) и (2) требуют проведения некоторых действий, которые влияют на действительную программу проведения эксперимента или на применяемую процедуру рандомизации. В отличие от этого, последний из трех классов является остаточным. Естественно, что экспериментатор может умышленно принять решение о том, что некоторым фактором можно пренебречь и считать, что для устранения его влияния не нужна даже рандомизация. Так, в примере 38.1 мы почти наверняка пренебрежем цветом глаз водителей. Однако фактор может попасть в класс (3) и просто потому, что на него не обратили внимания, как это было со временем дня в **38.10**.

Существенным элементом искусства экспериментатора является выбор факторов, которые рандомизируются при проведении эксперимента. Если экспериментатор аккуратен, то он будет рандомизировать все факторы, которые предполагаются причинно важными, но которые не включены в структуры эксперимента. Но все экспериментаторы вынуждено пренебрегают некоторыми предположительно причинными факторами. Если бы это было не так, то процедура рандомизации приводила бы к очень большим усложнениям. Таким образом, выбор факторов, которые будут рандомизироваться, является, по существу, вопросом умения правильно разобраться в явлении.

38.12 В **38.7—8** мы видели, что генеральная совокупность, про которую можно делать обоснованные выводы из рандомизированного эксперимента, является гипотетической генеральной совокупностью, зависящей от самого процесса рандомизации. Однако экспериментатор должен делать свои выводы в отношении реального мира. В примере 38.1 гипотетическая генеральная совокупность, о которой делаются выводы в результате эксперимента, включает в себя все возможные размещения доз алкоголя между водителями, принимающими участие в эксперименте. Однако, насколько широко можно обобщить эти выводы, чтобы распространить их на такие более обширные генеральные совокупности, как

(а) всех водителей автомобилей мужчин,

(б) всех водителей автомобилей (включая женщин)?

Если водители, принимающие участие в эксперименте, были отобраны из всех водителей мужчин в результате случайного выбора (не обязательно *простого* случайного выбора), то мало кто из экспериментаторов будет колебаться, чтобы обобщить полученные в результате эксперимента данные на генеральную совокупность (а). Аналогично мало кто из экспериментаторов поспешит распространить эти данные на генеральную совокупность (б) без дальнейших знаний о водителях-женщинах, полученных, возможно, из других экспериментов.

Предположим, однако, что принимающие участие в эксперименте водители отбирались из служащих одной и той же корпорации. Даже если они отбирались при этом случайным образом, мы сумеем распространить наши выводы самое большее на генеральную совокупность, состоящую из всех водителей, работающих в этой корпорации. Обычно корпорации принимают участие в эксперименте или из-за собственной заинтересованности, или из-за желания сотрудничать с некоторым научным учреждением. Если отсутствует иной процесс подбора корпорации, то дальнейшие обобщения результатов эксперимента нуждаются в оправдании. Только в том случае, когда экспериментальный материал является (или предполагается) эквивалентным случайному выбору из более обширной генеральной совокупности, можно перенести результаты эксперимента на эту более обширную генеральную совокупность.

В конце концов в этой связи нельзя избежать обсуждения возможности обобщений, поскольку речь идет об обобщении по пространству (например, на другие страны) и по времени.

38.13 Мы очень кратко познакомились с некоторыми фундаментальными вопросами, касающимися выводов из экспериментов. Для более широкого ознакомления (в значительной степени не носящего технический характер) читателю рекомендуется книга Кокса (1958а), так же как упоминавшаяся книга Фишера (1935).

Мешающие факторы: блочные эксперименты

38.14 Мы видели, что существенная черта современного экспериментирования состоит в «рандомизации» влияния тех факторов, которые не учтены структурой эксперимента. В практических ситуациях одним из таких факторов является то, что сами экспериментальные объекты не могут быть физически однородны. Так, сельскохозяйственные эксперименты проводятся на участках земли, которые, хотя и расположены близко друг от друга, но не имеют одни и те же характеристики плодородия. Более того, если число участков, необходимых для проведения эксперимента, растет, то, по-видимому, будет расти и вариация их плодородия. При этом становится разумным использовать несколько маленьких групп аналогичных участков, а не одну большую группу довольно неоднородных участков. Аналогично при генетических исследованиях оказывается разумным проводить эксперименты над животными для членов одного и того же приплода, причем приплод обычно бывает довольно малым по количеству членов, так что нужно использовать несколько приплодов. Здесь отдельные животные, так же как участки в сельскохозяйственных экспериментах, являются экспериментальными объектами, в то время как приплоды, так же как группы аналогичных участков, называются «блоками» экспериментальных объектов. Используя эту терминологию, можно сказать, что многие эксперименты являются блочными или, даже не боясь тавтологии, сказать (считая, что может быть всего один блок), что *все* эксперименты являются блочными экспериментами. При этом эффект различия между блоками исследуется только с той целью, чтобы устранить его влияние. Фактически блоки являются «мешающими» факторами в эксперименте.

Перейдем к формальному исследованию блочных экспериментов основанному на работе Точера (1952).

38.15 Предположим, что эксперимент проводится над экспериментальными объектами, объединенными в b блоков, каждый из которых содержит k $(k > 1)$ объектов, так что всего можно получить $bk = n$ наблюдений. Эксперимент проводится для исследования перекрестной (или иерархической, или смешанной) классификации некоторых факторов, представляющих интерес. Предположим, что классификация содержит t различных клеток. Так, например, в двухфакторной классификации с r строчками и c столбцами имеется $t = rc$ различных клеток. Эти t клеток будем называть «обработками». Задача состоит в том, чтобы выбрать, какие обработки провести над каждым из k объектов в каждом блоке.

Будем предполагать, что в каждом блоке никакая обработка не применяется больше чем к одному объекту*). Это предположение разумно, поскольку, если нужны еще наблюдения, то мало смысла дублировать обработки внутри одного блока, а не использовать их в других блоках. Из упомянутого предположения следует, что $t \geqslant k$.

38.16 Эксперимент можно полностью описать, определив *матрицу обработок* (treatment matrix) порядка $(k \times t)$ для каждого блока. Для матрицы $\boldsymbol{t}_j$, относящейся к j-му блоку, элемент, стоящий на пересечении l-й строки и i-го столбца, равен единице или нулю в зависимости от того, применяется ли i-я обработка к l-му объекту этого блока или нет. Очевидно, что в каждой строке присутствует один ненулевой элемент (поскольку к каждому объекту применяется одна обработка) и что в каждом столбце имеется не более одного ненулевого элемента (из-за предположения в **38.15**).

Если нужно описать только применение обработок к блокам, без ссылки на применение к конкретным объектам внутри блока, то, используя b матриц обработок $\boldsymbol{t}_j$, можно построить *матрицу инциденций* (incidente matrix) эксперимента $\boldsymbol{n}$, порядка $(t \times b)$, у которой (i, j)-й элемент n_{ij} равен единице или нулю в зависимости от того, применяется ли i-я обработка к j-му блоку.

38.17 Если j-й столбец матрицы $\boldsymbol{n}$ обозначить $\boldsymbol{n}_j$, а $\mathbf{1}_p$ — вектор $p \times 1$, состоящий из единиц, то

$$\boldsymbol{t}_j'\mathbf{1}_k = \boldsymbol{n}_j, \tag{38.1}$$

поскольку в левой части (38.1) проводится просто суммирование элементов столбцов матрицы $\boldsymbol{t}_j$. Далее, поскольку все элементы матрицы $\boldsymbol{t}_j$ равны нулю или единице, то

$$\boldsymbol{t}_j'\boldsymbol{t}_j = \operatorname{diag}(\boldsymbol{n}_j), \tag{38.2}$$

где $\operatorname{diag}(\boldsymbol{z})$ обозначает диагональную матрицу, диагональные элементы которой совпадают с элементами вектора $\boldsymbol{z}$.

Если i-я обработка применяется во всем эксперименте r_i раз $(r_i > 0)$, а $\boldsymbol{r}$ обозначает $(t \times 1)$-вектор, элементы которого равны r_i, то суммирование элементов строк матрицы $\boldsymbol{n}$ дает

$$\boldsymbol{n}\mathbf{1}_b = \boldsymbol{r}, \tag{38.3}$$

а суммирование элементов столбцов дает

$$\boldsymbol{n}'\mathbf{1}_t = k\mathbf{1}_b, \tag{38.4}$$

*) Это предположение, которое выполняется для всех хорошо известных планов экспериментов, может быть ослаблено (см. Точер (1952)).

поскольку каждый блок состоит из k объектов. Далее из (38.2—3) получаем

$$\sum_j t_j' t_j = \sum_j \operatorname{diag}(n_j) = \operatorname{diag}\left(\sum_j n_j\right) = \operatorname{diag}(n\mathbf{1}_b) = \operatorname{diag}(r). \quad (38.5)$$

38.18 В духе наших предыдущих обсуждений размещение обработок между объектами будет рандомизироваться независимо внутри каждого блока, но для начала мы не будем рассматривать влияния внутриблочной рандомизации на выводы, которые делаются в результате эксперимента. Мы вернемся к этому вопросу в **38.41**. Сначала будем смотреть на рандомизацию как на общую предупредительную меру против смещения и будем проводить наш анализ в терминах линейной модели (модель I), хорошо известной нам из предыдущих глав.

Линейная модель для блочных экспериментов

38.19 Очевидно, что линейная модель для блочных экспериментов записывается в виде

$$\mathsf{M}(y_{ij}) = \tau_i + \beta_j, \quad (38.6)$$

где τ_i — эффекты обработок, а β_j — эффекты блоков. При этом предполагается, что эффекты блоков и обработок аддитивны и не взаимодействуют.

В (19.19) мы видели, что единственными линейными функциями параметров, допускающими несмещенные оценки с помощью линейных функций наблюдений, являются линейные комбинации математических ожиданий наблюдений. А значит, в (38.6) можно оценить без смещения только линейные комбинации величин $(\tau_i + \beta_j)$. Это понятно и из того факта, что добавление любой константы ко всем τ_i и вычитание этой константы из всех β_j не меняет (38.6). Чтобы освободиться от невозможности раздельной оценки отдельных параметров τ_i, мы, как и в сингулярной модели в **19.13—15** т. 2 и в примере 19.9, введем линейное ограничение на параметры. Поскольку эффекты блоков β_j являются мешающими параметрами, естественно наложить ограничения только на них в виде

$$\sum_{j=1}^{b} \beta_j = 0.$$

Фактически мы добавляем к n величинам y_{ij} фиктивную случайную величину (нуль), «математическое ожидание» которой равно сумме β_j. Это позволяет нам отделить мешающие нам эффекты блоков от интересующих нас эффектов обработок.

38.20 В силу наложенного ограничения для несингулярного представления эффектов блоков нужны $b-1$ параметров.

Обозначим их $\alpha_1, \ldots, \alpha_{b-1}$ и представим в виде вектора $\boldsymbol{\alpha}$. Вектор $\boldsymbol{\alpha}$ можно получить из вектора $\boldsymbol{\beta}$, координаты которого равны β_j, следующим образом. Возьмем произвольную ортогональную $(b \times b)$-матрицу U_0, первая строчка которой равна $b^{-1/2}\mathbf{1}_b'$, а оставшиеся $b-1$ строчки (соответствующую подматрицу мы обозначим u) в остальном произвольны. Тогда

$$U_0\boldsymbol{\beta} = \begin{pmatrix} 0 \\ u\boldsymbol{\beta} \end{pmatrix}, \quad (38.7)$$

поскольку ограничения на β_j записываются в виде

$$\mathbf{1}_b'\boldsymbol{\beta} = 0. \quad (38.8)$$

Поскольку U_0 ортогональна, то $U_0^{-1} = U_0'$, а значит, из (38.7) имеем

$$\boldsymbol{\beta} = U_0'\begin{pmatrix} 0 \\ u\boldsymbol{\beta} \end{pmatrix} = u'u\boldsymbol{\beta}. \quad (38.9)$$

Положим теперь

$$\boldsymbol{\alpha} = u\boldsymbol{\beta}. \quad (38.10)$$

Тогда (38.7) и (38.9) принимают вид

$$\begin{aligned} U_0\boldsymbol{\beta} &= \begin{pmatrix} 0 \\ \boldsymbol{\alpha} \end{pmatrix}, \\ \boldsymbol{\beta} &= u'\boldsymbol{\alpha}. \end{aligned} \quad (38.11)$$

38.21 Перейдем к формальному решению модели блочных экспериментов по методу НК. Пусть y_j обозначает вектор из k наблюдений в j-м блоке, y — вектор, объединяющий все y_j, u представлено в виде векторов-столбцов $u_1, \ldots, u_b$, каждый из которых имеет порядок $b-1$. Тогда (38.6) равносильно соотношению

$$y = X\theta + \varepsilon, \quad (38.12)$$

где соответствующее разбиение X и θ имеет вид

$$\underset{bk \times (t+b-1)}{X} = \left(\begin{array}{c|c} t_1 & \mathbf{1}_k u_1' \\ t_2 & \mathbf{1}_k u_2' \\ \vdots & \vdots \\ t_b & \mathbf{1}_k u_b' \end{array}\right) \quad (38.13)$$

и

$$\underset{(t+b-1)\times 1}{\theta} = \begin{pmatrix} \boldsymbol{\tau} \\ \hline \boldsymbol{\alpha} \end{pmatrix}. \quad (38.14)$$

Как всегда, компоненты вектора ошибок ε в (38.12) предполагаются некоррелированными с нулевыми математическими ожиданиями и дисперсиями, равными σ^2.

38.22 Используя (38.5), (38.1) и соотношения $\mathbf{1}_k'\mathbf{1}_k = k$, $\boldsymbol{uu}' = \boldsymbol{I}_{b-1}$ получим из (38.13)

$$\boldsymbol{X}'\boldsymbol{X} = \left[\begin{array}{c|c} \sum_j \boldsymbol{t}_j'\boldsymbol{t}_j & \sum_j \boldsymbol{t}_j'\mathbf{1}_k\boldsymbol{u}_j' \\ \hline \sum_j \boldsymbol{u}_j\mathbf{1}_k'\boldsymbol{t}_j & \sum_j \boldsymbol{u}_j\mathbf{1}_k'\mathbf{1}_k\boldsymbol{u}_j' \end{array}\right] = \left(\begin{array}{c|c} \operatorname{diag}(\boldsymbol{r}) & \boldsymbol{nu}' \\ \hline \boldsymbol{un}' & k\boldsymbol{I}_{b-1} \end{array}\right). \tag{38.15}$$

Теперь мы можем обратить матрицу в (38.15). Обозначая

$$\boldsymbol{\Omega} = \{\operatorname{diag}(\boldsymbol{r}) - \boldsymbol{nu}'\boldsymbol{un}'/k\}^{-1}, \tag{38.16}$$

получим

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \left(\begin{array}{c|c} \boldsymbol{\Omega} & -\boldsymbol{\Omega nu}'/k \\ \hline -\boldsymbol{un}'\boldsymbol{\Omega}/k & \{\boldsymbol{I}_{b-1} + \boldsymbol{un}'\boldsymbol{\Omega nu}'/k\}/k \end{array}\right), \tag{38.17}$$

что проверяется простым перемножением правых частей в (38.15) и (38.17). Таким образом,

$$\boldsymbol{X}'\boldsymbol{y} = \left(\begin{array}{c} \sum_j \boldsymbol{t}_j'\boldsymbol{y}_j \\ \hline \sum_j \boldsymbol{u}_j\mathbf{1}_k'\boldsymbol{y}_j \end{array}\right) = \left(\begin{array}{c} \boldsymbol{T} \\ \boldsymbol{uB} \end{array}\right), \tag{38.18}$$

где $\boldsymbol{T} = \begin{bmatrix} T_1 \\ \cdot \\ \cdot \\ \cdot \\ T_t \end{bmatrix}$, а T_i — сумма y для всех объектов, к которым применялась i-я обработка, в то время как $\boldsymbol{B} = \begin{bmatrix} B_1 \\ \cdot \\ \cdot \\ \cdot \\ B_b \end{bmatrix}$, где $B_j = \mathbf{1}_k'\boldsymbol{y}_j$ — сумма y для всех объектов из j-го блока. Согласно (38.17—18) НК-оценки имеют вид

$$\hat{\boldsymbol{\theta}} = \left(\begin{array}{c} \hat{\boldsymbol{\tau}} \\ \hline \hat{\boldsymbol{\alpha}} \end{array}\right) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y},$$

где

$$\hat{\boldsymbol{\tau}} = \boldsymbol{\Omega}(\boldsymbol{T} - \boldsymbol{nu}'\boldsymbol{uB}/k) \tag{38.19}$$

и

$$\hat{\boldsymbol{\alpha}} = \boldsymbol{u}(\boldsymbol{B} - \boldsymbol{n}'\hat{\boldsymbol{\tau}})/k. \tag{38.20}$$

Из (38.20) и (38.11) получаем для исходных параметров блоков

$$\hat{\boldsymbol{\beta}} = \boldsymbol{u}'\hat{\boldsymbol{\alpha}} = \boldsymbol{u}'\boldsymbol{u}(\boldsymbol{B} - \boldsymbol{n}'\hat{\boldsymbol{\tau}})/k. \tag{38.21}$$

38.23 Для упрощения вычислений оценки (38.19) и (38.21) можно привести к более простому виду. Поскольку матрица $\boldsymbol{U}_0$ из **38.20** ортогональна, то

$$\boldsymbol{I}_b = \boldsymbol{U}_0'\boldsymbol{U}_0 = (b^{-1/2}\mathbf{1}_b')'(b^{-1/2}\mathbf{1}_b') + \boldsymbol{u}'\boldsymbol{u},$$

так что

$$\boldsymbol{u}'\boldsymbol{u} = \boldsymbol{I}_b - \mathbf{1}_b\mathbf{1}_b'/b. \tag{38.22}$$

Подставляя (38.22) в (38.16) и вспоминая (38.3), получим

$$\boldsymbol{\Omega} = \{\operatorname{diag}(\boldsymbol{r}) - \boldsymbol{nn}'/k + \boldsymbol{rr}'/(bk)\}^{-1}. \tag{38.23}$$

Оценка (38.19) аналогично переписывается в виде

$$\hat{\boldsymbol{\tau}} = \boldsymbol{\Omega}\{\boldsymbol{T} - \boldsymbol{nB}/k + \boldsymbol{r}G/(bk)\}, \tag{38.24}$$

где G обозначает полную сумму значений всех y, причем

$$G = \mathbf{1}_b'\boldsymbol{B} = \mathbf{1}_t'\boldsymbol{T}. \tag{38.25}$$

Правую часть (38.24) можно упростить и дальше. В самом деле, согласно (38.23), имеем

$$\boldsymbol{\Omega}^{-1}\mathbf{1}_t = \boldsymbol{r} - \boldsymbol{n}(\boldsymbol{n}'\mathbf{1}_t)/k + \boldsymbol{r}(\boldsymbol{r}'\mathbf{1}_t)/(bk). \tag{38.26}$$

Но из соотношений (38.3—4) и их следствия

$$\boldsymbol{r}'\mathbf{1}_t = bk \tag{38.27}$$

следует, что последние два члена в (38.26) совпадают, а значит, $\boldsymbol{\Omega}^{-1}\mathbf{1}_t = \boldsymbol{r}$. Таким образом,

$$\boldsymbol{\Omega r} = \mathbf{1}_t, \tag{38.28}$$

и (38.24) переписывается в виде

$$\hat{\boldsymbol{\tau}} = \boldsymbol{\Omega}(\boldsymbol{T} - \boldsymbol{nB}/k) + \mathbf{1}_t G/(bk). \tag{38.29}$$

В **38.25** мы увидим, что последний член в правой части (38.29), по существу, дает оценку генерального среднего, которое мы опустили в нашей линейной модели.

Подставляя далее (38.22) в (38.21), получим

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{I}_b - \mathbf{1}_b\mathbf{1}_b'/b)(\boldsymbol{B} - \boldsymbol{n}'\hat{\boldsymbol{\tau}})/k. \tag{38.30}$$

Это выражение снова можно упростить, поскольку, используя (38.25) и (38.3) и подставляя затем значение $\hat{\boldsymbol{\tau}}$ из (38.29), получим

$$\mathbf{1}_b'(\boldsymbol{B} - \boldsymbol{n}'\hat{\boldsymbol{\tau}}) = G - \boldsymbol{r}'\hat{\boldsymbol{\tau}} = G - \boldsymbol{r}'\boldsymbol{\Omega}(\boldsymbol{T} - \boldsymbol{nB}/k) - \boldsymbol{r}'\mathbf{1}_t G/(bk).$$

Но, как следует из (38.25), (38.4) и (38.27), это выражение равно нулю. А значит,

$$\boldsymbol{r}'\hat{\boldsymbol{\tau}} = G, \tag{38.31}$$

и (38.30) сводится просто к

$$\hat{\beta} = (B - n'\hat{\tau})/k. \qquad (38.32)$$

Кроме вычисления Ω из (38.23), для получения оценок параметров обработок и оценок параметров блоков нужны только векторы суммарных значений наблюдений по обработкам и по блокам, соответственно T и B и полная сумма значений G, которая в соответствии с (38.25) может быть получена из любого из этих векторов.

ДА для блочных экспериментов

38.24 Для проведения ДА для блочного эксперимента нужно вычислить остаточную СК. Она равна

$$S_0 = (y - X\hat{\theta})'(y - X\hat{\theta}) \equiv y'y - y'X\hat{\theta},$$

что сводится с использованием (38.18) и (38.11) к

$$S_0 = y'y - \begin{pmatrix} T \\ uB \end{pmatrix}' \begin{pmatrix} \hat{\tau} \\ \hat{\alpha} \end{pmatrix} = y'y - T'\hat{\tau} - B'u'\hat{\alpha} =$$

$$= y'y - T'\hat{\tau} - B'\hat{\beta}. \qquad (38.33)$$

В общем случае ДА не является ортогональным, поскольку внедиагональные матрицы в (38.17) не являются нулевыми. Поэтому, как и в примере 35.4, в **35.38** и **35.43** нужно найти остаточную СК, скажем S_1, для случая, когда между обработками нет разницы, а оцениваются только параметры блоков и единый для всех блоков параметр обработок. Тогда разность $S_1 - S_0$ будет представлять собой СК, соответствующую разнице между обработками.

38.25 В предположении, что между обработками нет разницы, оценка $\hat{\tau}$ сводится к $\hat{\tau}\mathbf{1}_t$. Тогда (38.24) сводится к

$$\hat{\tau}\mathbf{1}_t = \Omega\{T - nB/k + rG/(bk)\}.$$

В соответствии с (38.28) подставим сюда Ωr вместо $\mathbf{1}_t$. Умножая слева на Ω^{-1}, транспонируя и умножая затем справа на $\mathbf{1}_t$, получим

$$\hat{\tau}r'\mathbf{1}_t = \{T - nB/k + rG/(bk)\}'\mathbf{1}_t.$$

Но первые два члена в правой части совпадают, поскольку из (38.25) и (38.4) следует, что

$$\mathbf{1}_t'(T - nB/k) = 0. \qquad (38.34)$$

В результате получаем

$$\hat{\tau} = G/(bk)$$

или

$$\hat{\tau}\mathbf{1}_t = \mathbf{1}_t G/(bk), \qquad (38.35)$$

что интуитивно понятно, поскольку, в случае отсутствия разницы между обработками, генеральное среднее $G/(bk)$ является оценкой для всех обработок. Далее, используя (38.35) и (38.4), получим, что в рассматриваемом случае (38.32) сводится к

$$(\hat{\beta})_{\hat{\tau}=\hat{\tau}\mathbf{1}_t} = (B - \mathbf{1}_b G/b)/k. \qquad (38.36)$$

Подставляя (38.35—6) в (38.33) и используя (38.25), получим

$$S_1 = y'y - B'B/k, \qquad (38.37)$$

так что (как читателю предлагается проверить в упражнении 38.1)

$$S_1 - S_0 = T'\hat{\tau} + B'\hat{\beta} - B'B/k = (T - nB/k)'\Omega(T - nB/k) \qquad (38.38)$$

представляет собой СК, соответствующую разнице между обработками, в то время как, согласно (38.37), объединенная СК для блоков и генерального среднего равна $B'B/k$.

38.26 Мы можем собрать все полученные результаты в следующую таблицу.

Таблица ДА для общего блочного эксперимента

Источник дисперсии	СК	Число ст. св.
Разница между обработками (учитывающая эффекты блоков)	$T'\hat{\tau} + B'\hat{\beta} - B'B/k = (T - nB/k)'\Omega(T - nB/k)$	$t - 1$
Эффекты блоков (без учета разницы между обработками)	$B'B/k - G^2/(bk)$	$b - 1$
Остаточная	$y'y - T'\hat{\tau} - B'\hat{\beta}$	$bk - b - t + 1$
Генеральное среднее	$G^2/(bk)$	1
Полная	$y'y$	bk

(38.39)

Число ст. св. для остаточной СК получается как разность. Из (38.39) понятно, что наш анализ сводится просто к выделению члена с $t-1$ ст. св., соответствующего разнице между обработками, после того как уже выделен член, соответствующий $b-1$ линейно независимым параметрам блоков $\boldsymbol{\alpha}$ и генеральному среднему.

Планирование блочных экспериментов

38.27 При проведении описанного выше анализа самым сложным является вычисление матрицы $\boldsymbol{\Omega}$ из (38.23), которое проводится с использованием обращения матриц. Матрица $\boldsymbol{\Omega}$ зависит от матрицы инциденций $\boldsymbol{n}$ и вектора $\boldsymbol{r}$, который получается из $\boldsymbol{n}$ в соответствии с (38.3). Поскольку, согласно (19.16), $\sigma^2(X'X)^{-1}$ является дисперсионной матрицей вектора оценок $\hat{\boldsymbol{\theta}}$, из (38.17) следует, что

$$\boldsymbol{V}(\hat{\boldsymbol{\tau}}) = \sigma^2\boldsymbol{\Omega}. \tag{38.40}$$

Мы попытаемся найти *план* эксперимента (т. е. выбрать $\boldsymbol{n}$) так, чтобы дисперсионная матрица оценок параметров обработок имела некоторую желаемую форму. При определении этой формы мы будем накладывать интуитивно приемлемые условия, которые приведут нас к большинству обычно используемых планов экспериментов. Кифер (1958, 1959) рассматривает различные концепции оптимальности при планировании эксперимента и показывает, что планы, симметричные относительно всех обработок (такие, к которым мы придем, наложив симметричные условия на все оценки параметров обработок), оптимальны в большинстве из рассматриваемых смыслов. Карлин и Стадден (1966) дают обзор общей теории и приводят обобщения.

В частности, эти симметричные планы минимизируют обобщенную дисперсию*) (определитель матрицы (38.40)) и обеспечивают локальную оптимальность по мощности критерия проверки равенства всех параметров обработок. Стоит отметить, что эти свойства оптимальности не сохраняются, если сами планы можно выбирать с использованием случайной процедуры, хотя обобщенная дисперсия остается минимальной, когда размеры блоков k стремятся к бесконечности.

Некоторые теоретические выводы и обсуждение такого рода *случайного выбора планов* (включая *случайные сбалансированные*) содержатся в работах Дэмпстера (1960—1), Сэттерсвейта (1959), Буднэ (1959). Две последние статьи обсуждаются в работах Юдена, Кемпсорна, Тьюки, Бокса, Хантера (1959) и Энскомба (1959).

*) Из 19.8 т. 2 мы знаем, что *внутри* любого плана НК-оценки минимизируют обобщенную дисперсию средн линейных оценок. Упоминаемый выше результат относится к выбору *между* планами.

38.28 Если выбрать $\boldsymbol{n}$ так, чтобы сделать $\boldsymbol{\Omega}$ диагональной, то оценки параметров обработок становятся некоррелированными (ортогональными) и, кроме того, нужное нам матричное обращение становится тривиальным. Но матрица $\boldsymbol{\Omega}$ диагональна тогда и только тогда, когда диагональна обратная к ней матрица, а поскольку первый член в фигурных скобках в (38.23) уже диагонален, нам нужно потребовать только, чтобы матрица $\boldsymbol{M} = \boldsymbol{nn}' - \boldsymbol{rr}'/b$ была диагональной. Далее, используя (38.4) и (38.27), получим

$$\mathbf{1}_t'\boldsymbol{M}\mathbf{1}_t = \mathbf{1}_t'(\boldsymbol{nn}' - \boldsymbol{rr}'/b)\,\mathbf{1}_t = k\mathbf{1}_b'k\mathbf{1}_b - (bk)^2/b = 0. \tag{38.41}$$

Таким образом, сумма всех элементов матрицы $\boldsymbol{M}$ всегда равна нулю. Если матрица $\boldsymbol{M}$ должна быть диагональной, то все ее внедиагональные элементы должны равняться нулю, а значит, нулю должна равняться сумма ее диагональных элементов, которые мы обозначим M_{ii}, т. е. с использованием (38.3)

$$0 = \sum_{i=1}^{t} M_{ii} = \sum_{i=1}^{t}\left(\sum_{j=1}^{b} n_{ij}^2 - r_i^2/b\right) = \sum_{i=1}^{t}\sum_{j=1}^{b}(n_{ij} - r_i/b)^2. \tag{38.42}$$

Отсюда для всех i, j должно выполняться

$$n_{ij} - r_i/b = 0, \tag{38.43}$$

или в матричных обозначениях

$$\boldsymbol{n} = \boldsymbol{r}\mathbf{1}_b'/b. \tag{38.44}$$

Таким образом, условие диагональности $\boldsymbol{\Omega}$ сводится к тому, что каждый блок содержит одно и то же множество обработок, а суммарное число объектов, к которому применяется каждая обработка, кратно b. Поскольку каждая обработка встречается в каждом блоке 0 или 1 раз и по крайней мере 1 раз в эксперименте как целом, то отсюда следует, что каждая обработка встречается в каждом блоке ровно 1 раз. А значит, $k = t$ и

$$\boldsymbol{r} = b\mathbf{1}_t, \tag{38.45}$$

так что (38.44) сводится просто к

$$\boldsymbol{n} = \mathbf{1}_t\mathbf{1}_b'. \tag{38.46}$$

Таким образом мы получили матрицу инциденций плана *случайных блоков*.

Относящиеся сюда результаты содержатся в упражнениях 38.2—3.

Планы случайных блоков

38.29 Таким образом, после длинных рассуждений мы пришли к изучению плана случайных блоков, который упоминался в **36.39**, когда впервые рассматривалось рандомизированное размещение объектов между обработками. Структура эксперимента со случайными блоками чрезвычайно проста: t обработок случайным образом размещаются между $k = t$ объектами в каждом из b блоков. В силу диагональности матрицы $\mathbf{\Omega}$, откуда в **38.28** и был получен соответствующий план, НК-оценки параметров и общая таблица ДА для блочных экспериментов (38.39) существенно упрощаются. Подставляя (38.45—6) в (38.23), получим

$$\mathbf{\Omega} = \mathbf{I}_t/b, \tag{38.47}$$

в то время как подстановка (38.46) и (38.25) в (38.29) и (38.32) дает

$$\hat{\boldsymbol{\tau}} = \mathbf{T}/b, \tag{38.48}$$

$$\hat{\boldsymbol{\beta}} = \mathbf{B}/t - \mathbf{1}_b G/(bt). \tag{38.49}$$

Аналогично СК для разницы между обработками (38.38) принимает вид

$$S_1 - S_0 = \mathbf{T}'\mathbf{T}/b - G^2/(bt), \tag{38.50}$$

а остаточная СК сводится к

$$S_0 = \mathbf{y}'\mathbf{y} - \mathbf{T}'\mathbf{T}/b - \mathbf{B}'\mathbf{B}/t + G^2/(bt). \tag{38.51}$$

В упражнении 38.4 читателю предлагается проверить эти формулы.

38.30 Приведем теперь таблицу упрощенного ДА.

Таблица ДА для эксперимента со случайными блоками

Источник дисперсии	СК	Число ст. св.
Разница между обработками	$\mathbf{T}'\mathbf{T}/b - G^2/(bt)$	$t-1$
Разница между блоками	$\mathbf{B}'\mathbf{B}/t - G^2/(bt)$	$b-1$
Остаточная	$\mathbf{y}'\mathbf{y} - \mathbf{T}'\mathbf{T}/b - \mathbf{B}'\mathbf{B}/t + G^2/(bt)$	$(t-1)(b-1)$
Генеральное среднее	$G^2/(bt)$	1
Полная	$\mathbf{y}'\mathbf{y}$	bt

(38.52)

Сравнивая с (38.39), видим, насколько велико упрощение. Из симметрии таблицы (38.52) относительно обработок (символы $\mathbf{T}$ и t) и блоков (символы $\mathbf{B}$, b) становится понятно, почему блоки рассматриваются как (мешающие) факторы анализа, поскольку (38.52) формально совпадает с таблицей ДА для двухфакторной перекрестной классификации с одним наблюдением в клетке (см. пример 35.3 и упражнение 35.1). Как и всегда, в таком анализе отсутствует СК взаимодействий, хотя, как и в примере 35.3, можно из остаточной СК выделить член с одной ст. св. для проверки взаимодействий. Таким образом, можно проверить предположение модели о том, что обработки и блоки не взаимодействуют.

Пример 38.2

В **38.10** мы видели, что пренебрежение временем дня, когда проводятся испытания в эксперименте из примера 38.1, вызывает серьезные возражения. Если рассматривать теперь время дня как мешающий фактор и провести эксперимент с использованием случайных блоков (каждый блок соответствует некоторому времени дня), то упомянутые возражения отпадают.

Экономия степеней свободы: два мешающих фактора

38.31 Из (38.39) видно, что при использовании b блоков нужно выделять член с $b-1$ ст. св. Если из-за большой разницы между объектами даже внутри блоков нужно сделать число объектов k в блоке достаточно малым, то для получения заданного числа наблюдений нужно соответственно увеличить число блоков. Но тогда слишком большое число ст. св. будет относиться к блокам, и, возможно, придется искать другой план эксперимента, чтобы избежать этого, поскольку ситуация, когда bt мало, а b намного больше, чем t, оставляет желать лучшего.

Чтобы сэкономить число ст. св., теряемых на устранении блочных эффектов, нужно договориться, что не все блоки требуют отдельных параметров, как это было в **38.19**. Простой путь, приводящий к этому, заключается в следующем. Считается, что сами блоки образуют двухфакторную классификацию, скажем с n строчками и m столбцами. При этом, как и ранее, предполагается, что каждый блок содержит k объектов. Кроме того, предполагается, что $nm = b$, а для параметров блоков β_{lj} имеет место

$$\beta_{lj} = \rho_l + \gamma_j, \qquad l = 1, \ldots, n; \; j = 1, \ldots, m. \tag{38.53}$$

Так поступают, например, при использовании блоков для устранения двух мешающих факторов, соответствующих классификации блоков по строчкам и столбцам.

Как и ранее, для того чтобы можно было выделить эффекты обработок, наложим следующие ограничения:

$$\sum_{i=1}^{n} \rho_i = \sum_{j=1}^{m} \gamma_j = 0. \qquad (38.54)$$

При этом окажется, что блокам будет соответствовать только $(n-1)+(m-1)$ ст. св.

Теперь можно кратко описать анализ по методу НК, который вполне аналогичен анализу из **38.15—26.**

38.32 Как и в **38.16**, теперь имеется nm матриц обработок t_{ij}. Пусть n-матрица инциденций порядка $(t \times n)$, соответствующая строчкам, так что n_{il} равно единице или нулю в соответствии с тем, применяется или нет i-я обработка в l-й строке (вне зависимости от того, в каком столбце). Аналогично m-матрица инциденций столбцов, порядка $(t \times m)$. Как и в (38.3), имеет место

$$n\mathbf{1}_n = m\mathbf{1}_m = r. \qquad (38.55)$$

Чтобы выполнялось (38.54), соотношение (38.8) заменяется на

$$\mathbf{1}_n'\rho = \mathbf{1}_m'\gamma = 0, \qquad (38.56)$$

а вместо α, определенного в (38.10), имеем

$$\left.\begin{aligned} \sigma &= u\rho, \\ \delta &= v\gamma, \end{aligned}\right\} \qquad (38.57)$$

где σ имеет порядок $n-1$, δ — порядок $m-1$, а u и v являются аналогами u из **38.20.**

Соотношение (38.12) остается справедливым, только теперь y — вектор, состоящий из nm векторов y_{ij},

$$\underset{nmk \times (t+n+m-2)}{X} = \left[\begin{array}{c|c|c} t_{11} & \mathbf{1}_k u_1' & \mathbf{1}_k v_1' \\ \vdots & \vdots & \vdots \\ t_{ij} & \mathbf{1}_k u_i' & \mathbf{1}_k v_j' \\ \vdots & \vdots & \vdots \\ t_{nm} & \mathbf{1}_k u_n' & \mathbf{1}_k v_m' \end{array}\right], \qquad (38.58)$$

а

$$\underset{(t+n+m-2)\times 1}{\theta} = \begin{pmatrix} \tau \\ \cdots \\ \sigma \\ \cdots \\ \delta \end{pmatrix}. \qquad (38.59)$$

В результате получаем (см. (38.15))

$$X'X = \left(\begin{array}{c|c|c} \operatorname{diag}(r) & nu' & mv' \\ un' & mkI_{n-1} & 0 \\ vm' & 0 & nkI_{m-1} \end{array}\right), \qquad (38.60)$$

и если по аналогии с (38.23) обозначить

$$\Omega = \{\operatorname{diag}(r) - nn'/(mk) - mm'/(nk) + 2rr'/(nmk)\}^{-1}, \qquad (38.61)$$

то матрица, обратная к (38.60), имеет вид

$$(X'X)^{-1} = \left[\begin{array}{c|c|c} \Omega & -\Omega nu'/(mk) & -\Omega mv'/(nk) \\ \hline -un'\Omega/(mk) & \{I_{n-1} + un'\Omega nu'/(mk)\}/(mk) & un'\Omega mv'/(nmk^2) \\ \hline -vm'\Omega/(nk) & vm'\Omega nu'/(nmk^2) & \{I_{m-1} + vm'\Omega mv'/(nk)\}/(nk) \end{array}\right], \qquad (38.62)$$

так что соотношение (38.40) остается верным. Так же как в (38.18), можно положить

$$X'y = \begin{pmatrix} T \\ uR \\ vC \end{pmatrix}, \qquad (38.63)$$

где T, R и C — соответственно векторы суммарных значений для обработок, строк и столбцов. Как и ранее, полная сумма будет обозначаться G.

38.33 НК-оценки получаются умножением (38.62) на (38.63). Вместо (38.19) теперь получим

$$\hat{\tau} = \Omega\{T - nu'uR/(mk) - mv'vC/(nk)\}. \qquad (38.64)$$

Аналогично выводу (38.24) это выражение можно упростить и привести к виду

$$\hat{\tau} = \Omega\{T - nR/(mk) - mC/(nk) + 2rG/(mnk)\}. \qquad (38.65)$$

Аналогично выводу (38.32) получим

$$\left.\begin{aligned} \hat{\rho} &= (R - n'\hat{\tau})/(mk), \\ \hat{\gamma} &= (C - m'\hat{\tau})/(nk). \end{aligned}\right\} \qquad (38.66)$$

Таким образом, остаточная СК, так же как в (38.33), имеет вид

$$S_0 = y'y - T'\hat{\tau} - R'\hat{\rho} - C'\hat{\gamma}, \qquad (38.67)$$

а если, как и в 38.25,

$$\hat{\tau} = \hat{\tau}\mathbf{1}_t, \tag{38.68}$$

то остаточная СК примет вид

$$S_1 = \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{R}'\boldsymbol{R}/(mk) - \boldsymbol{C}'\boldsymbol{C}/(nk) + G^2/(nmk), \tag{38.69}$$

так что из (38.67) и (38.69) получим, что СК, соответствующая разнице между обработками, равна

$$S_1 - S_0 = \boldsymbol{P}'\boldsymbol{\Omega}\boldsymbol{P} - G^2/(nmk), \tag{38.70}$$

где $\boldsymbol{P}$ — матрица, стоящая в фигурных скобках в правой части (38.65). Соотношение (38.70) является аналогом (38.38). Таблицу ДА можно теперь построить по аналогии с (38.39), причем эффектам блоков будет соответствовать член, имеющий всего $n + m - 2$ ст. св. Проверку приведенных выше формул и построение таблицы мы оставляем читателю в качестве упражнения 38.5.

Латинские квадраты

38.34 Устранение двух мешающих факторов методами **38.31—3** чаще всего проводится при $n = m$ и $k = 1$. Тогда каждый блок состоит из одного объекта, а объекты расположены в виде квадратной таблицы. В **38.31—3** мы ничего не говорили о том, как t обработок размещаются между блоками. Предположим теперь, что $t = n = m$, так что мы имеем таблицу из t^2 объектов, между которыми случайным образом размещаются t обработок. Если предположить, что каждая обработка применяется ровно один раз в каждой строке и ровно один раз в каждом столбце, то мы получим расположение, известное как *латинский квадрат*. Ниже в (38.71) приведен пример с $t = 4$ и обработками, обозначенными A, B, C, D:

$$\begin{matrix} A & B & C & D \\ B & A & D & C \\ C & D & A & B \\ D & C & B & A \end{matrix} \tag{38.71}$$

В восемнадцатом веке Эйлер много занимался изучением латинских квадратов с чисто математической точки зрения. Тот факт, что латинские квадраты стали использоваться при планировании экспериментов в нашем столетии, служит замечательным примером того, как совершенно бесполезные на вид теории могут получить в конце концов большое практическое значение.

38.35 Если смотреть только на строки в (38.71), то мы получим план со случайными блоками с $b = t$. То же самое справедливо, если смотреть только на столбцы. Таким образом, можно воспользоваться таблицей ДА (38.52) с $b = t$, причем во второй строчке этой таблицы можно брать *любую* разность между блоками, соответствующую строчкам или столбцам в (38.71). Это дает основание ожидать, что можно разделить два этих эффекта и получить приведенную ниже таблицу.

Источник дисперсии	Число ст. св.
Разница между обработками	$t - 1$
Строчки	$t - 1$
Столбцы	$t - 1$
Остаточная	$(t - 1)(t - 2)$
Генеральное среднее	1
Полная	t^2

(38.72)

И в самом деле, это непосредственно выводится из результатов **38.33**, если положить $n = m = t$ и $k = 1$. Читателю предлагается проделать это в упражнении 38.6.

Если два мешающих фактора могут взаимодействовать, что противоречит предположению об их аддитивности, сделанному в (38.53), то это может привести к серьезным ошибкам (см. Шеффе (1959)).

38.36 Для плана со случайными блоками нет трудности при выборе одного плана случайным образом. Нужно только иметь случайные перестановки чисел от 1 до t (см. **9.14** и пример 9.7), чтобы с равными вероятностями сделать выбор среди $(t!)^b$ возможных размещений экспериментальных объектов по обработкам.

Однако в случае латинских квадратов выбор плана случайным образом не так прост, поскольку совсем не ясно, как много существует латинских квадратов заданного порядка, хотя из рассмотрения циклических перестановок элементов первой строки очевидно, что для любого порядка существуют латинские квадраты.

Число возможных латинских квадратов порядка t весьма велико при больших значениях t. Так, например, имеется 576 квадратов порядка 4; 161 280 квадратов порядка 5; и 812 851 200 квадратов порядка 6. Их число подсчитано для порядков вплоть до седьмого. Хотя известно много примеров квадратов более высоких порядков, но вычисление их числа, когда порядок

больше или равен восьми, все еще остается нерешенной задачей. Детали и примеры можно найти в «статистических таблицах» Фишера и Иэйтса.

Перестановкой строк и столбцов каждый квадрат может быть приведен к такому виду, когда верхняя строка и левый столбец расположены в порядке ABC и т. д. В этом случае говорят о «стандартном квадрате». Например, имеется четыре стандартных квадрата четвертого порядка:

$$\begin{matrix} A & B & C & D \\ B & A & D & C \\ C & D & B & A \\ D & C & A & B \end{matrix} \quad \begin{matrix} A & B & C & D \\ B & C & D & A \\ C & D & A & B \\ D & A & B & C \end{matrix} \quad \begin{matrix} A & B & C & D \\ B & D & A & C \\ C & A & D & B \\ D & C & B & A \end{matrix} \quad \begin{matrix} A & B & C & D \\ B & A & D & C \\ C & D & A & B \\ D & C & B & A \end{matrix} \tag{38.73}$$

Из каждого приведенного в (38.73) квадрата можно образовать 144 (=4!3!) квадратов, перестановкой всех столбцов и всех строк, за исключением первой. (Не имеет смысла делать перестановку первой строки, поскольку в результате получится повторение уже имеющегося квадрата с переобозначением букв $A, \ldots, D$, а такое переобозначение не играет никакой роли). Таким образом, как уже упоминалось, полное число квадратов четвертого порядка равно $4 \times 144 = 576$. В общем случае каждый стандартный квадрат порядка t порождает $t!(t-1)!$ квадратов.

Таким образом, необходимо только описать стандартные квадраты. Чтобы выбрать латинский квадрат случайным образом, надо сначала случайным образом выбрать стандартную форму, а затем переставить случайным образом строки и столбцы, причем самый удобный процесс рандомизации состоит в использовании таблиц случайных перестановок (см. **9.14** и пример 9.7). Для квадратов, порядок которых больше или равен восьми, для которых не описаны все стандартные типы, можно выбирать только среди тех типов, которые описаны. А значит, случайный выбор будет производиться не из всего множества возможных квадратов, а только из некоторого его подмножества.

Три и больше мешающих фактора: греко-латинские и ортогональные квадраты

38.37 Нетрудно обобщить понятие латинского квадрата, чтобы получить устранение трех и более мешающих факторов. Это можно сделать процессом наложения различных латинских

квадратов. Если латинский квадрат

$$\begin{matrix} A & B & C & D \\ C & D & A & B \\ D & C & B & A \\ B & A & D & C \end{matrix} \tag{38.74}$$

наложить на квадрат (38.71), заменив предварительно латинские буквы в (38.74) на соответствующие греческие, чтобы не возникло путаницы, то мы придем к расположению

$$\begin{matrix} A\alpha & B\beta & C\gamma & D\delta \\ B\gamma & A\delta & D\alpha & C\beta \\ C\delta & D\gamma & A\beta & B\alpha \\ D\beta & C\alpha & B\delta & A\gamma \end{matrix} \tag{38.75}$$

в котором каждая комбинация греческой и латинской буквы встречается ровно один раз. По этой причине квадраты (38.71) и (38.74) называются *ортогональными* (*латинскими*) *квадратами*. Их наложение (38.75) называется *греко-латинским квадратом*.

Очевидно, что при $t = 2$ не существует греко-латинского квадрата. Более интересным является то, что не существует греко-латинского квадрата при $t = 6$, хотя при $t = 6$ имеется 812 851 200 латинских квадратов. Эйлер предположил, что греко-латинские квадраты не существуют при $t = 4r + 2$, и понадобилось почти два столетия, чтобы опровергнуть его предположение и показать (см. Бозе и Шриканде (1959), Бозе, Шриканде и Паркер (1960)), что, за исключением $t = 2$ и $t = 6$, всегда существуют греко-латинские квадраты. В таблицах Фишера и Иэйтса приводятся примеры.

Греческие буквы в (38.75) можно использовать для идентификации третьего мешающего фактора (строки и столбцы, как и ранее, идентифицируют первые два, а латинские буквы относятся к обработкам). Тогда план устраняет влияние трех мешающих факторов точно так же, как латинский квадрат устраняет влияние двух. Таблица ДА является очевидным обобщением упражнения **38.6**, и мы оставляем читателю ее построение.

38.38 На (38.75) можно наложить еще латинский квадрат (с использованием третьего множества символов, например, чисел) так, чтобы каждая комбинация из двух множеств символов встречалась ровно один раз. Три таких латинских квадрата называются ортогональными в совокупности. Это спра-

ведливо для расположения

$$\begin{matrix} A\alpha 1 & B\beta 2 & C\gamma 3 & D\delta 4 \\ B\gamma 4 & A\delta 3 & D\alpha 2 & C\beta 1 \\ C\delta 2 & D\gamma 1 & A\beta 4 & B\alpha 3 \\ D\beta 3 & C\alpha 4 & B\delta 1 & A\gamma 2 \end{matrix} \qquad (38.76)$$

которое называется *гипер-греко-латинским квадратом*. Если греческие буквы и числа используются для идентификации третьего и четвертого мешающих факторов, то план устраняет влияние четырех мешающих факторов. Построение таблицы снова предоставляется читателю в упражнении 38.6.

38.39 На (38.76) нельзя наложить ни одного латинского квадрата с сохранением ортогональности. Существует не более $t-1$ взаимно ортогональных латинских квадрата порядка t. Простое доказательство этого факта оставляется читателю в виде упражнения 38.7. Множество из $t-1$ ортогональных квадратов порядка t, вида (38.76) называется *полным множеством* ортогональных латинских квадратов. Такие полные множества существуют, если t — простое число или степень простого числа (см. Х. Манн (1949)). Таким образом, при $t \leqslant 9$ полное множество существует всегда, за исключением значений $t=2$ и $t=6$, при которых, как мы видели в **38.37**, не существует даже двух ортогональных квадратов. При $10 \leqslant t \leqslant 30$ Барра (1965) приводит некоторые факты об известных множествах ортогональных квадратов. Полные множества описаны при $t \leqslant 7$. В таблицах Фишера и Иэйтса приводятся примеры для $t \leqslant 9$.

За дальнейшими теоретическими фактами и конструкциями латинских квадратов и состоящих из них ортогональных множеств, о которых мы здесь не упоминали, читатель отсылается к двум монографиям Вайда (1967a, b), к раннему отчету Манна (1949) о методах Р. Бозе и к обзорной статье Барра (1965), в которой приводится библиография по данному вопросу, продолжающая обзор Нортона (1939).

38.40 Практическое использование латинских квадратов в экспериментальной работе ограничено из-за условия, что число обработок должно совпадать с числом уровней каждого из мешающих факторов. Это условие становится еще более ограничительным, когда мы переходим к греко-латинскому квадрату и ортогональным множествам квадратов более высоких порядков. В связи с этим последние расположения используются довольно редко. Однако сами латинские квадраты часто используются в сельскохозяйственных экспериментах, когда строчки и столбцы квадратной таблицы соответствуют фактическому расположению экспериментальных участков (объектов). При этом градиент плодородия почвы вдоль этих двух направлений для занятой под эксперимент площади не будет оказывать влияния на обработки. Конечно, градиент плодородия может быть и в других направлениях, например по диагонали к квадратной таблице. Тогда применение латинского квадрата не приводит к устранению мешающих влияний. Однако обычно экспериментатор выбирает ориентацию своих строк и столбцов так, чтобы устранить известные или предполагаемые градиенты плодородия.

Понятно, что расположение в виде латинского квадрата можно использовать всякий раз, когда нужно устранить влияние двух географических или временных координат. Аналогично расположение с более высокими порядками можно применять, если имеется три или больше таких мешающих факторов.

Пример 38.3

В эксперименте, рассмотренном в примерах 38.1—2, можно предположить, что день недели, когда проводится испытание, также влияет на результат. Гипотеза состоит в том, что существует некоторая накопляемость усталости в течение рабочей недели, которая оказывает такое же воздействие, как обсуждавшееся ранее «время дня». Устранить влияние этих двух мешающих факторов можно, рассматривая столько промежутков времени дня, сколько дней в рабочей неделе (скажем, 5), и представляя эксперимент в виде плана латинского квадрата размера 5×5. Заметим, что для отдельного квадрата можно брать только 5 обработок (доз алкоголя). Однако ничто не мешает использовать столько (5×5)-квадратов, сколько требуется для проверки всех предложенных обработок. Нужно только, чтобы число обработок было кратно 5.

Если, помимо времени дня и дня недели, рассматривать место работы как третий мешающий фактор, влияющий на эксперимент, то можно выбирать принимающих участие в эксперименте водителей из пяти мест работы и представить эксперимент в виде греко-латинского квадрата. Но необходимость выбора именно пяти мест работы и именно пяти промежутков времени дня, только потому, что рабочая неделя состоит из пяти дней, часто делает соответствующие планы неудобными.

Рандомизация и устойчивость для случайных блоков и латинских квадратов

38.41 В **38.18** мы оставили в стороне вопрос о влиянии на выводы случайного размещения обработок между объектами внутри блоков при блочных экспериментах. Тот же вопрос возникает по отношению к случайному размещению обработок по

объектам в латинском квадрате (см. **38.34**). Теперь мы должны детально рассмотреть этот вопрос.

Приведенный вопрос можно рассматривать с двух точек зрения. Во-первых, поскольку при рандомизации мы проводим (в виде предостережения от смещенности выводов) некоторые физические действия при размещении обработок между объектами, можно спросить, как это повлияет на справедливость выводов, которые делаются с использованием линейной модели с гомоскедастичными некоррелированными ошибками, становящимися независимыми, если предположить их нормальность (что необходимо делается при проверке гипотез). С этой точки зрения приведенный вопрос состоит в устойчивости наших выводов.

Однако вопрос можно поставить более прямо, преследуя более широкие цели. Рандомизация порождает равновероятные множества наблюдений, а в главе 31 и далее мы видели, что рассмотрение перестановочных распределений может позволить нам заменить теорию нормального распределения на более общие (свободные от распределения) методы. Если выполняются предположения нормальности, то уменьшение эффективности при применении этих методов (по сравнению с методами, использующими нормальность) либо мало, либо вообще отсутствует, зато если предположения нормальности не выполняются, то их эффективность может быть намного больше. Можно спросить, не будет ли рандомизация играть ту же роль, т. е. не позволит ли она освободить нас от необходимости делать предположения нормальности.

38.42 Детальное рассмотрение теории рандомизации для случайных блоков и латинских квадратов содержится в предпоследней главе книги Шеффе (1959), где приведены ссылки на литературу. С точки зрения оценивания самыми интересными являются результаты о математических ожиданиях СрК в таблицах ДА, которые цитируются из работ Кемпсорна (1952) и Уилка и Кемпсорна (1957) (см. также Кокс (1958b)).

Для плана случайных блоков математическое ожидание СрК обработок получается из математического ожидания остаточного СрК вычитанием члена, зависящего от взаимодействий между блоками и обработками и добавлением обычного члена, зависящего от эффектов обработок. (Стоит отметить, что присутствие ошибок взаимодействий (см. **36.41**) между обработками и объектами *внутри* блоков не меняет ситуацию). Таким образом, остаточный СрК увеличивается из-за взаимодействий между блоками и обработками. Однако эта трудность исчезает, если (как это часто бывает и на самом деле) рассматривать эффекты блоков не как фиксированные, а как случайные величины. Остаточный СрК можно тогда соответствующим образом использовать для оценки величины СрК обработок.

Для латинского квадрата ситуация сложнее, поскольку здесь ошибки взаимодействий оказывают некоторое влияние на сравнение СрК обработок с остаточным СрК. По причинам, о которых говорилось в **36.42**, простые результаты здесь не получаются.

38.43 Рассмотрим теперь вопросы проверки гипотез для эффектов обработок при наличии рандомизации. Для случая, когда нет ошибок объектов (см. **36.41**), соответствующая теория была уже развита в **37.39—41**, где рассматривался критерий перестановок для эффектов столбцов при двухфакторной классификации с одним наблюдением в клетке. В **38.30** мы видели, что ситуация случайных блоков формально совпадает с указанной классификацией. Таким образом, при проверке гипотез об эффектах обработок для случайных блоков справедливы результаты **37.39—41**. При этом строчки играют роль блоков, а столбцы — обработок. А значит, можно в (38.52) применить обычный критерий ДА для обработок с числом ст. св., установленным методом **37.29**, как это указывалось в **37.40**. (Несколько выборочных экспериментов Кольера и Бейкера (1966) указывают, что мощность обычных F-критериев также устойчива к нарушению нормальности.) С другой стороны, можно, как и в **37.41** и упражнении 37.14, использовать свободные от распределения критерии, основанные на рангах или нормальных метках, причем, если число обработок или число блоков не слишком мало, то можно пренебречь поправкой на число ст. св.

Доксум (1967) и Холландер (1967) изучают проверку гипотез об обработках для экспериментов со случайными блоками против альтернатив об упорядочениях вида (31.156). Критерии, основанные на симметричном критерии Вилкоксона (см. **31.79**), обладают очень высокой эффективностью по сравнению с t-критерием Стьюдента при альтернативах сдвига (всегда не менее 0,864 и не менее 0,563 в нормальном случае). Аналог (31.151) много хуже.

Если присутствуют ошибки объектов, то, как показал Огава (1961, 1963), можно все еще оправдать приближенное применение стандартного F-критерия, если дисперсии эффектов объектов почти постоянны внутри блоков, а число блоков достаточно велико.

Для латинских квадратов даже при отсутствии ошибок объектов критерий перестановок для эффектов обработок менее удовлетворителен, чем для случайных блоков. То же было верно для оценок в **38.42**. Как и ранее математическое ожидание обычной для ДА статистики критерия совпадает для нормальной теории и для случая рандомизации, однако в последнем случае выражение для дисперсии более сложное. Вследствие этого остается мало доводов в пользу применения нормального

приближения в случае перестановочного распределения (Уэлч (1937); Шеффе (1959)).

38.44 Таким образом, доводы в пользу справедливости критериев нормальной теории для рандомизированных латинских квадратов весьма сомнительны, и их число даже уменьшается для большинства других, более сложных планов эксперимента. Поэтому вызывает сомнение распространенное безмятежное предположение о том, что рандомизированный случай всегда хорошо приближается нормальной теорией.

Теперь возникает принципиальный вопрос. Должна ли рандомизация явно включаться в теорию, лежащую в основе наших критериев и процедур оценки? На этот вопрос трудно не ответить утвердительно; поскольку рандомизация является важным моментом современного подхода к статистическим выводам (см. 38.7). Поэтому трудно оправдать то относительное пренебрежение этим довольно сложным разделом статистической теории.

Дисперсии разницы между обработками в блочных экспериментах

38.45 Вернемся теперь к задаче планирования для общего анализа блочных экспериментов, проведенного выше в **38.19—26.** В **38.27** мы требовали, чтобы оценки параметров обработок были ортогональны (как было получено в **38.28—30**, это ведет к плану случайных блоков). Вместо этого сформулируем теперь задачу планирования в терминах дисперсии разностей между этими оценками.

Пусть мы хотим, чтобы в результате эксперимента дисперсии разностей между оценками параметров i-й и l-й обработок равнялись, скажем, $2\sigma^2 d_{il}$. Обозначим (i, l)-й элемент дисперсионной матрицы оценок параметров обработок через $\sigma^2 w_{il}$. Согласно (38.40) эта матрица равна $\sigma^2\boldsymbol{\Omega}$, где $\boldsymbol{\Omega}$ определена в (38.23). Тогда

$$w_{ii} - 2w_{il} + w_{ll} = 2d_{il}, \tag{38.77}$$

так что

$$w_{il} = \frac{1}{2}(w_{ii} + w_{ll}) - d_{il}. \tag{38.78}$$

Если обозначить $\boldsymbol{w}$ вектор с элементами $\left\{\frac{1}{2}w_{ii}\right\}$, а $\boldsymbol{D}$ матрицу с элементами $\{-d_{il}\}$, где $d_{ii} = 0$ согласно (38.77), то (38.78) можно переписать в виде

$$\boldsymbol{\Omega} = \boldsymbol{w}\mathbf{1}_t' + \mathbf{1}_t\boldsymbol{w}' + \boldsymbol{D} = (\boldsymbol{w} \mid \mathbf{1}_t)\begin{pmatrix}\mathbf{1}_t' \\ \boldsymbol{w}'\end{pmatrix} + \boldsymbol{D}. \tag{38.79}$$

Нетрудно видеть, что соотношение (38.79) для любого скаляра c совпадает с

$$\boldsymbol{\Omega} = \left(\boldsymbol{w} - \frac{1}{2}c\mathbf{1}_t\right)\mathbf{1}_t' + \mathbf{1}_t\left(\boldsymbol{w} - \frac{1}{2}c\mathbf{1}_t\right)' + (\boldsymbol{D} + c\mathbf{1}_t\mathbf{1}_t'). \tag{38.80}$$

Поскольку само $\boldsymbol{w}$ (как функция от $\boldsymbol{\Omega}$) подлежит выбору при планировании блочного эксперимента, то без потери общности можно продолжать использовать (38.79), условившись, что $\boldsymbol{D}$ может быть заменено на любую матрицу, отличающуюся от $\boldsymbol{D}$ на скаляр, умноженный на $\mathbf{1}_t\mathbf{1}_t'$. Такие матрицы называют матрицами *класса* $\boldsymbol{D}$.

38.46 Итак $\boldsymbol{D}$ равна умноженной на константу матрице дисперсий разностей оценок параметров обработок, которую мы хотим получить. Пусть мы хотим получить эту матрицу такой, чтобы все внедиагональные элементы $\boldsymbol{D}$ были равны, так что мы хотим, чтобы все разности оценивались с одной и той же точностью. (Главная диагональ матрицы $\boldsymbol{D}$, разумеется, состоит из нулей). Согласно (38.77) мы получим такой вид матрицы $\boldsymbol{D}$, если выберем дисперсионную матрицу $\boldsymbol{\Omega}$ оценок параметров обработок так, чтобы были равны друг другу все ее диагональные элементы w_{ii} и равны друг другу внедиагональные элементы w_{il}. В этом случае сам вектор $\boldsymbol{w}$ равен скаляру, умноженному на $\mathbf{1}_t$, так что в (38.80) можно выбрать константу c так, чтобы $\boldsymbol{w} - \frac{1}{2}c\mathbf{1}_t = \mathbf{0}$. Тогда соотношение (38.80) сводится к

$$\boldsymbol{\Omega} = \boldsymbol{D}_c, \tag{38.81}$$

где $\boldsymbol{D}_c$ обозначает соответствующую матрицу из класса $\boldsymbol{D}$. Из (38.81) и (38.23) получим

$$\boldsymbol{D}_c^{-1} = \operatorname{diag}(\boldsymbol{r}) - \boldsymbol{n}\boldsymbol{n}'/k + \boldsymbol{r}\boldsymbol{r}'/(bk),$$

откуда

$$\boldsymbol{n}\boldsymbol{n}' = k\left\{\operatorname{diag}(\boldsymbol{r}) - \boldsymbol{D}_c^{-1} + \boldsymbol{r}\boldsymbol{r}'/(bk)\right\}. \tag{38.82}$$

Из (38.28) и (38.81) имеем

$$\boldsymbol{r} = \boldsymbol{D}_c^{-1}\mathbf{1}_t, \tag{38.83}$$

а из (38.83) с использованием (38.27) получим

$$bk = \mathbf{1}_t'\boldsymbol{D}_c^{-1}\mathbf{1}_t. \tag{38.84}$$

Подставляя (38.83—4) в (38.82), получим

$$\boldsymbol{n}\boldsymbol{n}' = k\left\{\operatorname{diag}(\boldsymbol{D}_c^{-1}\mathbf{1}_t) - \boldsymbol{D}_c^{-1} + \frac{\boldsymbol{D}_c^{-1}\mathbf{1}_t\mathbf{1}_t'\boldsymbol{D}_c^{-1}}{\mathbf{1}_t'\boldsymbol{D}_c^{-1}\mathbf{1}_t}\right\}. \tag{38.85}$$

Хотя (38.85) было получено при специальном предположении о том, что все внедиагональные элементы матрицы $\boldsymbol{D}$ равны, как непосредственно показал Точер (1952), это соотношение остается верным в общем случае для произвольной $\boldsymbol{D}$ (см. упражнение 38.8).

38.47 Возвращаясь к предположениям **38.46**, видим, что если все внедиагональные элементы матрицы $\boldsymbol{D}$ равны $-1/a$ (соответственно все дисперсии разностей равны $2\sigma^2/a$), то

$$\boldsymbol{D} = \frac{1}{a}(\boldsymbol{I}_t - \mathbf{1}_t\mathbf{1}_t'), \tag{38.86}$$

а поскольку

$$\boldsymbol{D}_c = \boldsymbol{D} + c\mathbf{1}_t\mathbf{1}_t',$$

то

$$\boldsymbol{D}_c = \frac{1}{a}\{\boldsymbol{I}_t + (ac-1)\mathbf{1}_t\mathbf{1}_t'\}. \tag{38.87}$$

Нетрудно проверить, что матрица, обратная к (38.87), равна

$$\boldsymbol{D}_c^{-1} = a\left\{\boldsymbol{I}_t - \frac{ac-1}{1+t(ac-1)}\mathbf{1}_t\mathbf{1}_t'\right\}.$$

Полагая $A = (1-ac)/\{1+t(ac-1)\}$, получим

$$\boldsymbol{D}_c^{-1} = a\{\boldsymbol{I}_t + A\mathbf{1}_t\mathbf{1}_t'\}, \tag{38.88}$$

а значит, (38.83) можно переписать в виде

$$\boldsymbol{r} = \boldsymbol{D}_c^{-1}\mathbf{1}_t = a(1+At)\mathbf{1}_t = r\mathbf{1}_t, \tag{38.89}$$

где $r = a(1+At)$. Отсюда следует, что каждая обработка применяется в эксперименте одно и то же число r раз. Используя (38.88—9), получим, что (38.85) можно переписать в виде

$$\boldsymbol{nn}' = k\{(r-a)\boldsymbol{I}_t + (a/t)\mathbf{1}_t\mathbf{1}_t'\}. \tag{38.90}$$

Здесь не фигурирует произвольная константа A, которая нужна только для определения $\boldsymbol{r}$.

Уравнение плана

38.48 Уравнение для $\boldsymbol{nn}'$ называется *уравнением плана*. Из определения $\boldsymbol{n}$ в **38.16** мы видим, что (i, l)-й элемент матрицы $\boldsymbol{nn}'$ указывает на то, сколько раз в эксперименте, как целом, встречаются вместе в одном блоке i-я и l-я обработки. В частности, при $i = l$ диагональные элементы матрицы $\boldsymbol{nn}'$ равны просто частотам r_i, с которыми i-я обработка применяется в эксперименте.

Планы сбалансированных неполных блоков

38.49 Проинтерпретируем теперь приведенное в (38.90) уравнение плана в свете **38.48**. Рассматривая уравнение для диагональных элементов, имеем

$$\{\boldsymbol{nn}'\}_{ii} = r = k\{(r-a) + a/t\},$$

а значит,

$$r = ak(t-1)/\{(k-1)t\}. \tag{38.91}$$

Для внедиагональных элементов (38.90) получим

$$\{\boldsymbol{nn}'\}_{il} = ka/t = \lambda. \tag{38.92}$$

Таким образом, в эксперименте, как в целом, в одном и том же блоке каждая пара обработок встречается ровно λ раз.

Из (38.91—2) получим

$$r(k-1) = \lambda(t-1). \tag{38.93}$$

Поскольку $t \geqslant k$, то из (38.93) следует $\lambda \leqslant r$, что и так ясно из определения этих чисел. Более того, поскольку t обработок встречаются в эксперименте по r раз, а в эксперименте участвуют b блоков, в каждом из которых k объектов, то

$$rt = bk. \tag{38.94}$$

С использованием (38.92—3) соотношение (38.90) можно переписать в виде

$$\boldsymbol{nn}' = (r-\lambda)\boldsymbol{I}_t + \lambda\mathbf{1}_t\mathbf{1}_t'. \tag{38.95}$$

Если $r = \lambda$, то матрица инциденций (38.46) плана случайных блоков удовлетворяет соотношению (38.95), поскольку, согласно (38.93—4), в этом случае $k = t$, $\lambda = b$. Это понятно и из самого определения r и λ. В дальнейшем мы исключим этот случай и будем рассматривать только случаи, когда $r > \lambda$, $t > k$.

Блочный эксперимент, удовлетворяющий уравнению плана (38.95) и условиям (38.93—4), называется планом *сбалансированных неполных блоков* (СНБ). Эти планы были введены Иэйтсом (1936а). Их характерная особенность состоит в том, что каждая из t обработок встречается в r из b блоков, состоящих из k объектов, а любая пара обработок встречается в одном блоке в λ из b блоков. Мы пришли к плану СНБ в **38.46—7**, исходя из требования, чтобы оценки всех разностей параметров обработок имели одну и ту же дисперсию $2\sigma^2/a$, где a определяется в (38.91—2).

Из условий (38.93—4) следует, что любые три из пяти констант t, k, λ, r, b определяют оставшиеся две. Как правило, используются первые три из этих констант.

38.50 Условия (38.93—5) являются необходимыми условиями (которым должны удовлетворять целые числа t, k, λ, r, b) для того, чтобы существовал план СНБ. Однако, вообще говоря, они не являются достаточными, поскольку может не существовать матрицы инциденций $\boldsymbol{n}$, удовлетворяющей уравнению плана (38.95) для $\boldsymbol{nn}'$ (см. упражнение 38.11). Можно привести дальнейшие необходимые условия. Например, из (38.95) следует, что $(t \times t)$-матрица $\boldsymbol{nn}'$ не вырождена, так что $(t \times b)$-матрица $\boldsymbol{n}$ должна иметь ранг t, а значит,

$$b \geqslant t \tag{38.96}$$

(отсюда и из (38.94) следует, что $r \geqslant k$). Этот результат впервые другим способом получил Р. Фишер. В своем исчерпывающем обзоре по данному вопросу Гуерин (1965) приводит и другие, более точные необходимые условия. Ханани (1961) доказал, что условия (38.93—5) являются не только необходимыми, но и достаточными при $k = 3$ или $k = 4$, а также при $k = 5$ в случае $\lambda = 1$, 4 или 20.

Если задан некоторый план СНБ, то существует план с теми же t и k, а λ, умноженным на произвольное целое число m. Этот план получается простым повторением исходного плана m раз; при этом r и b также увеличиваются в m раз. Мы всегда будем считать $m = 1$, т. е. брать минимально возможное λ при любом заданном плане. Однако при заданных t и k может существовать несколько таких «минимальных» λ, соответствующих различным планам СНБ. В приведенной ниже в **38.54** таблице фигурируют такие примеры. Более того, для одних и тех же значений t, k и λ может существовать несколько планов СНБ с существенно разной структурой (не изоморфных). Это интуитивно понятно из того факта, что значения r и λ являются только условиями первого и второго порядков на расположение обработок в блоках. В общем случае они не ограничивают частоты одновременного появления трех, четырех и т. д. обработок.

38.51 Если t и k фиксированы, то остается возможность выбора третьей константы плана. План СНБ всегда можно построить, просто взяв в качестве блоков каждое из $\binom{t}{k}$ сочетаний обработок. Тогда

$$b = \binom{t}{k}, \quad r = \binom{t-1}{k-1}, \quad \lambda = \binom{t-2}{k-2}. \tag{38:97}$$

Такой план СНБ называется *неприводимым*. Поскольку для проведения соответствующего эксперимента требуется

$$bk = rt = t^{(k)}/(k-1)!$$

наблюдений, то неприводимые планы применяются на практике, только когда k или $(t-k)$ малы.

Если $k = 2$, то из (38.93—4) следует, что всегда

$$b = \lambda \binom{t}{2}.$$

Для неприводимого плана (38.97) это выполняется только при $\lambda = 1$.

Если $k = t - 1$, то (38.97) сводится к $b = t$, $r = t - 1$, $\lambda = t - 2$. Это пример *симметричного* плана СНБ ($b = t, k = r$), который в указанном случае оказывается также неприводимым.

38.52 Поскольку в эксперименте каждая из t обработок встречается r раз, а каждый блок фигурирует k раз (см. (38.4)), то соблазнительно поменять в плане ролями блоки и обработки, положив $t' = b$, $b' = t$, $r' = k$ и $k' = r$ для получения *дуального* плана. Однако такой дуальный план сам будет планом СНБ тогда и только тогда, когда исходный план СНБ является симметричным (в качестве иллюстрации см. упражнение 38.12).

Если план СНБ можно разложить на r подмножеств блоков, так что в каждом подмножестве каждая обработка присутствует один раз, то план называется *разложимым* (*resolvable*). При этом в каждом подмножестве фигурирует *дубликат* множества обработок. Для разложимости, очевидно, требуется, чтобы $t = ck$, $b = cr$, где c — целое положительное число. Однако последние условия сами по себе не являются достаточными для разложимости плана.

Как показал Бозе (1942), для разложимых планов СНБ условие (38.96) можно заменить на более сильное

$$b \geqslant t + r - 1. \tag{39.98}$$

Последнее условие снова является только необходимым для существования разложимого плана СНБ, поскольку (38.98) всегда справедливо, если только $t = ck$, $b = cr$ (см. упражнение 38.13). Равенство в (38.98) достигается тогда и только тогда, когда для любых двух блоков из различных дубликатов внутри разложимого плана существует ровно k^2/t общих обработок. В этом случае план называется *афинно разложимым* планом СНБ (см. упражнение 38.18).

38.53 Манн (1949) приводит обзор методов построения планов СНБ, восходящих к Бозе. В работе Гуерина (1965) описываются эти и некоторые другие методы построения планов СНБ. Там же приведена исчерпывающая библиография

Таблица 38.1

Указатель известных планов СНБ для $t \leqslant 100$, $k \leqslant r \leqslant 20$
(неприводимые планы в таблицу не включены)

k \ t	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	25	26	27	28	29	31	33	35	36	37	39	40	41	43	45	49	52	57	61	64	65	66	73	76	81	91	96
3	2*	1*		1*	2*	3†	2†	1*		1*	2†		2‡	1*		1*			1†		1†			1†	1‡			1‡	1‡																
4			3*	3*	2*	6‡	3†	1*			1*	3‡		2†	3‡	3‡	2†		1*			1*		2‡				1†		1†				1‡	1‡		1‡								
5					4‡	2*		5†		4†	4†	5‡			4‡	1*			1*					2†		2‡					1*		1†				1†		1‡				1‡		
6							5†	5†		5†	(2,3)		5‡	5‡		(3,4)				3†				1*			2†													1‡					
7									6‡	3*						3*	4†					2*	3†									2†		1*											1†	
8											7†	7‡											4‡													1*		1*								
9													8‡	4‡		6†			3*		4†				3†			2*														1*		1*		
10															9‡	9‡								3*				5‡																		1*	
11																		5†																													
12																																	3†														
13																					6†									4†																
14																																															
15																								7†			(6,8)																				
16																																											4‡			
17																										8‡																					
18																																															
19																													9‡																		
20																																														4‡	

$k > \frac{1}{2}t$ см. текст

В таблице для любых (t, k) указано соответствующее значение (значения) λ. Эти значения могут быть увеличены в любое целое число раз простым повторением всего плана. Если указано более одного значения λ, то можно брать любую их линейную комбинацию с целыми неотрицательными коэффициентами, добавив нужное число повторений планов, соответствующих каждому из указанных λ.

*) См. таблицы Фишера и Иэйтса; $r \leqslant 10$.
†) См. таблицы С. Рао (1961) и Фишера и Иэйтса (6-е изд., 1963): $11 \leqslant r \leqslant 15$.
‡) См. Такэючи (1962): $16 \leqslant r \leqslant 20$.

по данному вопросу со списком фундаментальной серии работ Бозе, начиная с Бозе (1939). Вайда (1967b) изучает математические аспекты этого вопроса. Маллер (1965) приводит методы получения планов СНБ из полных множеств ортогональных латинских квадратов, когда t — целая степень простого числа (см. упражнения 38.10—11, где даны некоторые примеры такой конструкции, приведенные в более ранних таблицах Фишера и Иэйтса). Если t нечетное, а k не превосходит наименьшего простого делителя t, то план СНБ всегда можно построить методом, приведенным в упражнении 38.20. Если t простое, то этот метод справедлив для любого $k < t$.

38.54 В таблицах Фишера и Иэйтса приведен указатель всех известных планов СНБ для значений k и r с $r \leqslant 10$. Там же приведены комбинаторные методы получения конкретных планов. Кокрэн и Кокс (1957) приводят детальные способы выбора из этих планов. С. Рао (1961) описывает планы и комбинаторные методы их получения для $11 \leqslant r \leqslant 15$ (эти планы, а также их обобщения включены в 6-е издание таблиц Фишера и Иэйтса за 1963 г.). Спротт (1962) описывает планы при $16 \leqslant r \leqslant 20$; Такэючи (1962) при $t \leqslant 100$ и $r \leqslant 20$ ($r = 30$ для симметричных планов) приводит описание и дает методы построения (см. исправления Силлитто (1964) и Клатуэрти и Левицкий (1968)).

В таблице 38.1 приведены при $t \leqslant 100$, $k \leqslant r \leqslant 20$ значения λ, для которых известны планы СНБ, не являющиеся неприводимыми. Случай $k = 2$ опущен, поскольку (согласно **38.51**) в этом случае всегда существует неприводимый план с $\lambda = 1$. При $k = t - 1$ (опять согласно **38.51**) всегда существует симметричный неприводимый план с $\lambda = t - 2$, также не фигурирующий в нашей таблице. Далее в таблице приводятся только значения λ при $k \leqslant t/2$, поскольку (за исключением уже рассмотренного случая $k = t - 1$) план с $k' > t/2$ всегда можно получить из плана с $k < t/2$. В самом деле, рассмотрим для каждого блока те обработки, которые в нем не применяются, в качестве нового, дополнительного блока, содержащего $k' = t - k$ объектов. В результате получится *дополнительный* план, состоящий из дополнительных блоков. Из (38.93—4) получим, что

$$\frac{\lambda'}{\lambda} = \frac{(t-k)(t-k-1)}{k(k-1)} > 1.$$

Анализ планов СНБ

38.55 Анализ эксперимента, проведенного по плану СНБ, можно провести теперь, подставив просто в соотношения из **38.23—6** соответствующие значения, поскольку **38.23—6** справедливы для любых блочных экспериментов.

Подставляя (38.89) и (38.95) в (38.23), получим с использованием (38.93—4) для исключения r и b

$$\mathbf{\Omega}^{-1} = r\boldsymbol{I}_t - \frac{1}{k}\{(r-\lambda)\boldsymbol{I}_t + \lambda \mathbf{1}_t\mathbf{1}_t'\} + \frac{r^2}{bk}\mathbf{1}_t\mathbf{1}_t' = \\ = \frac{\lambda t}{k}\left\{\boldsymbol{I}_t + \frac{(t-k)}{t^2(k-1)}\mathbf{1}_t\mathbf{1}_t'\right\}. \qquad (38.99)$$

Матрица в (38.99) имеет точно тот же вид, что и в (38.87), а значит, как может непосредственно проверить читатель, обратная к ней, как и к (38.87), равна

$$\mathbf{\Omega} = \frac{k}{\lambda t}\left\{\boldsymbol{I}_t - \frac{(t-k)}{kt(t-1)}\mathbf{1}_t\mathbf{1}_t'\right\}. \qquad (38.100)$$

Подставляя (38.100) в (38.29) и используя (38.34), получим

$$\hat{\boldsymbol{\tau}} = \frac{k}{\lambda t}(\boldsymbol{T} - \boldsymbol{nB}/k) + \mathbf{1}_t G/(bk). \qquad (38.101)$$

Отсюда следует, что оценки параметров обработок больше не являются не зависящими от суммарных значений по блокам. Это и должно быть, поскольку различные множества блоков связаны с различными обработками. Из определения $\boldsymbol{n}$ и $\boldsymbol{B}$ видим, что $\boldsymbol{nB}/k$ является $(t \times 1)$-вектором, i-й элемент которого равен сумме средних значений для блоков, взятой по тем блокам, в которых применяется i-я обработка. А значит, представляющий непосредственный интерес вектор

$$\boldsymbol{T}_a = \boldsymbol{T} - \boldsymbol{nB}/k \qquad (38.102)$$

может быть назван вектором *исправленных* сумм для обработок. Соотношение (38.101) принимает вид

$$\hat{\boldsymbol{\tau}} = \frac{k}{\lambda t}\boldsymbol{T}_a + \mathbf{1}_t G/(bk). \qquad (38.103)$$

Используя (38.103) и (38.4), получим, что (38.32) сводится, таким образом, к

$$\hat{\boldsymbol{\beta}} = \boldsymbol{B}/k - \frac{1}{\lambda t}\boldsymbol{n}'\boldsymbol{T}_a - \mathbf{1}_b G/(bk). \qquad (38.104)$$

Далее, СК разницы между обработками в таблице ДА (38.39) равна $\boldsymbol{T}_a'\mathbf{\Omega}\boldsymbol{T}_a$, а с использованием (38.100) и (38.34) это сводится просто к

$$\boldsymbol{T}_a'\mathbf{\Omega}\boldsymbol{T}_a = \frac{k}{\lambda t}\boldsymbol{T}_a'\boldsymbol{T}_a. \qquad (38.105)$$

В результате получаем следующую таблицу ДА, являющуюся частным случаем (38.39).

Таблица ДА для эксперимента СНБ

Источник дисперсии	СК	Число ст. св.
Разница между обработками	$\frac{k}{\lambda t}\boldsymbol{T}_a'\boldsymbol{T}_a$	$t-1$
Эффекты блоков	$\boldsymbol{B}'\boldsymbol{B}/k - G^2/(bk)$	$b-1$
Остаточная	$\boldsymbol{y}'\boldsymbol{y} - \frac{k}{\lambda t}\boldsymbol{T}_a'\boldsymbol{T}_a - \boldsymbol{B}'\boldsymbol{B}/k$	$bk-t-b+1$
Генеральное среднее	$G^2/(bk)$	1
Полная	$\boldsymbol{y}'\boldsymbol{y}$	bk

(38.106)

Читатель может проверить, что (38.106) сводится к таблице случайных блоков (38.52), если положить $k = t$, $\lambda = b$.

Кокрэн и Кокс (1957) и Кемпсорн (1952) приводят подробные инструкции для проведения ДА для планов СНБ, обращая внимание на возможные упрощения в разложимом, симметричном и других частных случаях. В их работах учитывается также использование информации между блоками. Сейчас мы перейдем к рассмотрению этого вопроса (см. также С. Рао (1947)).

38.56 Можно предположить, что результаты **38.55** должны дополнить наше обсуждение анализа планов СНБ, однако это не так. Модель (38.6), которая служит основой для всех наших результатов, является линейной моделью с фиксированными эффектами, т. е., как и в главе 35, мы проводим анализ по методу НК, оставаясь в рамках модели I. При рассмотрении параметров обработок это предположение, вообще говоря, реалистично. Но, как мы видели в **38.14**, блоки в эксперименте являются обычно мешающими факторами, не имеющими прямого интереса. Конкретные блоки, используемые в эксперименте, не оказывают на него влияния. Поэтому при проведении анализа более естественно рассматривать эффекты блоков как случайные величины. В терминологии главы 36 нам следует, таким образом, рассмотреть *смешанную модель* с фиксированными эффектами обработок и случайными эффектами блоков. Это естественно ведет к другому анализу, который обычно называется *использованием информации между блоками*. Проводимый ниже анализ не ограничивается планами СНБ, а справедлив для любых блочных экспериментов.

Смешанная модель для использования информации между блоками

38.57 Не будем рассматривать теперь $b-1$ линейно независимых параметров блоков α из модели в **38.21**. Вместо этого предполагается, что эффектам блоков соответствует случайная величина, скажем β *). Как обычно, предполагается, что ошибки ε_{ij} имеют нулевое математическое ожидание и дисперсию σ^2. Если математическое ожидание β равно нулю, то β не влияет на математическое ожидание y, а дисперсия β налагается на обычную дисперсию ошибки ε_{ij}. В обозначениях **38.16—21** описываемая модель имеет вид

$$\mathsf{M}(y) = X\tau, \tag{38.107}$$

где

$$\underset{(bk\times t)}{X} = \begin{pmatrix} t_1 \\ t_2 \\ \cdot \\ \cdot \\ \cdot \\ t_b \end{pmatrix}. \tag{38.108}$$

Следует заметить, что все еще предполагается, что эффекты блоков и обработок не взаимодействуют.

Теперь ошибки в модели не являются больше некоррелированными, поскольку любым двум наблюдениям из одного блока соответствует одно и то же значение β. Если обозначить $\rho = \sigma_\beta^2/\sigma^2$, то легко видеть, что дисперсионная матрица y равна

$$V = \sigma^2 \begin{pmatrix} A & & & 0 \\ & A & & \\ & & \ddots & \\ 0 & & & A \end{pmatrix}. \tag{38.109}$$

Здесь вдоль главной диагонали матрицы V стоят b одинаковых матриц

$$A = I_k + \rho 1_k 1_k'. \tag{38.110}$$

Предположим сначала, что ρ известно. (Вопрос об оценивании ρ рассматривается в **38.62—4**.)

Для оценивания τ воспользуемся обобщенной НК-оценкой из **19.17** (т. 2):

$$\hat{\tau} = (X'V^{-1}X)^{-1} X'V^{-1}y, \tag{38.111}$$

*) В нарушение обычного соглашения здесь для обозначения случайной величины, соответствующей блокам, используется греческая буква, поскольку b и B уже использовались выше.

дисперсионная матрица которой равна

$$V(\hat{\tau}) = (X'V^{-1}X)^{-1}. \tag{38.112}$$

38.58 Матрица, обратная к (38.110), как и в (38.87), имеет вид

$$A^{-1} = I_k - \left(\frac{\rho}{1+k\rho}\right) 1_k 1_k'. \tag{38.113}$$

Используя (38.108—9) и (38.113), имеем

$$X'V^{-1}X = \frac{1}{\sigma^2}\sum_j t_j' \left\{ I_k - \left(\frac{\rho}{1+k\rho}\right) 1_k 1_k' \right\} t_j.$$

Подставляя сюда (38.5) и (38.1), получим

$$X'V^{-1}X = \frac{1}{\sigma^2}\left\{ \operatorname{diag}(r) - \left(\frac{\rho}{1+k\rho}\right)\sum_j n_j n_j' \right\}. \tag{38.114}$$

Поскольку из определения следует, что

$$\sum_{j=1}^{b} n_j n_j' = nn', \tag{38.115}$$

то (38.114) сводится к

$$X'V^{-1}X = \frac{1}{\sigma^2}\left\{ \operatorname{diag}(r) - \left(\frac{\rho}{1+k\rho}\right) nn' \right\}, \tag{38.116}$$

а дисперсионная матрица оценок в (38.112) является обратной к (38.116). Далее

$$X'V^{-1}y = \frac{1}{\sigma^2}\sum_j t_j' \left\{ I_k - \left(\frac{\rho}{1+k\rho}\right) 1_t 1_t' \right\} y_j,$$

что после подстановки сюда (38.18) и (38.1) сводится к

$$X'V^{-1}y = \frac{1}{\sigma^2}\left\{ T - \left(\frac{\rho}{1+k\rho}\right)\sum_j n_j B_j \right\} = \frac{1}{\sigma^2}\left\{ T - \left(\frac{\rho}{1+k\rho}\right) nB \right\}. \tag{38.117}$$

Используя (38.116—17), перепишем (38.111) в виде

$$\hat{\tau} = \left\{ \operatorname{diag}(r) - \left(\frac{\rho}{1+k\rho}\right) nn' \right\}^{-1} \left\{ T - \left(\frac{\rho}{1+k\rho}\right) nB \right\}. \tag{38.118}$$

38.59 Если $\sigma_\beta^2 \to 0$ $(\rho \to 0)$, то (38.118) сводится к

$$\lim_{\rho\to 0} \hat{\tau} = \{\operatorname{diag}(r)\}^{-1} T,$$

а дисперсионная матрица $\hat{\tau}$ равна тогда $\sigma^2\{\operatorname{diag}(r)\}^{-1}$. Это

в точности совпадает с результатом, который мы бы должны были получить, если бы вовсе не учитывали эффекты блоков (см. ниже (38.123)). Но так и должно быть, ибо при $\sigma_\beta^2 \to 0$ и $\mathbf{M}(\beta) = 0$ эффекты блоков исчезают.

В другом крайнем случае, когда $\sigma_\beta^2 \to \infty$ (и $\rho \to \infty$), соотношение (38.118) нужно переписать так, чтобы в нем не содержалось обращение матрицы, ибо в этом случае $\rho/(1+k\rho) \to 1/k$ и матрица $\{\mathrm{diag}(\boldsymbol{r}) - \boldsymbol{n}\boldsymbol{n}'/k\}$ сингулярна (проверка сингулярности проводится умножением справа на $\mathbf{1}_t$ и использованием (38.3—4)). Перепишем поэтому (38.118) при $\rho \to \infty$ в виде

$$\{\mathrm{diag}(\boldsymbol{r}) - \boldsymbol{n}\boldsymbol{n}'/k\}\,\hat{\boldsymbol{\tau}} = \boldsymbol{T} - \boldsymbol{n}\boldsymbol{B}/k. \qquad (38.119)$$

Умножая (33.24) слева на $\boldsymbol{\Omega}^{-1}$ и используя (38.31), получим, что оценка $\hat{\boldsymbol{\tau}}$ из (38.29) удовлетворяет соотношению (38.119). Таким образом, если эффекты блоков велики, то оценка (38.118) стремится к внутриблочной оценке (38.29). Хотя это может звучать парадоксально, но обращение к смешанной модели существенно влияет на оценку только тогда, когда эффекты блоков малы.

Легко видеть, что внутриблочные оценки остаются несмещенными в смешанной модели. В самом деле, поскольку для любого фиксированного множества эффектов блоков оценки являются несмещенными, то они безусловно должны остаться такими, когда эффекты блоков становятся случайными величинами. Если эффекты блоков велики, то стремятся применять более простой анализ между блоками, поскольку, как мы только что видели, изменение оценок в этом случае мало. Однако дисперсионной матрицей оценок будет теперь не (38.40), а обратная к (38.116).

38.60 Совсем не очевидно, что для случайных блоков оценки, полученные с использованием смешанной модели не меняются по сравнению с внутриблочными оценками, полученными для модели с фиксированными эффектами. В упражнении 38.14 показано, что для случайных блоков оценки (38.118) совпадают с оценками (38.48), но что дисперсионная матрица оценок изменяется так, что их дисперсии возрастают. (Это понятно из того факта, что оценка любого параметра обработки T_i/b является средним из b независимых наблюдений с дисперсией, которая, согласно (38.109—10), равна $\sigma^2(1+\rho)$). Такое изменение происходит из-за того, что в нашей модели отсутствует параметр генерального среднего. Если бы он был, то оценки в обеих моделях имели бы одно и то же распределение (см. упражнение 38.14).

38.61 Первоначальный подход Иэйтса (1939, 1940а) к использованию информации между блоками был несколько отличным. Сначала отмечалось, что, исходя из суммарных значений для блоков, можно построить междублочные оценки параметров обработок. Эти оценки оказывались некоррелированными с внутриблочными оценками. Для получения оценки с минимально допустимой этим методом дисперсией бралась взвешенная сумма двух приведенных оценок. (В упражнении 38.15 читателю предлагается рассмотреть обобщенный вариант такого подхода.) Но очевидно, что эти два разных подхода к использованию информации между блоками эквивалентны. В самом деле, взвешенная сумма двух компонент в упражнении 38.15 определяется со ссылкой на дисперсионную матрицу внутриблочных оценок в исходной модели (с фиксированными эффектами), а дисперсионная матрица междублочных оценок получена для смешанной модели. Тем не менее в упражнении 38.16 показано, что оба метода всегда приводят к одной и той же оценке.

38.62 МД-оценка $\hat{\boldsymbol{\tau}}$ в (38.118) является функцией отношения дисперсий $\rho = \sigma_\beta^2/\sigma^2$. Обычно бывает нужно оценить само ρ с помощью некоторого $\hat{\rho}$, а затем можно использовать $\hat{\boldsymbol{\tau}}(\hat{\rho})$. Оценим сначала отдельно σ_β^2 и σ^2, а затем в качестве оценки для ρ возьмем отношение полученных оценок.

Для нахождения подходящей оценки величин σ_β^2 и σ^2 вернемся к общему анализу из **38.19—24.** Теперь мы хотим найти СК, соответствующую блокам, а не разнице между обработками, как это было в **38.25.** В результате мы получим остаточную СК, скажем S_2, когда отсутствуют эффекты блоков. Разность $S_2 - S_0$ будет тогда равняться СК, соответствующей эффектам блоков.

38.63 Полагая $\hat{\boldsymbol{\beta}} = \mathbf{0}$ в (38.32), получим

$$\boldsymbol{B} = \boldsymbol{n}'\hat{\boldsymbol{\tau}}_0, \qquad (38.120)$$

где $\hat{\boldsymbol{\tau}}_0$ обозначает $(\hat{\boldsymbol{\tau}})_{\beta=0}$. Умножая слева на $\mathbf{1}_b'$ и используя (38.25) и (38.3), имеем

$$G = \boldsymbol{r}'\hat{\boldsymbol{\tau}}_0. \qquad (38.121)$$

Если подставить теперь (38.120—1) в (38.24), то с использованием (38.23) получим

$$\hat{\boldsymbol{\tau}}_0 = \boldsymbol{\Omega}\{\boldsymbol{T} - \boldsymbol{n}\boldsymbol{n}'\hat{\boldsymbol{\tau}}_0/k + \boldsymbol{r}\boldsymbol{r}'\hat{\boldsymbol{\tau}}_0/(bk)\} =$$
$$= \boldsymbol{\Omega}\{\boldsymbol{T} + [\boldsymbol{\Omega}^{-1} - \mathrm{diag}(\boldsymbol{r})]\,\hat{\boldsymbol{\tau}}_0\}. \qquad (38.122)$$

Разрешая (38.122) относительно $\hat{\boldsymbol{\tau}}_0$, получим

$$(\hat{\boldsymbol{\tau}})_{\beta=0} = \{\mathrm{diag}(\boldsymbol{r})\}^{-1}\boldsymbol{T}. \qquad (38.123)$$

В результате (38.33) сводится к

$$S_2 = y'y - T'\{\mathrm{diag}(r)\}^{-1}T. \qquad (38.124)$$

Используя первую строчку таблицы ДА (38.39) и (38.37), находим

$$S_2 - S_0 = T'\hat{\tau} + B'\hat{\beta} - T'\{\mathrm{diag}(r)\}^{-1}T =$$
$$= (T - nB/k)'\,\Omega\,(T - nB/k) + B'B/k - T'\{\mathrm{diag}(r)\}^{-1}T. \qquad (38.125)$$

В результате получилась искомая СК, соответствующая блокам. Таким образом, альтернативная к (38.39) таблица ДА имеет такой вид:

Источник дисперсии	СК	Число ст. св.
Эффекты блоков (с учетом разницы между обработками)	$(T - nB/k)'\,\Omega\,(T - nB/k) + B'B/k - T'\{\mathrm{diag}(r)\}^{-1}T$	$b-1$
Разница между обработками (без учета эффекта блоков)	$T'\{\mathrm{diag}(r)\}^{-1}T - G^2/(bk)$	$t-1$
Остаточная	$y'y - T'\hat{\tau} - B'\hat{\beta}$	$bk-b-t+1$
Генеральное среднее	$G^2/(bk)$	1
Полная	$y'y$	bk

(38.126)

Как и должно было быть, строки, соответствующие остаточной СК и генеральному среднему, не меняются, а оставшаяся СК перераспределяется между эффектами блоков и обработок поиному, поскольку здесь сначала принимаются во внимание разница между обработками, а не эффекты блоков. Следует вспомнить из примера 35.4, что порядок не имеет значения только в случае ортогонального анализа.

38.64 Как обычно, σ^2 можно оценить теперь остаточной СК S_0, поделенной на ее число ст. св., поскольку

$$\mathsf{M}(S_0) = (bk - b - t + 1)\,\sigma^2. \qquad (38.127)$$

Для оценивания σ_β^2 заметим сначала, что, согласно (38.124), СК, обусловленная блоками (скажем S_B), может быть переписана в виде

$$S_2 - S_0 = S_B = y'y - S_0 - T'\{\mathrm{diag}(r)\}^{-1}T,$$

так что, используя (38.127), получим

$$\mathsf{M}(S_B) =$$
$$= \mathsf{M}(y'y) - (bk - b - t + 1)\,\sigma^2 - \mathsf{M}[T'\{\mathrm{diag}(r)\}^{-1}T]. \qquad (38.128)$$

Из результатов упражнения 19.3 мы знаем, что если $\mathsf{M}(z) = 0$ и $V(z) = \sigma^2 W$, то $\mathsf{M}(z'Az) = \sigma^2\,\mathrm{tr}(AW)$, а значит, полагая без ограничения общности $\tau = 0$, получим из (38.109—10)

$$\mathsf{M}(y'y) = \sigma^2 bk\,(1 + \rho). \qquad (38.129)$$

Из модели (38.107—10) следует, с учетом свойств nn' из 38.48

$$\mathsf{M}(T) = \{\mathrm{diag}(r)\}\,\tau, \quad V(T) = \sigma^2[\mathrm{diag}(r) + \rho nn'].$$

Теперь снова можно воспользоваться результатом упражнения 19.3 и получить

$$\mathsf{M}[T'\{\mathrm{diag}(r)\}^{-1}T] = \sigma^2\,\mathrm{tr}[\{\mathrm{diag}(r)\}^{-1}\{\mathrm{diag}(r) + \rho nn'\}] =$$
$$= \sigma^2\,\mathrm{tr}[I_t + \rho\{\mathrm{diag}(r)\}^{-1}nn'] =$$
$$= \sigma^2\{t + \rho\,\mathrm{tr}[\{\mathrm{diag}(r)\}^{-1}nn']\}.$$

Используя определения, легко проверить, что

$$\mathrm{tr}[\{\mathrm{diag}(r)\}^{-1}nn'] = t, \qquad (38.130)$$

а значит,

$$\mathsf{M}[T'\{\mathrm{diag}(r)\}^{-1}T] = \sigma^2 t\,(1 + \rho), \qquad (38.131)$$

и (38.129) и (38.131) превращают (38.128) в

$$\mathsf{M}(S_B) = (b - 1)\,\sigma^2 + (bk - t)\,\rho\sigma^2. \qquad (38.132)$$

В результате из (38.132) и (38.127) получим

$$\mathsf{M}\left\{\frac{S_B - (b-1)\,S_0/(bk - b - t + 1)}{bk - t}\right\} = \rho\sigma^2 \equiv \sigma_\beta^2. \qquad (38.133)$$

Соотношения (38.133) и (38.127) задают искомые оценки.

Пусть $\hat{\rho}$ — их отношение. В рассматриваемом контексте мы интересуемся оцениванием τ, и $\hat{\rho}$ не обладает никакими свойствами оптимальности. Иэйтс (1939, 1940а) проводил усечение оценки σ_β^2, заменяя отрицательные значения нулем. В результате получалась оценка $\bar{\rho}$. Точер (1952) дает другие, более сложные оценки значения σ_β^2, которые являются несмещенными квадратичными МД-оценками при условии, что ошибки обладают нулевыми эксцессом и асимметрией. На самом деле нам нужна такая оценка значения ρ, чтобы при ее подстановке в оценку параметров обработок (38.118) сохранялась несмещенность.

Грейбилл и Уикс (1959) и Грейбилл и Сэшадри (1960) показали, что для планов СНБ оценка $\hat{\tau}$ из (38.118) остается несмещенной при использовании $\hat{\rho}$ или $\bar{\rho}$. Сэшадри (1966) показал, что дисперсия меньше, если использовать $\bar{\rho}$. Дж. Рой и Шах (1962) показали, что если оценка параметра ρ принадлежит к некоторому типу, определяемому в терминах характеристических чисел матрицы nn', то несмещенность $\hat{\tau}$ сохраняется для любого плана неполных блоков такого, что $r = r\mathbf{1}_t$, причем $\hat{\rho}$ принадлежит к требуемому типу. Шах (1964) приводит другие несмещенные оценки для нескольких хорошо известных планов экспериментов, включая план СНБ с $t > 5$.

Стейн (1966) рассматривает использование информации между блоками для планов СНБ, когда эффекты блоков возникают только из-за рандомизации.

Перестановочные распределения для планов СНБ

38.65 После общего обсуждения смешанной модели в **38.57—64** вернемся к модели с фиксированными эффектами для экспериментов СНБ и рассмотрим критерии перестановок для эффектов обработок.

Огава (1963) показал, что если имеются ошибки объектов (см. **36.41**), то стандартный F-критерий для обработок может быть оправдан, как приближение для перестановочного распределения, если b достаточно велико, а дисперсии ошибок объектов внутри блоков почти постоянны. *А значит*, его применение оправдано, когда b велико, а ошибки объектов отсутствуют.

Если вместо наблюденных значений использовать внутри блоков ранги, то для эффектов обработок в плане СНБ можно получить обобщение перестановочного распределения статистики критерия, рассмотренного в **38.43** и **37.39—41** для случайных блоков. Результаты, восходящие к Дербину (1951), приведены в упражнении 38.17. Ван Элтерн и Нётер (1959) показали, что по сравнению с обычным F-критерием для эффектов обработок, для критерия Дербина, использующего ранги, АОЭ равна $k/(k+1)$, умноженному на АОЭ критерия Вилкоксона (31.115). В нормальном случае получается $3k/\{\pi(k+1)\}$. Оказывается, что АОЭ зависит от объема блоков, а не от t. Для случайных блоков см. также П. Сен (1967).

Интересно отметить, что точные выражения можно в рассматриваемом случае получить, вообще говоря, только для первых двух моментов статистики критерия. Причина состоит в том, что, как упоминалось в **38.50**, условия СНБ не накладывают никаких ограничений на совместное появление трех и большего числа обработок.

Бенард и Ван Элтери (1953) приводят «хи»-квадрат-критерий перестановок для больших выборок в случае произвольных (не обязательно сбалансированных) планов неполных блоков, использующих ранги и допускающих повторение и потерю наблюденных значений.

Эксперименты с предпочтением

38.66 Планы СНБ представляют интерес в связи с экспериментами с предпочтением (preferenc experiments), когда часто невозможно измерить степень предпочтения, но известно ранжирование предпочтений. Если предпочтения должны быть выражены внутри b блоков, k элементов (обработок) которых выбраны из возможного числа t, то порядок, в котором рассматриваются эти элементы, может оказаться важным. Поэтому желательно, чтобы план СНБ принимал во внимание это влияние порядка. Простейший способ состоит в рассмотрении элементов в порядке, определяемом столбцами $(t \times t)$ латинского квадрата. Если использовать первые k столбцов квадрата, то они определяют t блоков по k элементов, причем каждый из возможных t элементов появляется ровно один раз в каждом соответствующем порядку положении. Если $b = ct$, где c — положительное целое число, то можно получить полное сбалансирование по порядку, используя указанным способом первые k столбцов c латинских квадратов размера $(t \times t)$.

Так, например, первые три столбца из (38.71) задают план СНБ с $t = 4$, $b = 4$, $k = 3$, $\lambda = 2$. Каждая из букв A, B, C, D появляется один раз в каждом столбце.

Используемые таким способом неполные латинские квадраты известны как планы квадратов Юдэна после того, как их открыл В. Юдэн. Разумеется сбалансированность по расположению может оказаться важной для любого эксперимента СНБ (не обязательно с предпочтениями), в котором «положение» может относиться к любой переменной, мешающий эффект которой мы хотим исключить из эксперимента. Так было в исходном рассмотрении латинских квадратов (см. **38.31** и **38.34**).

Парные сравнения

38.67 Частный случай плана СНБ при $k = 2$ (когда, как мы видели в **38.51**, план неприводим) обычно называется планом *парных сравнений* (*paired comparisons design*). Этот план особенно важен в экспериментах с предпочтением. В монографии Дэвида (1963), посвященной методам парных сравнений, имеется глава о соответствующих планах экспериментов. Возможно, самыми важными из них являются *планы связанных* (*linked*) *парных сравнений*, развитые Бозе (1956). В этих планах каждое из t суждений состоит в сравнении r пар элементов

(выбранных из полного числа n). Сравнение для каждой пары содержится в k суждениях, причем для любых двух суждений имеется ровно λ общих пар. Как указывают обозначения, это соответствует планам СНБ: суждения соответствуют «обработкам», пары элементов — блокам, «содержащим» k таких обработок. Всего имеется $b = \frac{1}{2} n(n-1)$ таких блоков. Новая черта планов связанных парных сравнений состоит в том, что требуется, чтобы каждый из n элементов фигурировал одинаково часто в r парах каждого суждения, т. е. фигурировал $2r/n = \alpha$ раз. Поэтому из (38.94) имеем

$$\alpha = k(n-1)/t. \tag{38.134}$$

Из-за дополнительного условия (38.134) из существования плана СНБ не следует существование плана связанных парных сравнений, хотя из существования последнего, очевидно, следует существование плана СНБ. Бозе (1956) дает методы (воспроизведенные Дэвидом (1963)) вывода планов связанных парных сравнений из планов СНБ.

Частично сбалансированные неполные блоки

38.68 Существенной чертой планов СНБ (см. **38.49**) является полная симметрия обработок, которые появляются всего r раз каждая и λ раз совместно с любой другой. Эта симметрия — естественное следствие симметричного требования в **38.46**—7 об одной и той же точности для всех оценок разностей между обработками. Сохранив условие появления каждой обработки ровно r раз, отбросим теперь требование того, чтобы λ было константой.

Предположим, что для каждой обработки оставшиеся $t-1$ обработок распадаются на m классов объема t_p, $\sum_{p=1}^{m} t_p = t-1$. Эти классы называются *ассоциированными классами* (*associate classes*), а любая обработка из p-го ассоциированного класса называется *p-ассоциированной* для данной обработки. Наложим теперь следующие условия:

(а) для двух обработок, p-ассоциированных с данной, существует ровно λ_p блоков, в которых они появляются совместно;

(б) если A — p-ассоциирована с B, то B — p-ассоциирована с A;

(в) если A и B — l-ассоциированы, то число обработок, которые одновременно p-ассоциированы для обработки A и q-ассоциированы для обработки B, постоянно (мы обозначим его P^l_{pq}).

План, удовлетворяющий этим условиям, называется планом *частично сбалансированных неполных блоков* (ЧСНБ). Такие планы впервые рассматривали Бозе и Нэйр (1939). Очевидно, что они содержат как частный случай планы СНБ; для этого нужно, чтобы $\lambda_p = \lambda$ для всех p.

38.69 Планы ЧСНБ содержат много констант. Это, как и ранее, t, r, b, k; m значений λ_p; m значений t_p и значения P^l_{pq}. Однако между этими константами существуют линейные соотношения, которые уменьшают их эффективное число.

Случай двух ассоциированных классов $m = 2$ подробно изучался Бозе, Клатуэрти, Шриканде (1954), которые приводят все известные планы для $r \leqslant 10$, $3 \leqslant k \leqslant 10$. Даже в этом простейшем случае имеются пять типов *схем ассоциирования* (т. е. схем, устанавливающих ассоциативные соотношения между обработками). Самым важным типом являются *делимые на группы* планы ЧСНБ, которые сами состоят из трех подтипов.

Гуерин (1965) дает обширный обзор основных результатов о существовании и конструировании планов ЧСНБ и приводит исчерпывающую библиографию. См. также математическое изложение Вайда (1967). Подробности соответствующих методов статистического анализа, включая использование информации между блоками, приводят Бозе, Клатуэрти, Шриканде (1954), Кемпсорн (1952) и С. Рао (1947).

Структурные обработки: решетчатые планы

38.70 При изложении планов экспериментов мы пока не предполагали, что между обработками могут существовать какие-либо соотношения. Предположим теперь, что обработки в эксперименте классифицированы в некоторые категории. Это имеет место, например, если обработки соответствуют rc комбинациям в двухфакторной перекрестной классификации. При этом сами обработки можно естественно расположить в виде таблицы с двумя входами. Блочные эксперименты, принимающие во внимание такие классификации, называются *решетчатыми планами* (*lattice designs*). Впервые они были введены Иэйтсом (1936b, 1939, 1940b). Эти планы имеют особенно большое значение, когда t велико, ибо таблица в **38.54** показывает, как мало имеется в этом случае расположений СНБ.

38.71 Предположим, что t обработок представлены в виде $(l \times k)$-таблицы с двумя входами, так что $t = lk$. Можно рассмотреть блоки, соответствующие l строчкам с k обработками и аналогично k столбцам с l обработками. В результате получится план с l блоками по k объектов и k блоками по l объектов. Такой план называется *прямоугольной решеткой* (*rectangular lattice*). Название возникло из-за того, что обработки

можно представить в виде точек решетки в $(l \times k)$-таблице. Чтобы вернуться в рамки нашего обсуждения, имеющего дело только с блоками равного объема, нужно ограничиться *квадратной решеткой*, когда $k = l$. При двухкратном повторении обработок, как это было выше, иногда говорят о *простой решетке*; при троекратном — о *тройной решетке* (*triple lattice*), при четырехкратном — о *четверной решетке* (*quadruple lattice*) и т. д.

Квадратные решетки не являются сбалансированными планами в том смысле, в каком таковыми являются СНБ. Хотя каждая обработка появляется $r = 2$ раза, неверно, что число блоков, в которых она появляется совместно с любой другой, одно и то же. Очевидно, что частота совместного появления λ_{ii} равна единице для обработок из одной и той же строки или из одного и того же столбца $(k \times k)$-таблицы и равна нулю для всех остальных пар обработок.

Можно очевидным образом обобщить двумерные $(k \times k)$-таблицы обработок на p-мерные таблицы, содержащие k^p обработок. Тогда без повторения обработок мы получим k^{p-1} блоков по k объектов. Такие *p-мерные решетчатые* планы снова не сбалансированы. Для некоторых приложений важна *кубическая решетка* $(p = 3)$.

38.72 Чтобы избежать усложнений анализа, связанных с отсутствием сбалансированности, можно построить *сбалансированные решетчатые* планы, рассматривая несколько экземпляров простых решетчатых расположений, которые изучались выше. При $t = k^2 = 9$ таблице

$$\begin{matrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{matrix} \qquad (38.135)$$

соответствует описанный в **38.71** план простой квадратной решетки с шестью блоками

$$\begin{array}{c|c|c||c|c|c} 1 & 4 & 7 & 1 & 2 & 3 \\ 2 & 5 & 8 & 4 & 5 & 6 \\ 3 & 6 & 9 & 7 & 8 & 9 \end{array} \qquad (38.136)$$

причем разбиение в (38.136) соответствует двум экземплярам, содержащим полный набор обработок. Если теперь добавить два других экземпляра

$$\begin{array}{c|c|c||c|c|c} 1 & 2 & 3 & 1 & 2 & 3 \\ 5 & 6 & 4 & 6 & 4 & 5 \\ 9 & 7 & 8 & 8 & 9 & 7 \end{array} \qquad (38.137)$$

то, как может проверить читатель, *объединение* (38.136—7) является сбалансированным. На самом деле в результате получается план СНБ с $t = 9$, $k = 3$, $\lambda = 1$, $b = 12$, $r = 4$.

Четыре полных экземпляра в (38.136—7) можно использовать для образования *множества решетчатых квадратов*, выводя эти экземпляры, как это описано в **38.71**, из (38.135) и следующей таблицы:

$$\begin{matrix} 1 & 5 & 9 \\ 6 & 7 & 2 \\ 8 & 3 & 4 \end{matrix} \qquad (38.138)$$

Эти планы более полезны, когда $k = t^{1/2}$ нечетно. В этом случае, как и в нашем примере, требуется $(k+1)/2$ квадратов. Если k четно, то для построения множества нужно $k+1$ квадратов.

38.73 Более подробная теория и анализ решетчатых планов, изложить которые здесь не позволяет объем книги, приведены в монографиях Кемпсорна (1952) и Кокрэна и Кокса (1957). Для нашего изложения важно, что они приводят нас к рассмотрению «структуризованных» обработок, как минимум с точностью до расположения по категориям. Следуя по этому пути, мы обратим внимание на обработки, которые являются комбинациями интересующих нас элементов, и придем к совсем новым вопросам.

Факторные эксперименты

38.74 Если обработки в эксперименте состоят из всех возможных комбинаций множества интересующих нас факторов, то он называется *факторным экспериментом*. Такой эксперимент формально совпадает с (полной) перекрестной классификацией, детально изученной в **35.15—33** и **35.40—4**. В предшествующих главах мы не использовали терминологию экспериментов, поскольку изложенный там анализ применим также и к неэкспериментальным ситуациям. Каждая переменная, определяющая классификацию (т. е. переменная, соответствующая строчкам, столбцам и т. д.), называется в эксперименте *фактором*. Значения, которые может принимать фактор, образующие клетки маргинальной классификации этого фактора, называются *уровнями* (*level*) этого фактора. В этом контексте изученную ранее $(r \times c)$-перекрестную классификацию можно описать как факторный эксперимент с двумя факторами, у одного из которых r уровней, а у другого c; $(r \times c \times l)$-перекрестная классификация — это факторный эксперимент с тремя факторами, по r, c и l уровней соответственно. Приведенные выше два примера часто называют просто $(r \times c)$-факторным и $(r \times c \times l)$-факторным экспериментами соответственно. Если

в первом из них $r = c$, то говорят о (r^2)-факторном эксперименте и т. д.

38.75 Если по плану случайных блоков проводится факторный эксперимент, то естественно возникает желание разделить СК обработок (с $t-1$ ст. св.) на компоненты, соответствующие главным эффектам и взаимодействиям. Так, например, в $(r \times c)$-факторном эксперименте с одним наблюдением в клетке должно быть $t = rc$ и СК, обусловленная разницей обработок (имеющая $rc-1$ ст. св.) должна быть, как в примере 35.3, разложена на члены с $r-1$, $c-1$, и $(r-1)(c-1)$ ст. св. В общем случае СК, обусловленная разницей между обработками, должна быть разбита на компоненты, соответствующие всем главным эффектам и взаимодействиям различных порядков, как и в главе 35. Это разбиение мы назовем ДА *для обработок*.

Из (38.52) видно, что СК, обусловленная разницей между обработками, равна $\{\boldsymbol{T}'\boldsymbol{T} - G^2/t\}/b$, так что $\boldsymbol{T}'\boldsymbol{T}/b$ играет теперь ту же самую роль для факторного эксперимента по плану случайных блоков, что и $\boldsymbol{y}'\boldsymbol{y}$ в главе 35. Таким образом, ДА для обработок можно проводить теми же методами, что и в главе 35, оперируя суммами по обработкам T_i и заменяя t на n. Нужно только помнить, что все компоненты СК нужно делить на b. Разумеется, это деление обусловлено тем, что каждое T_i представляет собой сумму b наблюдений.

38.76 Из упражнения 38.6 видно, что тот же самый простой ДА для обработок в терминах T_i можно проводить для факторных экспериментов по плану латинского квадрата, причем теперь вместо b делителем будет t, которое снова есть число наблюдений, дающих в сумме T_i. Аналогичное правило справедливо для обобщений в упражнении 38.6 на греко-латинский квадрат и ортогональные квадраты более высоких порядков.

Смешивание

38.77 До сих пор разбиение СК разницы между обработками было само по себе простым, поскольку в **38.75—6** мы рассматривали только факторные эксперименты (сами простой структуры) для простейших блочных планов. Откажемся теперь от обоих ограничений и вернемся к рассмотрению экспериментов с множеством соответствующих обработок для общего блочного плана.

Вглянув на СК разницы между обработками из таблицы ДА (38.39) для общего блочного эксперимента, читатель убедится, что теперь нельзя ожидать простоты **38.75—6**, поскольку размещение обработок между блоками существенно влияет на СК разницы между обработками (это влияние описывается

матрицей $\boldsymbol{B}$). Приведенные соображения остаются оправданными и для планов СНБ (таблица (38.106)), что следует из определения $\boldsymbol{T}_a$ в (38.102).

Нетрудно понять, что это неизбежно, поскольку в каждом блоке появляются не все обработки и влияние обработок оказывается невозможным отделить от влияния самих блоков. Даже сбалансированности (в смысле СНБ) недостаточно, чтобы предотвратить это. На самом деле возникает новая проблема, поскольку, как очевидно, существует опасность, что окажется невозможным оценить важные компоненты СК разницы между обработками (например, главные эффекты и взаимодействия первого порядка), поскольку они чрезвычайно тесно переплетены с эффектами блоков. В этом случае говорят, что они *смешаны* (*confounded*) с блоками.

38.78 Не ограничиваясь СК из таблицы ДА, рассмотрим в самом общем виде оценки p линейных функций от параметров обработок, скажем, $\boldsymbol{C}\boldsymbol{\tau}$, где $\boldsymbol{C}$ представляет собой $(p \times t)$-матрицу заданных коэффициентов. В частности, мы интересуемся *сравнениями* параметров. В этом случае (см. **35.58**) для любой строки матрицы $\boldsymbol{C}$ сумма элементов равна нулю. Следует помнить, что сравнения включают в себя просто разности, а также взаимодействия между любыми параметрами обработок.

Из сформулированной в (38.12—14) модели блочных экспериментов следует, что

$$\boldsymbol{C}\boldsymbol{\tau} \equiv (\boldsymbol{C} \vdots \boldsymbol{0})\,\boldsymbol{\theta}, \tag{38.139}$$

где $\boldsymbol{0}$ представляет собой $(p \times (b-1))$-матрицу, все элементы которой равны нулю. Эта матрица служит для того, чтобы устранить параметры блоков $\boldsymbol{\alpha}$.

Далее, чтобы при оценивании (38.139) вектор $\boldsymbol{L}\boldsymbol{y}$ оказался несмещенным, согласно (19.19) необходимо и достаточно, чтобы

$$\boldsymbol{L}\boldsymbol{X} = (\boldsymbol{C} \vdots \boldsymbol{0}),$$

т. е., используя (38.13),

$$\boldsymbol{L}\begin{bmatrix} t_1 \\ \cdot \\ \cdot \\ \cdot \\ t_b \end{bmatrix} = \boldsymbol{C},$$

$$\boldsymbol{L}\begin{pmatrix} 1_k u_1' \\ \cdot \\ \cdot \\ \cdot \\ 1_k u_b' \end{pmatrix} = \boldsymbol{0}. \tag{38.140}$$

Таким образом, если мы сумеем найти $(p \times bk)$-матрицу $\boldsymbol{L}$, удовлетворяющую (38.140), то между p линейными функциями $\boldsymbol{C}\boldsymbol{\tau}$ не будет смешивания.

38.79 Уравнения (38.140) налагают $p(t+b-1)$ условий на pbk элементов матрицы $\boldsymbol{L}$. При любом p этим условиям заведомо можно удовлетворить, если $t+b-1 \leqslant bk$, т. е. если $b \geqslant (t-1)/(k-1)$. При некоторых значениях p уравнениям *можно* удовлетворить и при $b < (t-1)/(k-1)$, поскольку условия на $\boldsymbol{L}$ не обязательно независимы. Это зависит от структуры блочного плана. Случай $t = bk$ (когда в эксперименте используется один экземпляр множества обработок) попадает в эту категорию, поскольку $t/k < (t-1)/(k-1)$ при $t > k$, а случай $t = k$ в этой ситуации тривиален.

Таким образом, мы видим, что если имеется достаточно блоков, то можно всегда, если этого пожелаем, избежать перемешивания для любого множества линейных функций от параметров обработок. Это интуитивно понятно из того факта, что если некоторые функции перемешаны для данного множества блоков, то можно специально добавить еще множество блоков, в котором они не перемешаны. Это справедливо, в частности, тогда, когда каждое множество блоков состоит из полного набора обработок. В литературе те функции, которые перемешаны в части эксперимента, а не во всем эксперименте, называется *частично смешанными*.

38.80 До сих пор мы смотрели на смешивание как на зло, от которого нужно освободиться. Несомненно, оно мешает нам, делая невозможным оценить некоторые функции от параметров обработок. Даже в случае частичного смешивания происходит существенное усложнение вычислений, которые могут оказаться весьма утомительными, в то время как, разумеется, точность оценок частично смешанных функций должна уменьшиться.

Однако в факторных экспериментах смешивание играет также положительную роль. В **35.44** указывалось, что взаимодействия более высоких порядков обычно в практических ситуациях играют малую роль. Поэтому часто при проведении эксперимента их можно специально перемешать с блоками. В результате соответствующая им СК (и ст. св.) будет включаться в остаточную СК. Разумеется, можно провести точно такой же процесс объединения, что и для несмешанного анализа в **35.44**. Главным здесь является то, что в рамках данного эксперимента (t, k и b) может оказаться невозможным оценить все желаемые линейные функции от параметров обработок (главные эффекты и взаимодействия). Если некоторые из них должны быть смешаны, то, вообще говоря, выгодно начинать с взаимодействий более высоких порядков и перемешивать как можно меньше главных эффектов и взаимодействий первого порядка.

Преследуя эту цель, Фишер (1942) рассмотрел факторный эксперимент с $2^m - 1$ факторами, каждый из которых имеет два уровня (всего 2^{2^m-1} комбинаций). Используя теорию абелевых групп, он доказал, что если $k \geqslant 2^m$, т. е. число блоков больше, чем число факторов, то при наличии одного набора обработок не нужно смешивать главные эффекты и взаимодействия первого порядка. Позднее он распространил результат (Фишер (1945)) на факторы с p^r уровнями, где p — простое число (см. также Манн (1949)). Кемпсорн (1952) дает детальное изложение предмета рассматривая также факторы с разным числом уровней. Кокрэн и Кокс (1957) обсуждают применения с подробными примерами расположений смешанных блоков. Иэйтс (1937) приводит много примеров с применениями в сельскохозяйственном экспериментировании, а Дэвьес, Бокс, Коннор, Кузин, Химсворс и Силлитто (1954) дают примеры промышленных экспериментов.

38.81 Одним из важных применений смешивания является факторный эксперимент с частичным дублированием (fractionally replicated), когда некоторые взаимодействия предполагаются равными нулю, чтобы можно было оценить оставшиеся главные эффекты и взаимодействия, используя только некоторые блоки плана с перемешанными блоками. Предположения, по существу, позволяют проводить анализ различия эффектов, которые иным способом неразличимы. Эти эффекты называются *безразличными* (*aliases*). Кемпсорн (1952) излагает теорию, а у Кокрэна и Кокса (1957) приведены некоторые обсуждения. Дэвьес с соавторами (1954) рассматривают планы с частичным дублированием в применении к промышленным экспериментам. Эти планы часто бывают полезны, когда нужно исследовать большое число факторов.

38.82 В настоящее время в распоряжении экспериментаторов имеется много других факторных планов с перемешиванием. *Расщепленные* планы (*split-plot designs*) смешивают главные эффекты, ставя в соответствие каждый уровень фактора объектам в блоке. Это расположение иногда удобно и даже необходимо для практического проведения эксперимента. Другие более сложные формы смешивания в факторных экспериментах (квазилатинские квадраты, «клеточные» квадраты) описываются и анализируются в уже упомянутых книгах.

Последовательность экспериментов: эволюционная операция

38.83 В книге Кенуя (1953а) особое внимание уделяется задаче планирования и анализа долгосрочных экспериментов и групп экспериментов (см. также Кокрэн и Кокс (1957) и Кэмпсорн (1952)).

Другие, но близкие вопросы, которые начали изучать недавно, состоят в выборе последовательности экспериментов для нахождения *оптимальной комбинации* факторов, например для максимизации выпуска конечного продукта (или некоторой другой величины) у процесса с фиксированными затратами, или, что эквивалентно, минимизации затрат при фиксированном выпуске. Метод состоит в подгонке *поверхности отклика* (*response surface*) к экспериментальным точкам по методу НК. Затем область экспериментирования сдвигается по направлению наискорейшего спуска до ближайшей стационарной точки, которая изучается с точки зрения исследования характера поверхности отклика. Для нахождения поверхности отклика предложен план эксперимента, обладающий специальными свойствами симметрии, названный *ротатабельным планом*. Читатель, интересующийся теорией этого метода эволюционного планирования и связанных с ним ротатабельных планов, отсылается к статьям Бокса и Уилсона (1951), Бокса и Хантера (1957), Бозе и Картера (1959), Гардинера, Грэндейджа и Хейдера (1959), Бозе и Дрэйпера (1959), Дрэйпера (1960a, b, c; 1961) и Бокса и Бенкена (1960). Менее теоретическое изложение приведено в последней главе Дэвьеса с соавторами (1954) и у Бокса (1957)*).

Методы эволюционного планирования не являются последовательными методами в собственном смысле этого слова. Их критиковали по той причине, что они не имеют строгого теоретико-вероятностного обоснования. Однако практическая важность этих методов не подвергается сомнению.

Планирование регрессии

38.84 Мы закончим главу разбором планов экспериментов, целью которых является проведение регрессионного анализа.

Как упоминалось в **38.2**, в примере 28.4 рассматривалась задача планирования эксперимента для изучения простой линейной регрессии. Оказалось, что для минимизации дисперсий НК-оценок двух параметров регрессии нужно половину наблюдений проводить при максимально возможном по модулю отрицательном значении x, а вторую половину при совпадающем по модулю, но положительном значении x. Мы указывали там,

*) См. также В. В. Налимов, Н. А. Чернова, Статистические методы планирования экстремальных экспериментов, «Наука», 1965; Планирование эксперимента, сборник под ред. Г. К. Круга, «Наука», 1966; Ч. Хикс, Основные принципы планирования эксперимента, «Мир», 1967; В. В. Федоров, Теория оптимального эксперимента, «Наука», 1971; Планирование оптимального эксперимента, Сборник под ред. М. Б. Малютова, вып. 48, МГУ, 1975. (*Прим. ред.*).

что это соответствует тому факту, что отрезок прямой наиболее эффективно определяется своими конечными точками.

Рассмотрим теперь более общую задачу размещения значений регрессора x, когда математическое ожидание y является полиномом от x. В **28.16—20** рассматривалась теория полиномиальной регрессии; при этом значения x считались заданными заранее. Теперь мы интересуемся тем, как выбрать значения x, чтобы параметры уравнения полиномиальной регресии оценивались оптимально.

38.85 Предположим, что значения x должны быть размещены внутри фиксированного интервала, который без ограничения общности будем считать интервалом $(-1, +1)$. На полиномиальный случай обобщаются приведенные выше интуитивные аргументы, ибо полином степени k определяется значениями в $k+1$ точке. Более того, в полиномиальном случае все еще можно ожидать, что по одной точке должно быть в каждом конце интервала. Эти интуитивные соображения оказываются справедливыми асимптотически для общей полиномиальной регрессии. Кифер (1959) показал, что если не принимать во внимание аналитические сложности, возникающие из-за того, что n целое (они не возникают при $n \to \infty$), то требуется не более $k+1$ различных значений x, из которых не более $k-1$ значений должны находиться внутри интервала. Характерно, однако, что точная теория, принимающая во внимание упомянутые сложности, не так проста. Нельзя ожидать, что здесь можно опереться на приведенные выше интуитивные аргументы.

38.86 Рассмотрим теперь выбор $k+1$ различных точек наблюдений $x_1 < x_2 < \ldots < x_{k+1}$ для минимизации обобщенной дисперсии оценок параметров. В рассматриваемом полиномиальном случае имеем

$$y = X\theta + \varepsilon,$$

где матрица X имеет вид

$$\underset{(n\times(k+1))}{X} = \begin{pmatrix} 1_{n_1} & 1_{n_1}x_1 & 1_{n_1}x_1^2 & \ldots & 1_{n_1}x_1^k \\ 1_{n_2} & 1_{n_2}x_2 & 1_{n_2}x_2^2 & \ldots & 1_{n_2}x_2^k \\ \vdots & \vdots & \vdots & & \vdots \\ 1_{n_{k+1}} & 1_{n_{k+1}}x_{k+1} & 1_{n_{k+1}}x_{k+1}^2 & \ldots & 1_{n_{k+1}}x_{k+1}^k \end{pmatrix}. \quad (38.141)$$

Здесь в точке x_i $(i = 1, \ldots, k+1)$ проводится n_i наблюдений, причем $\sum_{i=1}^{k+1} n_i = n$. Дисперсионная матрица $\hat{\theta}$ равна

$$V(\hat{\theta}) = \sigma^2 (X'X)^{-1} = \sigma^2 (Z'NZ)^{-1}, \quad (38.142)$$

где

$$\underset{((k+1)\times(k+1))}{\boldsymbol{Z}} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_{k+1} \\ x_1^2 & x_2^2 & \dots & x_{k+1}^2 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ x_1^k & x_2^k & \dots & x_{k+1}^k \end{pmatrix} \tag{38.143}$$

и

$$\underset{((k+1)\times(k+1))}{\boldsymbol{N}} = \begin{pmatrix} n_1 & & & & 0 \\ & n_2 & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \cdot \\ 0 & & & & n_{k+1} \end{pmatrix}. \tag{38.144}$$

Таким образом, мы видим, что поскольку имеется $k+1$ точка наблюдения и $k+1$ параметр, матрица $X'X$ становится произведением трех квадратных матриц. А значит,

$$\left|\frac{1}{\sigma^2}\boldsymbol{V}(\hat{\theta})\right|^{-1} = \frac{\sigma^{2(k+1)}}{|\boldsymbol{V}(\hat{\theta})|} = |\boldsymbol{X}'\boldsymbol{X}| = |\boldsymbol{Z}'||\boldsymbol{N}||\boldsymbol{Z}| = |\boldsymbol{N}||\boldsymbol{Z}|^2, \tag{38.145}$$

и для минимизации обобщенной дисперсии нужно максимизировать (38.145). Эту максимизацию мы проведем в два этапа. Каково бы ни было значение $|\boldsymbol{Z}|^2$,

$$|\boldsymbol{N}| = \prod_{i=1}^{k+1} n_i,$$

а значит, максимум, очевидно, достигается, если все n_i равны. Теперь остается провести максимизацию выбором x_i.

38.87 Нетрудно проверить, что

$$|\boldsymbol{Z}| = \prod_{\substack{i,j=1 \\ i<j}}^{k+1} (x_i - x_j),$$

так что

$$|\boldsymbol{Z}|^2 = \prod_{i<j} (x_i - x_j)^2. \tag{38.146}$$

Из того, что (38.146) представлено в виде произведения квадратов, сразу видно, что оно только увеличится, если экстремальные точки наблюдений передвинуть в концы интервала. Более того, максимальное значение выражения (38.146) только увеличится с ростом интервала. Таким образом, мы видим, что наблюдения всегда нужно проводить на максимально возможном интервале и по одной точке помещать в каждый конец этого интервала (мы все еще будем считать его интервалом

$(-1, +1)$). В каждой из $k+1$ точек нужно проводить равное число $n/(k+1)$ наблюдений.

38.88 Проведем теперь *непосредственно* максимизацию (38.146) при малых значениях k. При $k=1$ имеем линейный случай, и ничего больше делать не надо, поскольку результат **38.87** был подтвержден в примере 28.4. В квадратичном случае для максимизации (38.146) осталось выбрать x_2, поскольку $x_1 = -1$, $x_3 = +1$. В этом случае (38.146) сводится к

$$|\boldsymbol{Z}|^2 = 4(1 - x_2^2)^2$$

и достигает максимума при $x_2 = 0$. И без приведенных выкладок из соображений симметрии должно было быть понятно, что третья точка наблюдений должна быть нулем.

В кубическом случае используем явно симметрию и сведем задачу к выбору x_3 и $x_2 = -x_3$. Соотношение (38.146) превращается в

$$|\boldsymbol{Z}|^2 = 16x_3^2(1 - x_3^2)^4,$$

что достигает максимума при $x_3^2 = 1/5$.

Для полинома четвертой степени из симметрии следует, что $x_3 = 0$, и остается выбрать только x_4 и $x_2 = -x_4$. (38.146) сводится к

$$|\boldsymbol{Z}|^2 = 16x_4^6(1 - x_4^2)^4,$$

что достигает максимума при $x_4^2 = 3/7$.

Эти результаты, а также два других (для практических целей обычно больше и не требуется) собраны в следующей таблице.

Степень полинома	Точки наблюдения в $(-1, +1)$, в каждой из которых должно быть проведено $n/(k+1)$ наблюдений
1	± 1
2	± 1; 0
3	± 1; $\pm 0{,}4472$
4	± 1; $\pm 0{,}6547$; 0
5	± 1; $\pm 0{,}7651$; $\pm 0{,}2852$
6	± 1; $\pm 0{,}8302$; $\pm 0{,}4689$; 0

(38.147)

38.89 Хоул (1958) показал, что оптимальные точки наблюдений, максимизирующие (38.146), выражаются в терминах полиномов Лагранжа, определенных в **30.37**. Гест (1958) рассмотрел другой критерий оптимальности — минимизацию максимальной по всем точкам интервала дисперсии подогнанного полинома. Оптимальные значения оказались теми же самыми.

Гест показал, что оптимальные значения можно задать явно, как нули производной полинома k-го порядка. Тот же критерий оптимальности рассматривался в статье К. Смита (1918), который впервые вычислил значения из (38.147). По-видимому, это была первая детально решенная задача планирования, и самое удивительное заключается в том, что его статья была более или менее забыта в течение сорока лет. В статье Смита содержится также ряд диаграмм, сравнивающих дисперсии подогнанных полиномов в различных точках интервала для случаев, когда наблюдения проводятся: (а) оптимальным методом; (б) методом равномерного размещения наблюдений, который дает лучший результат в центре интервала и намного более плохой результат в экстремальных точках; (в) методом (б) с добавлением группы наблюдений в каждом конце интервала, чтобы уменьшить ухудшение, имеющее место для чисто равномерного размещения. Преимущество метода (в) состоит, разумеется, в том, что он не предполагает знания k и позволяет экспериментатору исследовать само значение k из наблюдений. Оптимальным же методом размещения нельзя воспользоваться для исследования более высоких значений k — об этом мы уже говорили в примере 28.4 для линейного случая. В любом случае кажется разумным использовать значения из (38.147), соответствующие наибольшему возможному k из тех, которые готов рассмотреть экспериментатор.

К. Смит (1918) рассматривает также влияние гетероскедастичности ошибок на оптимальное размещение. Хоул (1958) рассматривает некоторые специальные случаи коррелированных наблюдений.

38.90 Изучались также и другие критерии оптимальности. Хоул и Левайн (1964) и Левайн (1966) рассматривали вопрос размещения наблюдений в полиномиальной регрессии для минимизации дисперсии подогнанного полинома в специальной точке вне интервала наблюдений $(-1; +1)$. Оказалось, что это оптимальное размещение минимизирует максимальную дисперсию на интервале $(-1, x)$, если x достаточно велико. Гейлор и Свини (1965) изучали минимизацию максимальной дисперсии и средней дисперсии на любом интервале (не связанном с интервалом наблюдений) только для линейного случая. Дэвид и Аренс (1959) рассмотрели использование двух точек наблюдения в линейном случае для минимизации математического ожидания среднеквадратичной ошибки или максимума математических ожиданий квадрата ошибки, — последний случай отличается от минимизации максимальной дисперсии, поскольку допускается возможность, что истинная степень полинома может быть не 1, а 2. В очень общей статье Кифера и Волфовица (1959), не ограничиваясь полиномиальным случаем, применяются результаты теории игр для подсчета оптимального размещения при использовании нескольких критериев (см. также введение к статье Кифера (1959)). Хоул (1965a) применил эти методы для получения оптимального размещения в случае двумерных полиномов и тригонометрической регрессии. Хоул (1965b) для случая полиномиальной регрессии с одной переменной находит планы, минимизирующие дисперсию подогнанных значений для x, «экстраполированных» в интервал, лежащий непосредственно между двумя интервалами, в которых проводятся наблюдения. Он получает также соответствующие планы для полиномиальной регрессии от двух переменных, а также для обычной экстраполяции для случая двух переменных.

УПРАЖНЕНИЯ

38.1 В **38.25** проверить совпадение двух выражений (38.38) для СК, соответствующей разнице между обработками в блочном эксперименте.

38.2 Обобщая аргументы из **38.28**, показать, что если разбить оценки параметров обработок на группы и потребовать, чтобы корреляция между членами различных групп равнялась нулю, то каждый блок должен содержать одно и то же число обработок из данной группы. При этом эксперимент может быть разложен на множество независимых подэкспериментов с меньшим числом блоков в каждом.

(Точер, 1952.)

38.3 Пусть в **38.28** блочный эксперимент планируется так, чтобы сделать дисперсии оценок параметров обработок такими, какими они были бы, если бы не было эффектов блоков, т. е. $\sigma^2\{\mathrm{diag}(\boldsymbol{r})\}^{-1}$. Показать, что отсюда следует $\boldsymbol{M} = \boldsymbol{0}$, и значит, мы приходим к плану случайных блоков с матрицей инциденций, задаваемой (38.46).

(Точер, 1952.)

38.4 Проверить упрощенные формулы (38.47—51) для планов случайных блоков.

38.5 Проверить формулы из **38.32—3** для анализа по методу НК эксперимента, при котором сами блоки представлены в виде двухфакторной классификации. Привести таблицу ДА, аналогичную (38.39).

38.6 Опираясь на общие результаты **38.33** показать в **38.35**, что таблица ДА для плана латинского квадрата имеет вид

Источник дисперсии	СК	Число ст. св.
Разница между обработками	$\boldsymbol{T}'\boldsymbol{T}/t - G^2/t^2$	$t-1$
Строчки	$\boldsymbol{R}'\boldsymbol{R}/t - G^2/t^2$	$t-1$
Столбцы	$\boldsymbol{C}'\boldsymbol{C}/t - G^2/t^2$	$t-1$
Остаточная	$\boldsymbol{y}'\boldsymbol{y} - (\boldsymbol{T}'\boldsymbol{T} + \boldsymbol{R}'\boldsymbol{R} + \boldsymbol{C}'\boldsymbol{C})/t + 2G^2/t^2$	$(t-1)(t-2)$
Генеральное среднее	G^2/t^2	1
Полная	$\boldsymbol{y}'\boldsymbol{y}$	t^2

Обобщить это на греко-латинский квадрат и планы ортогональных квадратов более высоких порядков.

38.7 В 38.39 показать, что существует не более чем $t-1$ латинских квадратов порядка t, которые могут быть ортогональны в совокупности.

38.8 Проверить, что матрица, обратная к (38.79), имеет вид

$$\Omega^{-1} = D^{-1} - D^{-1}(w \vdots 1_t)\left(\begin{array}{c|c} 1 + 1_t' D^{-1} w & -1_t' D^{-1} 1_t \\ \hline -w' D^{-1} w & 1 + 1_t' D^{-1} w \end{array}\right)\begin{pmatrix} 1_t' \\ \hdashline w' \end{pmatrix} D^{-1}/\Delta,$$

где Δ — детерминант матрицы

$$\left(\begin{array}{c|c} 1 + 1_t' D^{-1} w & -1_t' D^{-1} 1_t \\ \hline -w' D^{-1} w & 1 + 1_t' D^{-1} w \end{array}\right).$$

Вывести отсюда, используя (38.23), что соотношение (38.85) справедливо для любой D.

(Точер, 1952.)

39.9 Показать, что если D определено так, чтобы разности параметров обработок оценивались с одинаковой потерей эффективности (сравнить с ситуацией, когда не нужно устранять эффекты блоков), то (38.85) ведет к уравнению (38.95) для плана СНБ. Показать, что если потеря эффективности равна нулю, то это сводится к плану случайных блоков (см. упражнение 38,3).

(Точер, 1952.)

38.10 Пусть задано полное множество ортогональных латинских квадратов порядка T (например, (38.76) для $T = 4$). Тогда для обозначения каждой из T^2 клеток таблицы нужна $T+1$ «ссылка», соответствующая, строчке, столбцу и одной из T букв одного из $T-1$ «алфавитов», образующих полное множество. Будем рассматривать теперь клетки таблицы как обработки и считать, что ссылкам соответствуют блоки в некотором эксперименте (совокупность блоков разбивается на $T+1$ множеств, состоящих из T элементов). Считая, что таблица задает размещение клеток по ссылкам, показать, что в результате получится план СНБ с

$$t = T^2, \quad b = T(T+1), \quad k = T, \quad r = T+1, \quad \lambda = 1.$$

Проверить, что в случае (38.76) результирующий план СНБ имеет вид

Номер блока	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Обработки в блоках	1	5	9	13	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
	2	6	10	14	5	6	7	8	6	5	8	7	7	8	5	6	8	7	6	5
	3	7	11	15	9	10	11	12	11	12	9	10	12	11	10	9	10	9	12	11
	4	8	12	16	13	14	15	16	16	15	14	13	14	13	16	15	15	16	13	14
Получены из	строчек				столбцов				латинских букв				греческих букв				чисел			

38.11 Для построенного в упражнении 38.10 плана СНБ рассмотрим $T+1$ дополнительных обработок, соответствующих множествам, на которые разбиваются блоки-ссылки, и применим к каждому блоку еще одну обработку в соответствии с множеством, к которому этот блок принадлежит. Добавим наконец еще один блок, в котором применяются все новые обработки. Показать, что в результате получается симметричный план СНБ с $t = b = T^2 + T + 1$, $k = r = T + 1$, $\lambda = 1$. Построить соответствующий расширенный план для плана, выведенного в упражнении 38.10 из (38.76).

Рассматривая случай $T = 6$ (см. 38.39) показать, что выполнение (38.93—5) недостаточно для существования плана СНБ.

38.12 В упражнении 38.10 показать, что для дуального плана, который получается, если положить $t' = b$, $b' = t$, $k' = r$ и $r' = k$, имеет место $\lambda < 1$, а значит, он не может быть планом СНБ. В упражнении 38.11 показать, что план, дуальный к расширению плана, выведенного из (38.76), имеет вид

Номер блока	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Обработки в блоках	1	1	1	1	2	2	2	2	3	3	3	3	4	4	4	4	1	5	9	13	17
	5	6	7	8	5	6	7	8	5	6	7	8	5	6	7	8	2	6	10	14	18
	9	10	11	12	10	9	12	11	11	12	9	10	12	11	10	9	3	7	11	15	19
	13	14	15	16	15	16	13	14	16	15	14	13	14	13	16	15	4	8	12	16	20
	17	18	19	20	20	19	18	17	18	17	20	19	19	20	17	18	21	21	21	21	21

Проверить, что это план СНБ.

38.13 Из (38.93) вывести, что для любого плана СНБ

$$r/(t-1) = (r-\lambda)/(t-k).$$

Используя (38.93—4), показать, что если $t = ck$, $b = cr$, как в абзаце, предшествующем (38.98), то

$$\frac{r(c-1)}{t-1} = \frac{r-\lambda}{k} = \frac{b-r}{t-1} = r - c\lambda = I,$$

где I — положительное целое число. Вывести отсюда (38.98)

(Мэрти, 1961.)

38.14 Используя (38.45—6), показать, что для случайных блоков оценка (38.118) совпадает с внутриблочной оценкой (38.29). Обе оценки сводятся к (38.48), так что в этом случае применение смешанной модели не изменяет оценки. Показать, что дисперсионная матрица оценок (38.112) равна $\frac{\sigma^2}{b}\{I_t + \rho 1_t 1_t'\}$, что отличается от внутриблочного результата ((38.40) и (38.47)). В частности, дисперсии оценок возрастают. Показать, что если за начало отсчета взять генеральное среднее выборки, то распределение величины $\hat{\tau}$ одинаково для двух моделей.

38.15 В 38.57 показать, используя (38.1), что

$$\mathbf{M}(B) = n'\tau$$

и

$$V(B) = k(1 + k\rho)\sigma^2 I_b.$$

Отсюда следует, что если матрица nn' несингулярна (это может быть, только если $b \geqslant t$), то можно несмещенно оценить параметры обработок, исходя из суммарных значений для блоков, положив

$$\hat{\tau}_B = (nn')^{-1} nB.$$

При этом

$$V(\hat{\tau}_B) = k(1 + k\rho)\sigma^2(nn')^{-1}.$$

Показать, что если за начало отсчета взять G (это соответствует вычитанию генерального среднего из параметров обработок), то $\hat{\tau}_B$ не коррелировано с определенным в (38.29) $\hat{\tau}$. Используя обобщенный метод НК из **19.17** (т. 2) и вид дисперсионной матрицы (38.40), показать отсюда, что линейной комбинацией с наименьшей дисперсией является

$$\hat{\tau}_I = \left\{ \operatorname{diag}(r) - \left(\frac{\rho}{1 + k\rho}\right) nn' + rr'/(bk) \right\}^{-1} \left\{ T - \left(\frac{\rho}{1 + k\rho}\right) nB \right\},$$

что согласуется с (38.29) при $\rho \to \infty$.

(См. Точер, 1952.)

38.16 Показать, что $\hat{\tau}_I$ из упражнения 38.15 совпадает с МД-оценкой $\hat{\tau}$ в (38.118) тогда и только тогда, когда $r'\hat{\tau} = G$. Используя (38.31) (где $\hat{\tau}$ определяется согласно (38.24)), показать, что это соотношение всегда выполняется.

38.17 В плане СНБ наблюдения внутри каждого блока заменяются на их ранги 1, ..., k. Пусть T_i — сумма полученных таким образом рангов, относящихся к i-й обработке и

$$S_t = \sum_{i=1}^{t} (T_i - T_.)^2 = \sum T_i^2 - t\left\{\frac{1}{2} r(k+1)\right\}^2.$$

Максимальное значение S_t равно $S_{\max} = \lambda^2 t(t^2 - 1)/12$. Обобщая **37.40—1** и упражнение 37.14, показать, что для $W = S_t/S_{\max}$ имеет место

$$\mathsf{M}(W) = \frac{k+1}{\lambda(t+1)},$$

$$\mathsf{D}(W) = \frac{2(k+1)^2}{\lambda^2 r t(t+1)^2}\left(r - 1 + \frac{\lambda - 1}{k - 1}\right)$$

и что величина

$$\left\{\frac{\lambda(t+1)}{k+1} - 1\right\} W/(1 - W)$$

имеет приблизительно распределение дисперсионного отношения с числом ст. св.

$$\nu_1 = t - 1 - \frac{2}{r}, \quad \nu_2 = (r - 1)\nu_1.$$

При $k = t$, $\lambda = r$ это сводится к результату упражнения 37.14, а план СНБ превращается в план случайных блоков.

(Дербин, 1951.)

38.18 Для любого плана СНБ, разложимого на r дубликатов t обработок показать, что среднее число обработок, общих для двух блоков из различных дубликатов, равно k^2/t, а сумма квадратов относительно среднего пропорциональна $b - t - r + 1$. Вывести отсюда (38.98) и доказать, что одинаковое число (т. е. k^2/t) обработок, общих для двух блоков из различных дубликатов, будет тогда и только тогда, когда в (38.98) выполняется равенство.

(Бозе, 1942.)

38.19 Проверить точки наблюдения, приведенные в таблице (38.147) для $k = 5$ и $k = 6$.

(К. Смит, 1918.)

38.20 Пусть t — нечетное, p — наименьший простой делитель t и $k \leqslant p$. Показать, что всегда можно сконструировать план СНБ следующим методом *равных разностей*. Обозначим обработки 0, 1, ..., $t - 1$ и построим $(t-1)/2$ исходных блока, содержащих обработки [0, 1, 2, ..., $k - 1$]; [0, 2, 4, ..., $2(k-1)$]; ...; $\left[0, \frac{1}{2}(t-1), (t-1), \ldots, \frac{1}{2}(k-1)(t-1)\right]$ (номера обработок получаются из приведенных здесь чисел взятием вычета по модулю t). Из каждого исходного блока образуем новый блок, добавив (по модулю t) к каждой обработке целое число r. Проделаем это для всех целых r, $1 \leqslant r \leqslant t - 1$. В результате для каждого исходного блока получится множество, состоящее из t блоков.

Показать, что для получившегося плана СНБ $b = \frac{1}{2}t(t-1)$, $r = \frac{1}{2}k(t-1)$ и $\lambda = \frac{1}{2}k(k-1)$.

(Гасснер, 1965.)

ГЛАВА 39

ТЕОРИЯ ВЫБОРОЧНЫХ ОБСЛЕДОВАНИЙ: ПЛАНИРОВАНИЕ

Оценивание средних для конечных совокупностей

39.1 В самом начале обсуждения случайного выбора в **9.1—4** т. 1 указывалось, что для применимости теории вероятностей к выборочной процедуре необходимо только, чтобы для каждой возможной выборки можно было определить вероятность ее выбора. При этом ни в коем случае не требовалось, чтобы процедура сводилась к простому случайному выбору, при котором каждая возможная выборка из генеральной совокупности имеет одну и ту же вероятность выбора. Мы указывали далее на различия между выбором без возвращения из конечной генеральной совокупности, при котором последовательные извлечения не независимы, и выбором с возвращением, при котором извлечения независимы.

Во всех трех томах почти всюду мы имели дело (и будем иметь дело) только с простым случайным выбором. Однако в этой главе и в следующей нас будут интересовать другие виды случайного выбора, которые на практике возникают при выборе из конечной совокупности, т. е. генеральной совокупности с конечным числом членов (которое мы будем всюду обозначать N). Даже при выборе с возвращением мы сталкиваемся тогда с новым явлением, которое возникает из-за конечности совокупности. Предполагается, что нам заранее известен перечень всех обследуемых элементов совокупности (в дальнейшем этот факт будет кратко формулироваться как «*опознаваемость* элементов совокупности»), а соответствующие им значения y мы узнаем только в результате извлечения. Так, что если одно и то же значение появляется в выборке дважды, то мы знаем, появился ли при этом дважды один и тот же элемент или два различных элемента с одним значением.

Непосредственной задачей теории выборочных обследований, как называется этот раздел теории, является почти всегда оценка средних (или, что эквивалентно, сумм) изучаемых случайных величин. Тем не менее мы излагаем ниже несколько более общую теорию. Так среднее любой функции от переменных (например, их квадратов) можно изучать тем же самым

методом. Изучением оценок дисперсий и других величин, связанных с генеральной совокупностью, мы будем заниматься лишь постольку, поскольку оно проясняет изучение нашего основного объекта, оценок средних. Результаты для долей получаются из результатов для средних, если считать, что переменные принимают только значения 0 или 1.

Итак, мы приступаем к изучению довольно узкого раздела статистической теории, но раздела, который интенсивно развивался не в силу математической привлекательности, а в силу практической важности. В последние годы в нескольких книгах приведен обзор и важные дополнения к обширной журнальной литературе. Особо стоит отметить книги Кокрэна (1963), Хансена, Гурвица и Мэдоу (1953) и Иэйтса (1960). Последняя из упомянутых книг содержит обширную библиографию теоретических и прикладных работ по выборочным обследованиям. Со времени первого издания в 1949 г. она была дополнена ссылками на работы, появившиеся вплоть до 1960 г. Мёрфи (1963) дает обзор дальнейших теоретических исследований.

Мы ограничимся обсуждением теоретических аспектов выборочных обследований. Наша цель — изложить их в контексте общей статистической теории. Нельзя надеяться, что при этом все важные результаты будут изложены красиво, но мы попытаемся минимизировать громоздкость, присущую этой области.

Случайный выбор с равными вероятностями без возвращения

39.2 Мы хотим оценить среднее μ *) величины y в конечной генеральной совокупности, состоящей из N элементов, основываясь на выборке из n элементов, извлеченной случайно, без возвращения с равными вероятностями выбора. Быть может, не совсем очевидно, что (при отсутствии какого-либо знания о виде совокупности) МД-оценкой величины μ является выборочное среднее. Для того чтобы показать это, используем метод наименьших квадратов из главы 19.

Рассмотрим модель

$$\underset{(n\times 1)}{\boldsymbol{y}} = \underset{(n\times 1)}{\mathbf{1}}\ \underset{(1\times 1)}{\mu} + \underset{(n\times 1)}{\boldsymbol{\varepsilon}}, \tag{39.1}$$

где $\mathbf{1}$ — вектор, все компоненты которого равны единице. Ошибки ε_i соответствуют отклонению наблюденных значений от среднего генеральной совокупности. Они не являются некор-

*) В этой и следующей главах мы для удобства обозначаем через μ (без индекса) среднее генеральной совокупности, обозначаемое ранее μ_1'. При этом μ_r, как обычно, обозначает центральный момент порядка r для $r > 1$. Соответствующие выборочные моменты обозначаются m (только в этих главах) и m_r, $r > 2$.

релированными, поскольку выбор не независим. Из соображений симметрии следует, что ковариации между любыми парами наблюдений в выборке совпадают. Обозначим их $\rho\mu_2$. Именно эти соображения симметрии позволяют нам надеяться, что наилучшей оценкой для μ будет выборочное среднее. Итак, дисперсионная матрица ошибок равна

$$\boldsymbol{V}(\varepsilon) = \mu_2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & \cdot & & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & & \cdot & \rho \\ \rho & \cdots & \rho & 1 \end{bmatrix} = \mu_2 \boldsymbol{V}. \tag{39.2}$$

Согласно **19.17** и упражнению 19.5 несмещенной, линейной МД оценкой μ является НК-оценка

$$\hat{\mu} = (\mathbf{1}'\boldsymbol{V}^{-1}\mathbf{1})^{-1}\mathbf{1}'\boldsymbol{V}^{-1}\boldsymbol{y} \tag{39.3}$$

с дисперсией

$$\mathsf{D}(\hat{\mu}) = \mu_2(\mathbf{1}'\boldsymbol{V}^{-1}\mathbf{1})^{-1}. \tag{39.4}$$

Чтобы применять (39.3—4), нужно вычислить $\boldsymbol{V}^{-1}$. Простым умножением на $\boldsymbol{V}$ проверяется, что

$$\boldsymbol{V}^{-1} = \frac{1}{(1-\rho)\{1+(n-1)\rho\}} \begin{bmatrix} \{1+(n-2)\rho\} & -\rho & \cdots & -\rho \\ -\rho & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & -\rho \\ -\rho & \cdots & -\rho & \{1+(n-2)\rho\} \end{bmatrix}, \tag{39.5}$$

а значит,

$$\mathbf{1}'\boldsymbol{V}^{-1} = \{1+(n-1)\rho\}^{-1}\mathbf{1}',$$

$$\mathbf{1}'\boldsymbol{V}^{-1}\mathbf{1} = n\{1+(n-1)\rho\}^{-1},$$

и (39.3—4) принимает вид

$$\hat{\mu} = m, \tag{39.6}$$

$$\mathsf{D}(\hat{\mu}) = \frac{\mu_2}{n}\{1+(n-1)\rho\}. \tag{39.7}$$

Пока что мы не вычислили корреляцию ρ. Мы умышленно отложили это вычисление, ибо вышеприведенные выкладки справедливы при любом ρ. В изучаемом же случае ρ легко вычислить, заметив, что если $n = N$, то $\mathsf{D}(\hat{\mu})$ обращается в нуль, поскольку выбрана вся совокупность. Таким образом, из (39.7) следует, что если $n = N$, то $\rho = -1/(N-1)$. Но понятно, что корреляция между парой выборочных значений не зависит от

размера выборки. Таким образом, вообще для случайной выборки без возвращения

$$\rho = -1/(N-1), \tag{39.8}$$

так что (39.7) принимает вид

$$\mathsf{D}(\hat{\mu}) = \frac{\mu_2}{n}\frac{N-n}{N-1}. \tag{39.9}$$

Моменты выборочного среднего и их несмещенные оценки

39.3 Фактически (39.9), а также значения третьего и четвертого центральных моментов случайной величины m уже были получены в примере 12.8, с использованием комбинаторных рассуждений. При этом с алгебраической точки зрения оказалось удобным проводить вычисления в терминах k-статистик выборки и совокупности, k_r и K_r соответственно, что мы будем делать и в этой главе. Мы будем называть теперь «дисперсией» совокупности, величину

$$\sigma^2 = K_2 = \frac{N}{N-1}\mu^2,$$

а «дисперсией» выборки

$$s^2 = k_2 = \frac{n}{n-1}m_2,$$

оставив для μ_2 и m_2 название «вторые моменты». В этих обозначениях перепишем результаты примера 12.8 (которые включают (39.9)) в виде*)

$$\mathsf{M}(m) = \mu,$$

$$\mathsf{D}(m) = \sigma^2\left(\frac{1}{n} - \frac{1}{N}\right),$$

$$\mathsf{M}(m-\mu)^3 = K_3\left\{\left(\frac{1}{n^2} - \frac{1}{N^2}\right) - \frac{3}{N}\left(\frac{1}{n} - \frac{1}{N}\right)\right\},$$

$$\mathsf{M}(m-\mu)^4 = K_4\left\{\left(\frac{1}{n^3} - \frac{1}{N^3}\right) - \frac{4}{N}\left(\frac{1}{n^2} - \frac{1}{N^2}\right) + \frac{6}{N^2}\left(\frac{1}{n} - \frac{1}{N}\right)\right\} +$$

$$+ 3\frac{(N-1)}{(N+1)}\left(\sigma^4 - \frac{K_4}{N}\right)\left\{\left(\frac{1}{n^2} - \frac{1}{N^2}\right) - \frac{2}{N}\left(\frac{1}{n} - \frac{1}{N}\right)\right\}. \tag{39.10}$$

Напомним, что, согласно (12.109),

$$\mathsf{M}(k_p) = K_p. \tag{39.11}$$

*) Здесь мы опустили индекс N у оператора математического ожидания M. В этой и следующей главах он всегда подразумевается.

Поэтому, чтобы получить несмещенные оценки центральных моментов величины m, можно в (39.10) подставить s^2 вместо σ^2 и k_3 вместо K_3. В результате получим

$$\hat{\mathsf{D}}(m) = s^2\left(\frac{1}{n} - \frac{1}{N}\right),$$
$$\{\mathsf{M}(\widehat{m-\mu})^3\} = k_3\left\{\left(\frac{1}{n^2} - \frac{1}{N^2}\right) - \frac{3}{N}\left(\frac{1}{n} - \frac{1}{N}\right)\right\}. \qquad (39.12)$$

Несмещенная оценка для четвертого момента получается несколько сложнее, поскольку соответствующее выражение в (39.10) нелинейно по K_r. Нам нужна несмещенная оценка для $\sigma^4 = K_2^2$. Из упражнения 12.11 сразу имеем

$$\mathsf{M}\left\{\frac{k_2^2 - \dfrac{(N-n)(Nn-n-N-1)}{n(n-1)N(N+1)}k_4}{\dfrac{2(N-n)}{(n-1)(N+1)}+1}\right\} = K_2^2,$$

что можно переписать в виде

$$\mathsf{M}\left\{\left(\frac{n-1}{n+1}\right)\left(\frac{N+1}{n-1}\right)k_2^2 - \frac{(N-n)(Nn-n-N-1)}{n(n+1)N(N-1)}k_4\right\} = K_2^2. \qquad (39.13)$$

Подставляя в правую часть последнего равенства в (39.10) случайную величину, заключенную в фигурные скобки в (39.13), вместо σ^4 и k_4 вместо K_4, получим несмещенную оценку четвертого момента случайной величины m.

39.4 Ничто не мешает нам в любом заданном случае оценить первые четыре момента m в соответствии с **39.3**, а затем подобрать распределение Пирсона, чтобы получить оценку выборочного распределения. Ситуация здесь абсолютно аналогична ситуации при оценивании методом максимального правдоподобия в **18.20** т. 2. Однако здесь, так же как и там при подгонке распределения по малой выборке, редко используются моменты. И это происходит не по лености, а из-за счастливого обстоятельства, что для размеров выборок, встречающихся на практике, центральная предельная теорема делает ненужной соответствующую работу. Хотя мы и не будем доказывать предельную нормальность m, стоит отметить, что в процессе предельного перехода появляется специфическое для нашей схемы условие. Нельзя просто устремить n к бесконечности, поскольку оно не может быть больше, чем N. Поэтому Мэдоу (1948), который получил центральную предельную теорему для этого случая, считает что как n, так и N возрастают при условии, что n/N остается отделенным от единицы. Из (39.10) легко проверить, что коэффициенты асимметрии и эксцесса для m стремятся при таком предельном переходе к соответствующим значениям для нормальной величины. Гаек (1960) приводит

необходимые и достаточные условия предельной нормальности m.

Таким образом, для распределения m можно использовать теорию стандартных ошибок, как это описано в **9.26—9** т. 1, и обычным образом проводить проверку гипотетического значения μ_0 для среднего генеральной совокупности или построение доверительных пределов для μ. Необходимо только, чтобы n было достаточно велико. Если n/N мало, то мы будем эффективно действовать, поступая так же, как и при простом случайном выборе.

Достаточность в теории выборочных обследований: опознаваемость элементов

39.5 В **17.31** т. 2 было изложено понятие достаточности. Достаточная статистика содержит всю имеющуюся в выборке информацию о параметре. Теперь в теории выборочных обследований мы не уточняем вид распределения генеральной совокупности, но, следуя Басу (1958) и Патхаку (1964с), в качестве параметра совокупности можем взять N-мерный вектор значений переменной y. Любая выборка из n наблюдений (с возвращением или без) доставляет информацию об n или меньшем числе компонент N-мерного вектора параметров. Любые две выборки (извлеченные по одной и той же схеме выбора), которые содержат одно и то же множество, состоящее из d ($\leqslant n$) элементов совокупности, доставляют одну и ту же информацию о параметре вне зависимости от того, появляются ли эти d элементов в двух выборках с одинаковыми частотами или нет. Поэтому такие выборки называются *эквивалентными*. Множество всех возможных выборок, получаемых при данной схеме выбора, можно разными способами разбить на подмножества. Если при разбиении каждое подмножество состоит из эквивалентных выборок, то такое разбиение называется *достаточным*. Если некоторые подмножества достаточного разбиения можно объединить и в результате получится другое достаточное разбиение, то последнее называют меньшим достаточным разбиением. Если имеется достаточное разбиение такое, что любое другое достаточное может быть сведено к нему упомянутым процессом объединения подмножеств, то это разбиение называется минимальным достаточным разбиением (см. **23.16** т. 2).

Любая статистика (вектор) t (т. е. любая функция от выборочных значений y) индуцирует разбиение множества всех возможных выборок на подмножества, состоящие из всех выборок с одним и тем же значением t. Если индуцируемое таким образом разбиение достаточно, то сама t является достаточной статистикой, поскольку ее значения определяют подмножества

эквивалентных выборок. При заданном t условное распределение любой другой статистики не даст нам никакой дополнительной информации о параметрах (т. е. свободно от параметров).

Теперь применима общая теория достаточных статистик из глав 17 и 23. В частности, справедлив результат Рао — Блекуэлла из **17.35** о том, что если задана любая несмещенная оценка некоторой функции параметров, то ее можно улучшить, взяв вместо нее условное ожидание при заданной достаточной статистике.

Как ни прост результат, из него получается неожиданное следствие. Рассмотрим простейший случай случайной выборки с возвращением из совокупности, состоящей из N элементов, для оценки среднего значения совокупности. Интуитивной оценкой является выборочное среднее m. Однако эту оценку можно улучшить. В самом деле, среди n выбранных элементов имеются такие, которые выбраны более одного раза. Как мы видели, в выборке имеется $d \leqslant n$ *различных* элементов, которые образуют базис эквивалентных выборок. Поэтому векторная статистика t, состоящая из значений $y^{(1)}, \ldots, y^{(d)}$, соответствующих этим различным элементам выборки, является достаточной. Условное ожидание m при заданном t, очевидно, является средним значением различных элементов выборки. Обозначим его m_d. Эта оценка имеет меньшую дисперсию чем m. Радж и Хамис (1958) явно вычисляют уменьшение дисперсии (см. упражнение 39.1). Здесь, как и в большинстве случаев, вычисление дисперсии оценки выборочного обследования, улучшенной по методу Рао — Блекуэлла, довольно сложно.

Как правило, при выборке с возвращением, если при построении любой оценки учитывать только *разные* выбранные элементы, а не все выбранные элементы, то точность оценки всегда улучшается. Однако дисперсию оценки, учитывающей все элементы, вычислять обычно проще, и она всегда является верхней границей для дисперсии оценки, учитывающей только разные элементы. Так что первую дисперсию можно смело использовать вместо второй. Если n/N велико, то уменьшение дисперсии редко будет достаточно большим, чтобы оправдать дополнительную работу, требующуюся для вычисления.

Патхак (1961a, b, 1962a, b, 1964a, b, c) провел ряд исследований о применении метода Рао — Блекуэлла к различным задачам теории выборочных обследований. См. также Дж. Рао (1966).

Выбор без возвращения с неравными вероятностями

39.6 Обобщим теперь схему случайного выбора без возвращения, позволив вероятностям выбора различаться для различных элементов и от извлечения к извлечению. Пусть ${}_{(r)}p_i$ — вероятность того, что i-й элемент (со значением y_i) выбирается при r-м извлечении, $\sum_{i=1}^{N} {}_{(r)}p_i = 1$, i меняется от 1 до N, а r — от 1 до n. Положим

$$\pi_i = \sum_{r=1}^{n} {}_{(r)}p_i, \tag{39.14}$$

$$\pi_{ij} = \sum_{\substack{r,\, s=1 \\ r \neq s}}^{n}\sum {}_{(r)}p_i \cdot {}_{(s)}p_j, \tag{39.15}$$

причем вероятности, соответствующие более позднему моменту времени в правой части (39.15), взяты при условии, что произошло событие, отвечающее другому сомножителю *). По определению π_i является полной вероятностью того, что y_i будет выбрано при n извлечениях, а π_{ij} — совместной вероятностью того, что выбрано как y_i, так и y_j $(i \neq j)$. Очевидно, что множество всех вероятностей ${}_{(r)}p_i$, как условных, так и безусловных, которое мы будем называть *схемой выбора*, определяет π_{ij} и π_i. Но одно и то же множество π_{ij} и π_i может быть связано с различными схемами выбора. Обычно трудно подобрать значения ${}_{(r)}p_i$, чтобы получить нужные значения π_{ij} и π_i, однако для схем выбора, описанных в упражнениях 39.5—6, соответствующие уравнения можно разрешить. Фелледжи (1963) дает рекурсивный метод получения ${}_{(r)}p_i$, равных при всех извлечениях. См. также Брюэр (1967).

Для выбора без возвращения с равными вероятностями, который мы уже рассматривали, имеем

$$\pi_i = \frac{n}{N}, \qquad \pi_{ij} = \frac{n(n-1)}{N(N-1)}.$$

В общем случае читатель может проверить, что из (39.14—15) следует

$$\sum_{i=1}^{N} \pi_i = n; \qquad \sum_{\substack{j=1 \\ j \neq i}}^{N} \pi_{ij} = (n-1)\pi_i; \qquad \sum_{\substack{i,\, j=1 \\ i \neq j}}^{N}\sum \pi_{ij} = n(n-1). \tag{39.16}$$

Примем теперь следующее соглашение об обозначениях. Индексы при выборочных значениях $y_1, \ldots, y_n$ соответствуют порядку, в котором эти значения извлекаются. Естественно, что эти индексы, вообще говоря, не совпадают с индексами, соот-

*) Следует обратить внимание читателя на некорректность обозначений, используемых в правой части (39.15). На самом деле для записи *условной* вероятности ${}_{(s)}p_j$ при $s > r$ (соответственно ${}_{(r)}p_i$ при $r > s$) нужно было бы использовать другие обозначения. Однако автор предпочитает ограничиться соответствующей оговоркой в тексте. (*Прим. перев.*)

ветствующими номеру элемента совокупности, т. е. $y_1, \ldots, y_N$. Так, например, y_3 в выборке (т. е. значение, полученное в результате третьего извлечения) не совпадает с y_3 в совокупности (т. е. со значением, соответствующим элементу с номером 3). Однако при рассмотрении симметрических функций от выборочных значений мы в целях простоты не будем обращать внимание на это различие.

При взятии математических ожиданий будем рассматривать выборку *как целое* (не учитывая порядок извлечения) и считать $nm = \sum_{i=1}^{n} y_i$ случайной величиной. Имеется $\binom{N}{n}$ возможных выборок. Мы предположим, что они записаны в некотором порядке и обозначены $S_1, \ldots, S_r, \ldots, S_{\binom{N}{n}}$. По определению

$$n\mathsf{M}(m) = \mathsf{M}\left(\sum_{i=1}^{n} y_i\right) = \sum_{r=1}^{\binom{N}{n}} \mathsf{P}\{S_r\}\left(\sum_{i=1}^{n} y_i\right)_r.$$

Стоящую в правой части этого равенства сумму $\binom{N}{n}$ членов, каждый из которых представляет собой сумму n значений, можно перегруппировать в сумму N членов (соответствующих элементам совокупности), причем каждый член, отвечающий данному элементу совокупности, будет равен сумме $\binom{N-1}{n-1}$ значений в соответствии с тем, какие $(n-1)$ из $(N-1)$ возможных элементов входят наряду с этим элементом в выборку. Итак,

$$n\mathsf{M}(m) = \sum_{i=1}^{N}\left[\sum_{r=1}^{\binom{N-1}{n-1}} \mathsf{P}\{S_r\}\right]_i y_i = \sum_{i=1}^{N} \pi_i y_i. \qquad (39.17)$$

Обозначим $\tau = \sum_{i=1}^{N} \pi_i y_i$. Таким образом, как и следовало ожидать, выборочное среднее не является, вообще говоря, несмещенной оценкой среднего совокупности. Кроме того, имеем

$$n^2\mathsf{D}(m) = \mathsf{D}\left(\sum_{i=1}^{n} y_i\right) = \mathsf{M}\left\{\left(\sum_{i=1}^{n} y_i\right)^2\right\} - \tau^2 =$$
$$= \mathsf{M}\left(\sum_{i=1}^{n} y_i^2\right) + \mathsf{M}\left(\sum_{\substack{i,j=1\\ i\neq j}}^{n} y_i y_j\right) - \tau^2.$$

Первое математическое ожидание вычисляется точно так же, как в (39.17), и равно $\sum_{i=1}^{N} \pi_i y_i^2$. Второе математическое ожидание является суммой $\binom{N}{n}$ членов, каждый из которых есть сумма $n(n-1)$ значений. Эту сумму можно перегруппировать, представив в виде суммы $N(N-1)$ членов по $\binom{N-2}{n-2}$ значений, в соответствии с тем, какие элементы, помимо двух заданных (из общего числа $N(N-1)$ пар), входят в выборку. Таким образом,

$$\mathsf{M}\left(\sum_{\substack{i,j=1\\ i\neq j}}^{n} y_i y_j\right) = \sum_{r=1}^{\binom{N}{n}} \mathsf{P}\{S_r\}\left(\sum_{\substack{i,j=1\\ i\neq j}}^{n} y_i y_j\right)_r =$$
$$= \sum_{\substack{i,j=1\\ i\neq j}}^{N}\left[\sum_{r=1}^{\binom{N-2}{n-2}} \mathsf{P}\{S_r\}\right]_{i,j} y_i y_j = \sum_{\substack{i,j=1\\ i\neq j}}^{N} \pi_{ij} y_i y_j,$$

а значит,

$$\mathsf{D}\left(\sum_{i=1}^{n} y_i\right) = n^2\mathsf{D}(m) = \sum_{i=1}^{N} \pi_i y_i^2 + \sum_{\substack{i,j=1\\ i\neq j}}^{N} \pi_{ij} y_i y_j - \tau^2 =$$
$$= \sum_{i=1}^{N} \pi_i (1-\pi_i) y_i^2 + \sum_{\substack{i,j=1\\ i\neq j}}^{N} (\pi_{ij} - \pi_i\pi_j) y_i y_j. \qquad (39.18)$$

39.7 Вычисление математических ожиданий (39.17—18), которое мы сейчас провели, является частным случаем следующего общего результата. Для любой функции g от наблюдений

$$\mathsf{M}\left\{\sum_{i=1}^{n} g(y_i)\right\} = \sum_{i=1}^{N} \pi_i g(y_i),$$
$$\mathsf{M}\left\{\sum_{\substack{i,j=1\\ i\neq j}}^{n} g(y_i, y_j)\right\} = \sum_{\substack{i,j=1\\ i\neq j}}^{N} \pi_{ij} g(y_i, y_j). \qquad (39.19)$$

Соотношения (39.19) уже не требуют нового доказательства, ибо изложенное выше доказательство не изменится, если y_i и $y_i y_j$ заменить на произвольные функции.

Теперь можно получить несмещенную линейную оценку для μ. Если значению y_i, соответствующему i-му элементу совокупности, приписать один и тот же вес w_i вне зависимости от того, когда выбран этот элемент, то мы хотим получить

$$\mathsf{M}\left\{\sum_{i=1}^{n} w_i y_i\right\} = \mu,$$

а значит, в соответствии с (39.19) при $g(y_i) = w_i y_i$

$$\sum_{i=1}^{N} \pi_i w_i y_i = \mu = \frac{1}{N} \sum_{i=1}^{N} y_i.$$

Это соотношение должно выполняться тождественно по μ, так что можно приравнять коэффициенты при y_i. В результате получим $w_i = 1/(N\pi_i)$, а значит, оценка

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{n} \frac{y_i}{\pi_i} \tag{39.20}$$

является единственной несмещенной оценкой указанного вида. Далее дисперсия величины $\hat{\mu}$ получается из (39.18) заменой y_i на $y_i/(N\pi_i)$. Таким образом,

$$N^2 \mathrm{D}(\hat{\mu}) = \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} y_i^2 + \sum_{\substack{i,\, j=1 \\ i \neq j}}^{N} \sum \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j} y_i y_j. \tag{39.21}$$

39.8 Используя (39.19), можно получить несмещенную оценку для (39.21). Нетрудно проверить, что такой оценкой является

$$N^2 \hat{\mathrm{D}}_1(\hat{\mu}) = \sum_{i=1}^{n} \frac{(1-\pi_i)}{\pi_i^2} y_i^2 + \sum_{\substack{i,\, j=1 \\ i \neq j}}^{n} \sum \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j \pi_{ij}} y_i y_j. \tag{39.22}$$

Эта оценка предложена Хорвитцем и Д. Томпсоном (1952), которые первыми сформулировали задачу, изложенную в **39.6.** С другой стороны, используя (39.16), можно переписать (39.21) в виде

$$N^2 \mathrm{D}(\hat{\mu}) = \frac{1}{2} \sum_{\substack{i,\, j=1 \\ i \neq j}}^{N} \sum (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2. \tag{39.23}$$

А значит, по (39.19) другой несмещенной оценкой для (39.21) является оценка

$$N^2 \hat{\mathrm{D}}_2(\hat{\mu}) = \frac{1}{2} \sum_{\substack{i,\, j=1 \\ i \neq j}}^{n} \sum \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2, \tag{39.24}$$

предложенная Иэйтсом и Гранди (1953).

Если бы мы могли выбрать π_i пропорционально y_i, то дисперсия (39.23) и ее оценка (39.24) обратились бы тождественно в нуль. (В упражнении 39.3 показано, что это не обязательно так для альтернативной оценки дисперсии (39.22).) На практике наилучшее приближение к этой ситуации получается в случае, когда мы знаем некоторую переменную, сильно коррелированную с y_i. Фелледжи (1963) и Дж. Рао (1963) обсуждают методы получения желаемых значений π_i и их влияние на выборочную дисперсию оценки. См. также Ханурав (1967).

Гаек (1964) исследует асимптотическую нормальность, дисперсию и оценку дисперсии для величины (39.20) в случае выбора с отбрасыванием (rejective sampling), когда производится выбор с возвращением с неравными вероятностями, однако если какой-нибудь элемент извлекается второй раз, то вся выборка не принимается во внимание.

С первого взгляда не совсем понятно, какая из приведенных двух оценок дисперсии предпочтительнее. Легко показать, что оценка (39.22) может принимать отрицательные значения (см. упражнение 39.3), однако (39.24) также может принимать отрицательные значения (см. упражнение 39.4). Оценка (39.24) неотрицательно определена только для таких схем, в которых $\pi_i \pi_j - \pi_{ij} \geqslant 0$ для всех i и j. Тем не менее для двух схем выбора, которые применяются на практике, оценка (39.24), как показывают упражнения 39.5—6, никогда не отрицательна, что неверно для (39.22). Поэтому не вызывает сомнения, что (39.24) более предпочтительна, чем (39.22), ибо последняя оценка, так сказать, более предрасположена принимать отрицательные значения, а значит, эта оценка величины (39.21) (математическое ожидание которой необходимо положительно) имеет большую дисперсию.

Похоже, что не существует оценки дисперсии (39.21), которая неотрицательно определена для всех схем выбора, однако, насколько нам известно, до сих пор нет доказательства этого факта.

39.9 Однако, взяв линейную оценку, отличную от (39.20), мы можем сделать так, что всегда будет существовать неотрицательная оценка для дисперсии взятой оценки. Но для этого мы не должны ограничиваться линейными функциями вида $\sum_{i=1}^{n} w_i y_i$, которые рассматривались при определении $\hat{\mu}$ в (39.20), поскольку, как мы видели, $\hat{\mu}$ является *единственной* несмещенной линейной функцией такого вида.

При построении линейных по множеству выборочных значений оценок среднего совокупности, коэффициенты, соответствующие каждому выборочному значению можно считать зависящими от следующих трех факторов:

(а) от номера того элемента, при выборе которого получается это значение;

(б) от извлечения, при котором это значение было выбрано;

(в) от всего множества элементов совокупности, вошедших в выборку, а не от отдельных элементов, как в (а).

Коэффициенты могут зависеть от любого одного из этих трех факторов, от любых двух и от всех трех факторов (а), (б), (в).

В результате получается семь общих классов линейных оценок, последний из которых включает все предыдущие. Эти классы изучались и анализировались Годэмбом (1955, 1965) и Купом (1963). Получено, что в упомянутом самом общем классе нельзя указать несмещенную оценку с минимальной дисперсией. Это утверждение интуитивно правдоподобно на том основании, что, зная значения, соответствующие элементам совокупности, можно всегда построить линейную несмещенную оценку с нулевой дисперсией, но эта оценка, очевидно, должна зависеть от самих значений, соответствующих элементам совокупности. Как показывает упражнение 39.3, примером такой оценки является (39.20). По существу отсутствие здесь оптимальной линейной оценки следует из того факта, что мы можем распознать элементы совокупности и назначаем вероятности выбора.

До сих пор мы рассматривали только коэффициенты, зависящие лишь от одного фактора (а). Рассмотрим теперь оценки с коэффициентами, зависящими от всех трех факторов (а), (б), (в).

39.10 Пусть теперь $y_{(r)}$ — значение, соответствующее элементу, выбранному при r-м извлечении. Пусть $p_{(r)}$ — условная вероятность выбрать этот элемент при этом извлечении, при условии, что он не был выбран до этого. Положим

$$z_1 = y_{(1)}/p_{(1)}, \qquad z_u = \sum_{r=1}^{u-1} y_{(r)} + y_{(u)}/p_{(u)}, \qquad u = 2, \ldots, n. \tag{39.25}$$

Каждая величина nz_u является несмещенной оценкой для $N\mu$, поскольку, согласно (39.19),

$$\mathsf{M}(z_1) = \sum_{i=1}^{N} y_i,$$

а при $u \geqslant 2$

$$\mathsf{M}(z_u) = \mathsf{M}\left[\sum_{r=1}^{u-1} y_{(r)} + \mathsf{M}\left\{\frac{y_{(u)}}{p_{(u)}} \mid y_{(1)}, \ldots, y_{(u-1)}\right\}\right] =$$

$$= \mathsf{M}\left[\sum_{r=1}^{u-1} y_{(r)} + \left\{N\mu - \sum_{r=1}^{u-1} y_{(r)}\right\}\right] = \mathsf{M}\{N\mu\} = N\mu,$$

а значит, любая линейная функция

$$z = \sum_{u=1}^{n} c_u z_u, \qquad \sum_{u=1}^{n} c_u = 1 \tag{39.26}$$

является несмещенной оценкой для $N\mu$. Наибольшей симметрией из таких функций обладает среднее

$$\bar{z} = \frac{1}{n}\sum_{u=1}^{n} z_u = \frac{1}{n}\left\{\sum_{u=1}^{n} \frac{y_{(u)}}{p_{(u)}} + \sum_{u=2}^{n}\sum_{r=1}^{u-1} y_{(r)}\right\}. \tag{39.27}$$

Как следует из упражнения 39.10, при равных вероятностях выбора (39.27) не сводится к Nm. Дисперсия оценки $\bar{z}$ равна

$$\mathsf{D}(\bar{z}) = \frac{1}{n^2}\left\{\sum_{u=1}^{n} \mathsf{D}(z_u) + \sum_{\substack{u,\,v=1\\ u\neq v}}^{n}\sum \operatorname{cov}(z_u, z_v)\right\}. \tag{39.28}$$

Вычисляя ковариацию z_u и z_v в два этапа, сначала при фиксированном z_u, а потом при изменении z_u, получим сразу, что $\operatorname{cov}(z_u, z_v) = 0$ при $u \neq v$. А значит,

$$\mathsf{M}(z_u z_v) = \mathsf{M}(z_u)\,\mathsf{M}(z_v) = N^2\mu^2. \tag{39.29}$$

Таким образом, нам осталось вычислить только дисперсии в (39.28). В общем случае вычисление $\mathsf{D}(z_u)$ весьма громоздко (см. упражнение 39.9), но для $\mathsf{D}(\bar{z})$ легко получить оценку. В самом деле,

$$\mathsf{D}(\bar{z}) = \mathsf{M}(\bar{z}^2) - N^2\mu^2,$$

так, что используя (39.29), получим, что несмещенной оценкой для $\mathsf{D}(\bar{z})$ является $\bar{z}^2 - z_u z_v$ для любых $u \neq v$. Усреднив эти оценки по всем $n(n-1)/2$ различным парам u и v, получим оценку

$$\hat{\mathsf{D}}(\bar{z}) = \bar{z}^2 - \frac{1}{n(n-1)}\sum_{u\neq v} z_u z_v, \tag{39.30}$$

которая совпадает с

$$\hat{\mathsf{D}}(\bar{z}) = \frac{1}{n(n-1)}\sum_{u=1}^{n}(z_u - \bar{z})^2 \geqslant 0. \tag{39.31}$$

Оценка (39.31) имеет вид s^2/n, где s^2 — вторая k-статистика для наблюдений z_u. Таким образом, при описанном подходе, предложенном Раджем (1956), задача выбора y_i (с неравными вероятностями и без возвращения) сводится к выбору z_u и последующему вычислению их среднего значения и оцениванию дисперсии этого среднего точно так же, как если бы z_u выбирались с равными вероятностями и с возвращением. Это утверждение является частным случаем общего результата, приведенного в упражнении 39.12.

39.11 Нам осталось решить, какую из оценок (39.20) или (39.27) использовать. В общем случае об их относительной

эффективности известно мало, но Радж (1956) описывает два выборочных эксперимента с $n=2$, в которых (39.27) значительно предпочтительнее.

Применяя, однако, метод Рао — Блэкуэлла улучшения оценок (см. **39.5**), можно получить оценки, дисперсия которых заведомо не больше, чем у (39.27). Вычислим (39.27) (которое мы теперь переобозначим $\bar{z}_{(s)}$, чтобы подчеркнуть зависимость от порядка извлечения в выборке) для каждой из $n!$ возможных перестановок, соответствующих порядку, в котором могли быть извлечены выбранные значения. Усредним теперь эти $n!$ значений с весами, равными относительным вероятностям этих перестановок, в результате получим улучшенную оценку $\bar{z}_s$. Точно так же можно пересмотреть оценку (39.31), чтобы получить улучшенную оценку величины $\mathsf{D}(\bar{z}_{(s)})$. Эти результаты, принадлежащие Мёрфи, являются непосредственным следствием результата Рао — Блэкуэлла из **17.35** и приведены в виде упражнений 39.30—1. Хотя и не показано, что дисперсия величины $\bar{z}_s$ имеет неотрицательную оценку для всех схем выбора, что справедливо для $\bar{z}_{(s)}$, но Мёрфи показал, что это так при $n=2$. Более того, Мёрфи (1957) показал, что в уже упомянутых выборочных экспериментах Раджа (1956) оценка $\bar{z}_s$ более эффективна, чем $\bar{z}_{(s)}$, и что отклонения несмещенной оценки дисперсии величины $\bar{z}_s$ меньше, чем у (39.31) для $\bar{z}_{(s)}$. Наконец, улучшенная оценка для $\mathsf{D}(\bar{z}_{(s)})$ также получает заметное улучшение эффективности по сравнению с (39.31).

Таким образом, имеются существенные теоретические основания для использования улучшенной оценки $\bar{z}_s$. Однако с ростом n задача вычислений становится очень громоздкой, поскольку нужно рассматривать $n!$ различных перестановок выбранных элементов. А даже для вычисления только одного $\bar{z}_{(s)}$, определенного в (39.27), нужно вычислить n условных вероятностей, что требует большой работы, если размер совокупности N и размер выборки n велики. Если упростить схему выбора, как в упражнении 39.31, считая что вероятности пропорциональны при всех извлечениях, то выражение для $\bar{z}_s$ упрощается, но все еще остается громоздким для вычислений.

Неравные вероятности при выборе с возвращением

39.12 Если выбор проводится с возвращением, то извлечения независимы и существует возможность, что любое y_i будет извлечено более одного раза. Определим π_i и π_{ij}, как и ранее, согласно (39.14—15), но теперь нужно определить еще

$$\pi_{ii} = \sum_{\substack{r,\, s=1 \\ r \neq s}}^{n}\sum {}_{(r)}p_i \cdot {}_{(s)}p_i.$$

Теперь π_{ij} — вероятность того, что как y_i, так и y_j выбраны по крайней мере один раз, а π_{ii} — вероятность того, что y_i выбрано не менее двух раз. (39.16) заменяется на

$$\sum_{i=1}^{N}\pi_i = n; \quad \sum_{j=1}^{N}\pi_{ij} = (n-1)\pi_i; \quad \sum_{i=1}^{N}\sum_{j=1}^{N}\pi_{ij} = n(n-1), \tag{39.32}$$

причем суммы берутся по всем i и j без ограничений. Аналогично (39.19) заменяется на

$$\begin{aligned} \mathsf{M}\left\{\sum_{i=1}^{n} g(y_i)\right\} &= \sum_{i=1}^{N}\pi_i g(y_i), \\ \mathsf{M}\left\{\sum_{\substack{i,\, j=1 \\ i\neq j}}^{n}\sum g(y_i, y_j)\right\} &= \sum_{i,\, j=1}^{N}\sum \pi_{ij} g(y_i, y_j), \end{aligned} \tag{39.33}$$

причем снова в последней двойной сумме в (39.33) на i и j не накладывается никаких ограничений. Как может проверить читатель, в (39.18) и (39.21) π_{ij} может теперь иметь равные индексы, а члены $\pi_i\pi_j$ входят в двойную сумму только с разными индексами. Если вместо (39.16) использовать теперь (39.32), то получим, что (39.23) останется без изменения.

В частном случае, когда вероятности выбора одни и те же при каждом извлечении, то p_i можно писать без стоящего впереди индекса, и теория упрощается. В этом случае $\pi_i = np_i$, $\pi_{ij} = n(n-1)p_ip_j$, оценка (39.20) сводится к

$$\hat{\mu} = \frac{1}{Nn}\sum_{i=1}^{n}\frac{y_i}{p_i}, \tag{39.34}$$

а (39.23) — к

$$N^2\mathsf{D}(\hat{\mu}) = \frac{1}{2n}\sum_{i,\, j=1}^{N}\sum p_ip_j\left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2. \tag{39.35}$$

Используя (39.33), получим несмещенную оценку для (39.35) в виде

$$N^2\hat{\mathsf{D}}_2(\hat{\mu}) = \frac{1}{2n^2(n-1)}\sum_{i,\, j=1}^{n}\sum\left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2.$$

Учитывая (2.27) т. 1, преобразуем это выражение к виду

$$N^2\hat{\mathsf{D}}_2(\hat{\mu}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{p_i} - \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i}\right)^2 = \frac{s^2_{y/p}}{n}, \tag{39.36}$$

где s^2 — вторая выборочная k-статистика для y_i/p_i, определяе-

мая, как и в **39.3**. Простой вид (39.36) основан на том, что $y_i/(Np_i)$ являются несмещенными оценками μ некоррелированными из-за того, что выбор проводится с возвращением, так что применимо упражнение 39.12.

Планирование выборки: расслоение

39.13 В **39.8** мы видели, что на практике нельзя надеяться найти схему выбора, которая уменьшала бы выборочную дисперсию до ее достижимого минимума, равного нулю. Однако, как там показано, если мы имеем некоторую информацию (хотя, возможно, и довольно нечеткую) о переменной y или о некоторой другой переменной, коррелированной с ней, то эта информация позволяет нам получить существенное улучшение по сравнению с простым случайным выбором. *Планирование выборки* (т. е. подбор π_{ij}, а значит, и π_i) преследует цель уменьшить, насколько это возможно, дисперсию оценки. (В дальнейшем мы модифицируем это утверждение, чтобы учесть изменение стоимости различных процедур выбора.) Если, как это обычно и бывает, мы имеем дело с несколькими величинами одновременно, то самое большее на что можно надеяться, это подобрать некоторый «компромиссный» набор π_{ij}, для которого все используемые оценки имели бы достаточно малую дисперсию. Именно эта необходимость компромисса дает толчок к рассмотрению различных классов схем выбора, цель которых состоит в уменьшении дисперсии.

39.14 Рассмотрим сначала (39.23), когда π_i фиксированы, а подбираются только π_{ij} (удовлетворяющие ограничениям, налагаемым (39.16)). Выражение (39.23) представляет собой сумму $N(N-1)/2$ членов, каждый из которых равен коэффициенту $(\pi_i\pi_j - \pi_{ij})$, умноженному на неотрицательную величину $\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$, в которой значения π_i теперь фиксированы. Если значения y_i неизвестны, то, как мы уже говорили, нельзя сказать, какой набор π_{ij} будет оптимальным. Но вне зависимости от значений y_i мы видим, что если положить π_{ij} равным $\pi_i\pi_j$, то соответствующий коэффициент обратится в нуль. Далее в **39.6** мы видели, что для обычного выбора без возвращения с равными вероятностями $\pi_i = \frac{n}{N}$, $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$ для всех $i \neq j$, а значит, $\pi_i\pi_j - \pi_{ij} = \frac{n(N-n)}{N^2(N-1)} > 0$. А значит, в этом случае каждый из $N(N-1)/2$ членов в (39.23) вносит положительный вклад в дисперсию. Отсюда видно, что если положить все π_i равными n/N, а некоторые π_{ij} равными $\pi_i\pi_j = n^2/N^2$, то мы уменьшаем до нуля вклад в дисперсию нашей оценки от

членов, соответствующих этим парам i, j. Но в силу (39.16) сумма π_{ij} равна $n(n-1)$, а значит, небольшое увеличение некоторых π_{ij} от значения $\frac{n(n-1)}{N(N-1)}$ до $\frac{n^2}{N^2}$ должно компенсироваться уменьшением других π_{ij}, что ведет к увеличению вклада в дисперсию соответствующих членов. Однако если эти последние уменьшенные π_{ij} связаны с меньшими значениями $|y_i - y_j|$, а увеличенные π_{ij} связаны с бо́льшими значениями $|y_i - y_j|$, то можно ожидать, что в результате получится уменьшение дисперсии оценки. Более того, поскольку те π_{ij}, которые увеличиваются, изменяются незначительно, то компенсирующее уменьшение других π_{ij} может быть тоже малым, особенно если число тех π_{ij}, которые уменьшаются, не меньше, чем тех, которые увеличиваются.

Таким образом, мы пришли к следующему принципу улучшения дисперсии оценки в случае равных π_i. Увеличить π_{ij} от $\frac{n(n-1)}{N(N-1)}$ до $\frac{n^2}{N^2}$, если $|y_i - y_j|$ велико и уменьшить π_{ij}, если $|y_i - y_j|$ мало. Этот принцип делается более понятным, если обратить внимание на то, что если $\pi_{ij} = \pi_i\pi_j$, то y_i и y_j должны выбираться *независимым* образом. Поэтому мы исследуем процедуры, в которых схема выбора разбивается на две или большее число подсхем, причем выбор внутри каждой из подсхем производится *независимо* от других. Возвращаясь к обозначениям **39.6**, это утверждение можно сформулировать в терминах схемы выбора в следующем виде:

$$\left.\begin{array}{ll}
\text{при } 1 \leqslant i \leqslant N_1 & \begin{cases} {}_{(r)}p_i > 0 & \text{для } 1 \leqslant r \leqslant n_1, \\ {}_{(r)}p_i = 0 & \text{для других } r; \end{cases} \\
\text{при } N_1 < i \leqslant N_1 + N_2 & \begin{cases} {}_{(r)}p_i > 0 & \text{для } n_1 < r \leqslant n_1 + n_2, \\ {}_{(r)}p_i = 0 & \text{для других } r; \end{cases} \\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\
\text{при } \sum_{l=1}^{k-1} N_l < i \leqslant N & \begin{cases} {}_{(r)}p_i > 0 & \text{для } \sum_{l=1}^{k-1} n_l < r \leqslant n, \\ {}_{(r)}p_i = 0 & \text{для других } r. \end{cases}
\end{array}\right\} \tag{39.37}$$

В (39.37) N элементов совокупности разбиты на k групп, состоящих из N_l элементов, $\sum_{l=1}^{k} N_l = N$. Выборка аналогично разбита на k соответствующих выборок размеров n_l, $\sum_{l=1}^{k} n_l = n$. Выбор в каждой подсовокупности проводится независимо от других.

Из нашего принципа теперь следует, что y_i должны так разбиваться на группы, чтобы члены разных групп отличались как можно больше (поскольку в (39.23) парам из разных групп соответствуют нулевые коэффициенты), а элементы одной группы должны отличаться как можно меньше (поскольку они соответствуют парам, для которых π_{ij} уменьшается).

Выбор элементов в подгруппах проводится независимо, а результаты объединяются для получения оценки параметра полной совокупности. Совокупность элементов, принадлежащих подгруппе, называется *слоем* (*strata*). Эта терминология связана с геологическими исследованиями, в которых производится естественное разбиение на слои пород. Для детального изучения *расслоенного выбора*, как будет называться схема выбора (39.37), нам будет удобно переформулировать задачу заново.

Расслоенный случайный выбор

39.15 Как и в **39.14** предположим, что совокупность разбита на k слоев и l-й слой содержит N_l элементов, $\sum_{l=1}^{k} N_l = N$. Рассмотрим частный случай общей схемы (39.37), считая, что выбор в каждом слое проводится независимо от других слоев и что в l-м слое проводится выборка n_l элементов без возвращения с равными вероятностями. Такая схема называется *расслоенным случайным выбором* без возвращения. В силу независимости выбора для различных слоев теория достаточно проста. Обозначим значения, соответствующие элементам l-го слоя, через y_{li} и условимся, что первый индекс l всегда будет относиться к слою.

Очевидно, что для всех i в l-м слое имеет место $\pi_i = n_l/N_l$. А значит, несмещенная оценка (39.20) принимает вид

$$\hat{\mu} = \frac{1}{N}\sum_{l=1}^{k}\sum_{i=1}^{n_l}\frac{y_{li}}{n_l/N_l} = \frac{1}{N}\sum_{l} N_l m_l, \tag{39.38}$$

где m_l — среднее выборки в l-м слое, истинное среднее которого обозначается μ_l *). Поскольку случайные величины m_l независимы, то дисперсия величины $\hat{\mu}$ равна

$$\mathrm{D}(\hat{\mu}) = \frac{1}{N^2}\sum_{l} N_l^2 \mathrm{D}(m_l) = \frac{1}{N^2}\sum_{l} N_l^2 \frac{\sigma_l^2}{n_l}\left(1 - \frac{n_l}{N_l}\right), \tag{39.39}$$

*) Мы не боимся, что читатель перепутает это с обозначением для моментов μ_r, поскольку было сказано, что индекс l всегда связывается со слоем.

где (39.10) применяется к каждому слою в отдельности, а σ_l^2 — дисперсия l-го слоя. Если аналогично обозначить через s_l^2 выборочную дисперсию в l-м слое, то из (39.12) получим следующую несмещенную оценку для (39.39):

$$\hat{\mathrm{D}}(\hat{\mu}) = \frac{1}{N^2}\sum_{l} N_l^2 \frac{s_l^2}{n_l}\left(1 - \frac{n_l}{N_l}\right). \tag{39.40}$$

Если мы хотим, чтобы все π_i были равны, как это было в **39.14**, где мы пришли к понятию расслоенного выбора, то нужно, чтобы n_l/N_l было одним и тем же для всех l. Этот случай называется выбором с равными долями (ВРД)*). Но нам не обязательно ограничивать наше изложение этим случаем, поскольку (39.38—40) справедливы при любом выборе n_l, так что мы можем рассматривать случай и неравных долей. На самом деле нас будет интересовать вопрос о том, как наиболее эффективно выбрать n_l или более точно, какими, при фиксированном $n = \sum_l n_l$ должны быть целые числа n_l, чтобы минимизировать дисперсию (39.39). Наше изложение следует Армитеджу (1947), хотя основной результат впервые был получен Чупровым в 1920 г. и независимо Нейманом (1934).

Выбор объема слоев выборки

39.16 Дисперсия (39.39) может быть переписана в виде

$$\frac{1}{N^2}\sum_{l} N_l^2 \frac{\sigma_l^2}{n_l}\left(1 - \frac{n_l}{N_l}\right) \equiv \frac{(N-n)}{nN^2}\sum_{l} N_l\sigma_l^2 + \\ + \frac{1}{N^2}\sum_{l} n_l\left(\frac{N_l\sigma_l}{n_l} - \frac{\sum_l N_l\sigma_l}{n}\right)^2 - \frac{1}{nN}\sum_{l} N_l\left(\sigma_l - \frac{\sum_l N_l\sigma_l}{N}\right)^2. \tag{39.41}$$

Проверка этого равенства проводится простыми алгебраическими выкладками. Из трех членов в правой части (39.41) только второй зависит от объемов слоев выборки n_l, причем этот член неотрицателен, а значит, дисперсия минимальна, если n_l выбрать так, чтобы этот член обратился в нуль, т. е. при

$$\frac{N_l\sigma_l}{n_l} - \frac{\sum_l N_l\sigma_l}{n} = 0$$

*) В отечественной литературе употребляется также термин «пропорциональный выбор». (*Прим. перев.*)

или

$$\frac{n_l}{N_l}=\frac{\sigma_l}{\sum\limits_l N_l\sigma_l/n}. \qquad (39.42)$$

Таким образом, дисперсия минимальна, если *доля выбранных элементов* в каждом слое пропорциональна квадратному корню из дисперсии этого слоя. Такой выбор n_l мы назовем *размещением по слоям с минимальной дисперсией*, а оценку, соответствующую этому случаю, обозначим $\hat{\mu}_{МД}$. Объемы выборок, определяемые таким способом, как правило, получаются дробными, и на практике нужно брать ближайшие к ним целые числа.

Из вышесказанного следует, что минимальное значение дисперсии задается соотношением (39.41) с опущенным вторым членом в правой части, что дает после упрощения

$$\mathsf{D}(\hat{\mu}_{МД})=\frac{1}{N^2}\left\{\frac{1}{n}\left(\sum_l N_l\sigma_l\right)^2-\sum_l N_l\sigma_l^2\right\}. \qquad (39.43)$$

Как показано в упражнении 39.21, значение дисперсии ненамного изменяется при малых отклонениях n_l от значений, определяемых (39.42).

39.17 Предположим, с другой стороны, что мы взяли все n_i равными, как при первоначальном изложении, положив для всех l

$$\frac{n_l}{N_l}=\frac{n}{N}. \qquad (39.44)$$

Из (39.38) сразу следует, что в этом случае $\hat{\mu}$ сводится к среднему полной выборки. Обозначим это среднее $\hat{\mu}_{ВРД}$.

Теперь легко заметить, что второй и третий члены в правой части (39.41) равны по величине и противоположны по знаку, а значит, сокращаются. Таким образом, значение дисперсии, получающееся в результате ВРД, определяемого (39.44), равно одному первому члену в (39.41). Это выражение больше, чем минимально возможное, за исключением случая, когда последний член правой части (39.41) обращается в нуль, т. е. в случае

$$\sigma_l=\sum N_l\sigma_l/N \quad \text{для всех} \quad l.$$

Это возможно только тогда, когда дисперсии всех слоев равны. Естественно, что при этом условии (39.44) и (39.42) совпадают.

39.18 Сравним дисперсию при ВРД (см. **39.17**) с дисперсией, получающейся при случайном выборе с равными вероятностями без расслоения. В последнем случае дисперсия задается соотношением (39.10), которое можно переписать в виде *)

$$\mathsf{D}(m_R)=\sigma^2\left(\frac{1}{n}-\frac{1}{N}\right)\equiv$$
$$\equiv\frac{N-n}{nN(N-1)}\left\{\sum_l(N_l-1)\sigma_l^2+\sum_l N_l(\mu_l-\mu)^2\right\}.$$

Это тождество справедливо в силу следующего тождества дисперсионного анализа:

$$(N-1)\sigma^2\equiv\sum_l(N_l-1)\sigma_l^2+\sum_l N_l(\mu_l-\mu)^2,$$

которое есть не что иное, как (35.25) в других обозначениях. В **39.17** мы видели, что при распределении по слоям, соответствующим ВРД,

$$\mathsf{D}(\hat{\mu}_{ВРД})=\frac{N-n}{nN^2}\sum N_l\sigma_l^2, \qquad (39.45)$$

и это выражение надо сравнить с $\mathsf{D}(m_R)$. Их разность равна

$$\mathsf{D}(m_R)-\mathsf{D}(\hat{\mu}_{ВРД})=$$
$$=\frac{N-n}{nN}\left[\left\{\frac{\sum\limits_l(N_l-1)\sigma_l^2}{N-1}-\frac{\sum\limits_l N_l\sigma_l^2}{N}\right\}+\frac{\sum\limits_l N_l(\mu_l-\mu)^2}{N-1}\right]^2. \qquad (39.46)$$

Член в фигурной скобке в правой части (39.46) отрицателен, поскольку он равен

$$\frac{1}{N(N-1)}\sum_l(N_l-N)\sigma_l^2<0. \qquad (39.47)$$

Таким образом, из (39.46) сразу следует, что если последний член в правой части этого равенства обращается в нуль, т. е. если все μ_l равны, то

$$\mathsf{D}(m_R)<\mathsf{D}(\hat{\mu}_{ВРД}),$$

так что расслоение с ВРД приводит к *увеличению* дисперсии. Более того, в **39.17** мы уже видели, что если все σ_l равны, то ВРД совпадает с размещением, обладающим свойством МД. Таким образом, если все σ_l равны и все μ_l равны, то

$$\mathsf{D}(m_R)<\mathsf{D}(\hat{\mu}_{ВРД})=\mathsf{D}(\hat{\mu}_{МД}). \qquad (39.48)$$

В этих условиях при *любом* распределении по слоям получается увеличение дисперсии оценки среднего всей совокупности.

*) Индекс R соответствует случайному выбору без возвращения с равными вероятностями.

Даже если μ_l различаются, но мало, и σ_l мало отличаются, все еще возможна ситуация, аналогичная (39.48), т. е. может оказаться

$$\mathsf{D}(m_R) < \mathsf{D}(\hat{\mu}_{\text{МД}}) < \mathsf{D}(\hat{\mu}_{\text{ВРД}}).$$

Более подробно см. Армитедж (1947); часть его результатов содержится в упражнениях 39.14—15.

39.19 Случаи, аналогичные (39.48), редко встречаются в выборочной практике, ибо они связаны с неравенством (39.47). А левая часть (39.47) имеет при фиксированном k порядок N^{-1}, в то время как другой член в квадратных скобках правой части (39.46), равный $\sum N_l(\mu_l - \mu)^2/(N-1)$ имеет порядок единицы, если N_l — величины того же порядка, что и N. А значит, при $N \to \infty$ и N_l/N фиксированных, членом в фигурных скобках в правой части (39.46) можно пренебречь, а второй член неотрицателен, так что

$$\mathsf{D}(m_R) \geqslant \mathsf{D}(\hat{\mu}_{\text{ВРД}}) \geqslant \mathsf{D}(\hat{\mu}_{\text{МД}}), \tag{39.49}$$

причем равенство в левой части достигается в пределе тогда и только тогда, когда все μ_l равны, а равенство в левой части тогда и только тогда, когда равны все σ_l.

Соотношение (39.49) соответствует тому, что мы ожидали, исходя из интуитивных соображений в **39.14**. Как там, так и в более общей ситуации в **39.16—18** мы видели, что именно различие между μ_l приводит к улучшению дисперсии при использовании расслоенной выборки с ВРД. (Дальнейшее улучшение при распределении с МД связано с различием среди σ_l). Это согласуется с общим выводом в конце **39.14**, где рассматривался только случай ВРД, о том, что слои должны как можно больше различаться.

Использование слоев в планировании выборочных обследований по понятным соображениям аналогично использованию блоков при планировании экспериментов (см. **38.14**). В обоих случаях группировка (элементов, экспериментальных объектов) проводится с целью устранения ошибок (выборочных, экспериментальных), возникающих из-за различия групп (слоев, блоков). Однако наряду с общностью методов имеется разница в целях. При обследованиях интересуются в первую очередь оценкой среднего полной совокупности, в то время как при изучении экспериментальных данных генеральное среднее редко представляет интерес. Эта разница является отражением того факта (см. **38.3**), что основой для анализа экспериментов служит гипотетическая, а не реальная генеральная совокупность.

Размещение с минимальной дисперсией при фиксированной полной стоимости

39.20 Перед тем как закончить обсуждение вопроса о размещении выборки по слоям, мы обобщим формулу (39.42) для распределения с МД, приняв во внимание различие в стоимости проведения выбора в разных слоях. Мы специально отложили это обобщение, поскольку теперь мы можем вывести его как специальный случай общего результата о распределении с минимальной дисперсией при фиксированной полной стоимости, которой, как мы увидим, полезен и в других отношениях.

Предположим, что в данной задаче выбора дисперсия используемой оценки имеет вид

$$V = v_0 + \sum_{l=1}^{k} \frac{v_l}{w_l}, \tag{39.50}$$

где v_0 и v_l являются функциями только от величин, определяемых совокупностью (но не выбором), а w_l не являются функциями величин v. В общем случае полная стоимость проведения выборочного обследования представима в виде

$$C = c_0 + \sum_{l=1}^{k} w_l c_l, \tag{35.51}$$

где c_0 — «стоимость накладных расходов», а c_l — стоимость внутри l-го класса. Используя неравенство Коши (см. **2.7**), запишем следующее соотношение:

$$(V - v_0)(C - c_0) = \sum_l \frac{v_l}{w_l} \sum_l w_l c_l \geqslant \left\{ \sum_l (v_l c_l)^{1/2} \right\}^2. \tag{39.52}$$

Равенство в (39.52) достигается тогда и только тогда, когда выполняется условие

$$\left(\frac{v_l}{w_l}\right) \Big/ (w_l c_l) = \text{const} \quad \text{для всех} \quad l.$$

Крайний правый член в (39.52) не зависит от w_l, а значит, выбирая w_l удовлетворяющим этому условию, которое мы перепишем в виде

$$w_l^2 \propto v_l / c_l \quad \text{для всех} \quad l, \tag{39.53}$$

мы получим минимально возможное значение $(V - v_0)(C - c_0)$, т. е. минимальное V при фиксированном C (минимальное C при фиксированном V).

39.21 В предшествующем изложении дисперсия задавалась соотношением (39.39), а значит, стоимость проведения выборочного обследования равна

$$C = c_0 + \sum_{l=1}^{k} n_l c_l, \tag{39.54}$$

где c_0 — стоимость накладных расходов, а c_l — стоимость наблюдений в l-м слое. Сравнивая (39.39) с (39.50), убеждаемся, что здесь

$$v_0 = \frac{1}{N^2} \sum_l N_l \sigma_l^2, \qquad v_l = \frac{N_l^2 \sigma_l^2}{N^2}, \qquad w_l = n_l,$$

а значит, (39.53) дает следующее размещение с МД при фиксированном C:

$$n_l \propto \frac{N_l \sigma_l}{N c_l^{1/2}}. \tag{39.55}$$

Теперь доли выбранных элементов n_l/N_l пропорциональны квадратному корню из дисперсии слоя, деленному на квадратный корень из стоимости одного наблюдения в этом слое. Наша исходная формула распределения с МД (39.42), не учитывающая стоимости, является, разумеется, частным случаем (39.52), когда все c_l равны.

Кокэн и Хан (1967) дают численный метод минимизации C, когда исследуются несколько переменных, каждая из которых оценивается с предписанной максимальной дисперсией.

Формирование слоев

39.22 Несомненно, читатель заметил, что в нашем изложении, начиная с **39.15** и далее, предполагалось, что заранее заданы k фиксированных слоев, и единственная задача состоит в распределении выборки по этим слоям. Но на каком-то этапе должны быть сформированы сами эти слои. Как это сделать лучше, чтобы в конечном счете минимизировать дисперсию оценки? Рассмотрим сначала этот вопрос при фиксированном k, а потом изучим влияние переменного k на ответ.

С учетом выводов **39.19** при формировании слоев нужно максимизировать разницу между средними слоев. Таким образом, если бы мы знали распределение переменной y совокупности, то для образования k слоев нам нужно было бы выбрать $k-1$ точек разбиения ее области изменения. Как же выбирать эти точки? Этот вопрос носит чисто теоретический характер в том смысле, что на практике истинное распределение y никогда не известно. Однако мы можем ориентироваться по предыдущим

наблюдениям или можем знать значения переменной сильно коррелированной с y, так что ответ на этот вопрос имеет и практический интерес. Основные результаты получены Далэниусом (1950).

39.23 Рассмотрим сначала случай, когда применяется ВРД. Не учитывая констант n, N и пренебрегая разницей между N_l и $N_l - 1$, перепишем дисперсию (**39.45**) в виде

$$\mathrm{D}(\hat{\mu}_{\text{ВРД}}) \propto V = \sum_{l=1}^{k} \int_{c_{l-1}}^{c_l} (y - \mu_l)^2 f(y)\, dy, \tag{39.56}$$

где $f(y)$ — задает распределение y в совокупности, $a \leqslant y \leqslant b$ и $a = c_0 < c_1 < c_2 < \ldots < c_{k-1} < c_k = b$, так что $c_1, \ldots, c_{k-1}$ — точки разбиения области значений y, определяющие границы слоев.

Для нахождения величин c, минимизирующих (39.56), положим

$$0 = \frac{\partial V}{\partial c_l} = \{(c_l - \mu_l)^2 f(c_l) - (c_l - \mu_{l+1})^2 f(c_l)\}, \qquad l = 1, \ldots, k-1,$$

так что, если $f(c_l) \neq 0$, то решением будет

$$(c_l - \mu_l)^2 = (c_l - \mu_{l+1})^2,$$

а поскольку $\mu_l < c_l < \mu_{l+1}$, отсюда следует, что

$$c_l = \frac{1}{2}(\mu_l + \mu_{l+1}). \tag{39.57}$$

Таким образом, точки разбиения нужно выбирать так, чтобы они были равны полусумме средних значений тех двух слоев, границей которых они являются. При заданных $f(y)$ и k нетрудно провести соответствующие вычисления.

39.24 С другой стороны, если после формирования слоев мы хотим применять МД-размещение объемов выборок в слоях, то точки разбиения нужно выбирать так, чтобы минимизировать (39.43). Подставляя (39.42) в (39.39), читатель легко убедится, что второй член в фигурных скобках в (39.43) играет роль только в том случае, если доли выбранных элементов n_l/N_l в слоях не являются малыми. Мы будем считать, что эти доли малы и пренебрегать ими. Поэтому без учета констант n и N, соотношение (39.43) перепишется в виде

$$\mathrm{D}(\hat{\mu}_{\text{МД}}) \propto D^2 = \left[\sum_{l=1}^{k} \left\{ \int_{c_{l-1}}^{c_l} f(y)\, dy \cdot \int_{c_{l-1}}^{c_l} (y - \mu_l)^2 f(y)\, dy \right\}^{1/2}\right]^2. \tag{39.58}$$

Поскольку нам нужно минимизировать D, мы положим $(l = 1, \ldots, k-1)$

$$0 = \frac{\partial D}{\partial c_l} = \frac{f(c_l)\int\limits_{c_{l-1}}^{c_l}(y-\mu_l)^2 f(y)\,dy + \int\limits_{c_{l-1}}^{c_l} f(y)\,dy\cdot(c_l-\mu_l)^2 f(c_l)}{2\left\{\int\limits_{c_{l-1}}^{c_l} f(y)\,dy\cdot\int\limits_{c_{l-1}}^{c_l}(y-\mu_l)^2 f(y)\,dy\right\}^{1/2}} -$$

$$- \frac{f(c_l)\int\limits_{c_l}^{c_{l+1}}(y-\mu_{l+1})^2 f(y)\,dy + \int\limits_{c_l}^{c_{l+1}} f(y)\,dy\cdot(c_l-\mu_{l+1})^2 f(c_l)}{2\left\{\int\limits_{c_l}^{c_{l+1}} f(y)\,dy\cdot\int\limits_{c_l}^{c_{l+1}}(y-\mu_{l+1})^2 f(y)\,dy\right\}^{1/2}}.$$

Если $f(c_l) \neq 0$, то после сокращения на множитель N_l и возвращения к исходным обозначениям это сведется к

$$\frac{\sigma_l^2 + (c_l-\mu_l)^2}{\sigma_l} = \frac{\sigma_{l+1}^2 + (c_l-\mu_{l+1})^2}{\sigma_{l+1}}. \qquad (39.59)$$

Если σ_l^2 одно и то же для всех слоев, то (39.59), как и должно быть, сводится к (39.57). Однако в общем случае удовлетворить этим условиям не так просто, как условиям (39.57), поскольку, как и следовало ожидать, сюда входят не только средние, но и дисперсии слоев. Поэтому мы будем искать приближение для (39.59).

39.25 Если k велико, то можно предположить, что $f(y)$ постоянно внутри каждого слоя и равно, скажем, f_l. Тогда

$$N_l/N = \int\limits_{c_{l-1}}^{c_l} f(y)\,dy \doteqdot f_l \int\limits_{c_{l-1}}^{c_l} dy = f_l(c_l - c_{l-1}),$$

и

$$\sigma_l^2 \doteqdot \frac{1}{12}(c_l - c_{l-1})^2,$$

т. е. приблизительно равно дисперсии равномерного распределения. Рассуждая точно так же, как в **39.24**, получим, что для выражения, которое нужно минимизировать в (39.43), справедливо

$$\frac{1}{N}\sum_l N_l\sigma_l \doteqdot \frac{1}{\sqrt{12}}\sum_l f_l(c_l - c_{l-1})^2 = \frac{1}{\sqrt{12}}\sum_l \{f_l^{1/2}(c_l - c_{l-1})\}^2. \qquad (39.60)$$

Если определить теперь преобразование

$$z(y) = \int\limits_a^y \{f(t)\}^{1/2}\,dt, \qquad (39.61)$$

то (39.60) можно переписать в виде

$$\frac{1}{N}\sum_l N_l\sigma_l \doteqdot \frac{1}{\sqrt{12}}\sum_l \{z(c_l) - z(c_{l-1})\}^2.$$

Таким образом, нужно минимизировать сумму квадратов величин $z(c_l) - z(c_{l-1})$ при условии, что их сумма фиксирована и равна $z(b) - z(a)$. Очевидно, что минимум достигается, когда все $z(c_l) - z(c_{l-1})$ равны. А значит, нужно применить преобразование (39.61) и определить точки разбиения как корни уравнений

$$z(y) = \frac{r}{k}\{z(b) - z(a)\}, \qquad r = 1, \ldots, k-1. \qquad (39.62)$$

Далэниус и Ходжес (1957, 1959), которые получили это приближение, указали, как проводить вычисления и получили более точные приближения. Другое приближение, полученное Экманом (1959), можно получить, если записать просто

$$\sum_l N_l\sigma_l \doteqdot \frac{1}{\sqrt{12}}\sum_l N_l(c_l - c_{l-1}). \qquad (39.63)$$

Из предшествующих рассуждений следует, что $\sum_l \{N_l(c_l - c_{l-1})\}^{1/2} = \text{const}$, а значит, минимум (39.63) достигается, если положить $\{N_l(c_l - c_{l-1})\}^{1/2}$ равными, или, что эквивалентно,

$$N_l(c_l - c_{l-1}) = \text{const}. \qquad (39.64)$$

Кокрэн (1961) приводит численные примеры применения (39.62) и (39.64) для малых k (равных 2, 3 и 4) и показывает, что они хорошо согласуются при применении к восьми распространенным асимметричным распределениям. Им рассматриваются также менее удовлетворительные приближения для (39.59).

Следует отметить, что для обоих приближенных решений (39.62) и (39.64) значения $N_l\sigma_l$ постоянны для всех слоев. Но в этом случае из формулы (39.42) для распределения с МД следует, что n_l должны быть одинаковыми для всех слоев. Таким образом, можно ожидать, что для расслоения с одинаковым объемом выборки в каждом слое получится дисперсия, близкая к минимально возможной, если точки деления между слоями выбрать так, чтобы минимизировать (39.39). В упражнении 39.18 читателю предлагается получить соответствующий

результат. Кокрэн (1961) показал, что при этом ответ мало отличается от случая, когда используется (39.62) или (39.64).

С. Гхош (1963b) обобщил решение задачи о выборе границы слоев на случай, когда этот выбор основывается на значении двух (коррелированных) величин.

В упражнениях 40.4—6 рассматривается *случайное* образование слоев.

39.26 Если мы проводим выбор, применяя всегда ВРД или распределение с МД, а объемы слоев велики, то (39.49) убеждает нас в том, что мы никогда не получим увеличения дисперсии, разбивая слои на подслои. Так что логически мы приходим к выводу, что выборка из n наблюдений должна производиться из $k = n$ слоев по одному наблюдению в каждом слое, или, если мы хотим использовать (39.40) для оценки дисперсии (что требует как минимум два наблюдения в каждом слое), то из $k = [n/2]$ слоев. Если n мало, а мы достаточно точно знаем изучаемое распределение, то так поступать целесообразно, но в противном случае необходимые для этого вычисления вряд ли будут оправданы. Имеется много эмпирических доказательств, что при увеличении k минимально возможная дисперсия уменьшается все более и более медленно, так что часто 2, 3 или 4 являются почти наилучшими значениями. Это происходит из-за того, что наше знание изучаемого распределения довольно неточно. Кокрэн (1963) и Далэниус (1953) приводят численные примеры. Детальное эмпирическое изучение влияния формирования слоев и распределения выборки по слоям на дисперсию оценки проведено Гессом, Сэти и Балакришнаном (1966).

39.27 Отметим наконец, что влияние любого расслоения на выборочную дисперсию всегда можно оценить после выбора. Нужно только использовать (39.40), чтобы оценить дисперсию величины $\hat{\mu}$ и сравнить это с оценкой дисперсии для m_R, приведенной в удобном виде в **39.18**. Из этой формулы видно, что задача сводится к получению оценки для

$$B = \sum_l N_l(\mu_l - \mu)^2 = \sum_l N_l\mu_l^2 - \frac{1}{N}\left(\sum_l N_l\mu_l\right)^2 =$$
$$= \sum_l N_l\mu_l^2 - \frac{1}{N}\left\{\sum_l N_l^2\mu_l^2 + \sum_{l\neq r}\sum N_lN_r\mu_l\mu_r\right\}. \qquad (39.65)$$

Поскольку

$$\mathbf{M}(m_l^2) - \mu_l^2 = \mathbf{D}(m_l) = \frac{\sigma_l^2}{n_l}\left(1 - \frac{n_l}{N_l}\right),$$

то

$$\mathbf{M}\left\{m_l^2 - \frac{s_l^2}{n_l}\left(1 - \frac{n_l}{N_l}\right)\right\} = \mu_l^2,$$

а значит, согласно (39.65) несмещенной оценкой для B является

$$\hat{B} = \sum_l\left(N_l - \frac{N_l^2}{N}\right)\left\{m_l^2 - \frac{s_l^2}{n_l}\left(1 - \frac{n_l}{N_l}\right)\right\} - \frac{1}{N}\sum_{l\neq r}\sum N_lN_rm_lm_r =$$
$$= \sum_l N_lm_l^2 - \frac{1}{N}\left(\sum_l N_lm_l\right)^2 - \frac{1}{N}\sum_l (N - N_l)(N - n_l)\frac{s_l^2}{n_l}. \qquad (39.66)$$

Таким образом, согласно **39.18** оценкой дисперсии при нерасслоенном случайном выборе является

$$\hat{\mathbf{D}}(m_R) = \frac{N-n}{nN(N-1)}\left\{\sum_l (N_l - 1)s_l^2 + \hat{B}\right\}, \qquad (39.67)$$

где $\hat{B}$ определено в (39.66).

Планирование выборки: группировка (clustering)

39.28 К принципу расслоения мы пришли в **39.14** при обсуждении влияния изменения вероятностей π_{ij} при фиксированных π_i, равных n/N. Мы видели, что по сравнению со значениями, которые получаются при выборе с равными вероятностями, некоторые π_{ij} выгодно несколько увеличить (и для компенсации уменьшить другие). Естественно возникает вопрос, а не стоит ли продолжить это увеличение и сделать некоторые π_{ij} настолько большими, насколько это возможно. Из определения в **39.6** следует, что всегда $\pi_{ij} \leqslant \pi_i, \pi_j$ так, что если $\pi_i = n/N$ для всех i, то $\pi_{ij} \leqslant n/N$.

Предположим теперь, что мы разбили N элементов совокупности на N_1 групп, каждая из которых содержит N_2 элементов (так что $N_1N_2 = N$), и что для любой пары i, j из любой заданной группы справедливо $\pi_{ij} = \pi_i = \pi_j = n/N$. Внутри каждой группы имеется $N_2(N_2 - 1)$ пар, а значит, всего имеется $N_1N_2(N_2 - 1) = N(N_2 - 1)$ пар i, j, для которых так увеличиваются π_{ij}. Из (39.16) следует, что сумма всех $N(N-1)$ вероятностей π_{ij} в совокупности должна равняться $n(n-1)$. Так что между оставшимися $N(N-1) - N(N_2-1) = N(N-N_2)$ парами i, j, для которых π_{ij} не увеличивается до n/N, нужно так распределить значения π_{ij}, чтобы их сумма равнялась $n(n-1) - N(N_2-1)\cdot\frac{n}{N} = n(n-N_2)$. Если считать все эти π_{ij} одинаковыми, то каждое будет равняться $\frac{n(n-N_2)}{N(N-N_2)}$. Предположим,

что n мы выбрали кратным N_2, скажем, $n = n_1 N_2$. Тогда $\pi_{ij} = \frac{n_1(n_1-1)}{N_1(N_1-1)}$. Из **39.6** мы видим, что это совпадает со значениями π_{ij}, которые получились бы, если бы мы проводили выборку объема n_1 из совокупности, состоящей из N_1 элементов, применяя случайный выбор с равными вероятностями. Если теперь учесть, что из равенства π_{ij} и π_i внутри групп следует, что все группы либо выбираются *целиком*, либо нет, то нетрудно понять, что описанное планирование выборки заключается в следующем. Надо разбить совокупность на N_1 равных групп (cluster) и случайным образом с равными вероятностями выбрать n_1 из них.

39.29 Частным случаем группового выбора является *систематический выбор*, при котором элементы совокупности представлены в виде единой последовательности (исходя из физических соображений или просто в порядке записи), состоящей из $N = N_1 N_2$ элементов. Среди первых N_1 из них выбирается случайным образом с равными вероятностями выбора один элемент. Если при этом извлечен p-й элемент, то систематическая выборка включает элементы с номерами p, $p+N_1$, $p+2N_1$, ..., $p+(N_2-1)N_1$. При этом максимально возможное число выборок равно N_1, причем объем каждой выборки равен N_2. Формально имеется полное совпадение между систематическим выбором и групповым выбором, описанным в **39.28**, но в литературе они различаются благодаря нескольким специфическим чертам.

Во-первых, тот факт, что систематический выбор порождает группы, не смежные с физической точки зрения, резко контрастирует с другими случаями группировки, когда, как показывает само название, используется множество соседних элементов из физической совокупности. Во-вторых систематический выбор в порядке записи, которая реально проводится в случайном порядке, часто используется как более простая замена выбора с равными вероятностями. В этом случае метод называют иногда *квазислучайным выбором*. Но самый важный отличительный фактор состоит в том, что при систематическом выборе обычно извлекается только одна выборка из возможного числа N_1, т. е. $n_1 = 1$. Это сразу делает невозможным оценку значения дисперсии, если не имеется какой-нибудь дополнительной информации. Кажется более простым наложить условие $n_1 \geqslant 2$, чтобы была возможна обоснованная оценка значения дисперсии.

Обсуждение и библиография по систематическому выбору приведены у Кокрэна (1963) и Иэйтса (1960), которые внесли существенный вклад в эту теорию.

39.30 Как влияет групповой выбор на $\mathrm{D}(\hat{\mu})$ в (39.23)? Вклад от пар i, j, принадлежащих к одной группе, теперь отрицателен, поскольку для этих пар $\pi_i\pi_j - \pi_{ij} = -\frac{n}{N}\left(1-\frac{n}{N}\right)$. Вклад от пар, принадлежащих разным группам, положителен, поскольку для таких пар $\pi_i\pi_j - \pi_{ij} = \frac{n_1(N_1-n_1)}{N_1^2(N_1-1)}$. Теперь можно применить аргументы **39.14**. Если пары элементов с большими значениями $|y_i - y_j|$ отнести к одному классу, где π_{ij} максимально возможные, то следует ожидать уменьшения (39.23) и возможно более существенного, чем в **39.14**, поскольку большие значения π_{ij} дают теперь отрицательный вклад, а не вносят нулевой вклад, как это было в **39.14**.

Таким образом, мы приходим к общему принципу разбиения на группы, который полностью противоположен принципу расслоения, рассмотренному в **39.14**. Нужно разбивать совокупность на группы внутренне *неоднородные*, чтобы уменьшить дисперсию группового выбора по сравнению с дисперсией выбора с равными вероятностями.

Как и в случае расслоения, оставим теперь в стороне общие рассуждения и перейдем к конкретному изложению деталей. Но предварительно сделаем два общих замечания.

Основное различие между расслоенным и групповым выбором состоит в том, что в каждом слое производится выбор, в то время как группы сами являются предметами выбора. Именно на этом факте основывается противоположность принципов для двух методов. При расслоенной выборке значение дисперсии зависит только от различия элементов внутри слоев и слои формируются так, чтобы минимизировать это различие. При групповом выборе дисперсия зависит только от различия между группами, поскольку группы извлекаются целиком, поэтому группы формируются, чтобы уменьшить именно это различие.

Второе замечание, которое мы считаем необходимым здесь привести, хотя об этом явно говорилось в **39.14** и **39.28**, касается того, что как при расслоенной выборке, использующей ВРД, так и при групповом выборе с равными объемами полная вероятность извлечения π_i не меняется и равна n/N, как и для случайного выбора с равными вероятностями. Эти методы основаны только на изменении совместных вероятностей извлечения π_{ij}. Конечно при расслоенном выборе с переменной долей выбранных элементов сами π_i тоже меняются. Это же верно и для группового выбора с группами неравного объема. Ниже мы рассмотрим эту более общую ситуацию. Кроме того, нам будет удобно провести одновременно обобщение и в другом направлении.

Многошаговый выбор

39.31 Групповой выбор предполагает предварительную группировку элементов совокупности, а затем извлечение некоторых групп. Естественно рассмотреть более общую ситуацию, когда на следующих шагах проводится выбор внутри извлеченных групп. Возвращаясь, насколько это возможно, к обозначениям **39.28**, дадим четкую формулировку этой более общей ситуации.

Совокупность из N элементов разбивается на N_1 *множеств первого порядка* (которые раньше назывались группами). Множество первого порядка с номером i содержит N_{i2} *множеств второго порядка*, j-е из которых содержит N_{ij3} *множеств третьего порядка*. Этот иерархический процесс можно продолжить и дальше, но мы не будем рассматривать более трех шагов, считая, что множества третьего шага совпадают с элементами совокупности. Иногда для наших целей достаточно будет ограничиться двумя шагами.

При трех шагах имеем

$$N = \sum_{i=1}^{N_1} \sum_{j=1}^{N_{i2}} N_{ij3}.$$

На первом шаге выбирается n_1 множеств первого порядка (из общего числа N_1). Метод выбора пока не уточняется. Внутри i-го выбранного множества выбирается n_{i2} множеств второго порядка для j-го, из которых проводится выбор n_{ij3} множеств третьего порядка, так что объем выборки равен *)

$$n = \sum_{i=1}^{n_1} \sum_{j=1}^{n_{i2}} n_{ij3},$$

Предполагается, что выбор на любом шаге может быть с неравными вероятностями, что выбор на любом шаге не зависит от выбора на других шагах и что выбор внутри любого множества на данном шаге не зависит от выбора внутри других множеств на этом же шаге.

39.32 Сначала нам нужно определить вид несмещенной оценки. Здесь применима общая теория выбора с неравными вероятностями из **39.6—8** и, в частности, (39.20) задает несмещенную оценку $\hat{\mu}$, а выражения (39.23) и (39.24) остаются верными для значения дисперсии величины $\hat{\mu}$ и для ее оценки. При этом π_i и π_{ij} в этих выражениях должны соответствовать теперь полным вероятностям извлечения, учитывающим все шаги выбора. Переобозначим теперь значения y_i элементов совокупности через y_{ijk}; при этом каждый индекс соответствует множеством данного шага. Так, y_{846} является значением, соответствующим элементу, принадлежащему восьмому множеству первого шага, четвертому (внутри этого восьмого) множеству второго шага и шестому множеству третьего шага (которое принадлежит упомянутому четвертому). В этих обозначениях несмещенная оценка (39.20) принимает вид

*) Получающийся при этом объем выборки n является случайной величиной. (*Прим. перев.*).

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{n_1} \sum_{j=1}^{n_{i2}} \sum_{k=1}^{n_{ij3}} \frac{y_{ijk}}{\pi_{(ijk)}}, \tag{39.68}$$

где $\pi_{(ijk)}$ — теперь *) полная вероятность извлечения заданного значения y_{ijk}.

Предположим, например, что выбор на каждом шаге проводится с равными вероятностями. Тогда

$$\pi_{(ijk)} = \frac{n_1}{N_1} \frac{n_{i2}}{N_{i2}} \frac{n_{ij3}}{N_{ij3}},$$

и (39.68) сводится к

$$\hat{\mu} = \frac{N_1}{Nn_1} \sum_{i=1}^{n_1} \frac{N_{i2}}{n_{i2}} \sum_{j=1}^{n_{i2}} \frac{N_{ij3}}{n_{ij3}} \sum_{k=1}^{n_{ij3}} y_{ijk}. \tag{39.69}$$

Если теперь $N_{i2} = N_2$ для всех i, а $N_{ij3} = N_3$ для всех i, j и, кроме того, аналогично $n_{i2} = n_2$, а $n_{ij3} = n_3$, то (39.69) сводится к

$$\hat{\mu} = \frac{N_1 N_2 N_3}{N n_1 n_2 n_3} \sum_i \sum_j \sum_k y_{ijk} = m, \tag{39.70}$$

т. е. к полному выборочному среднему, поскольку в этом случае $N = N_1 N_2 N_3$ и $n = n_1 n_2 n_3$. Соотношение (39.70) интуитивно понятно в силу полной симметрии.

39.33 Аналогично, если рассматривать двухшаговый выбор, то в формуле (39.68) нужно отбросить индекс k и суммирование по нему. Оценкой для случая равных вероятностей будет снова (39.69) с заменой $(N_{ij3}/n_{ij3}) \sum_k y_{ijk}$ на y_{ij}. Сведение к m для симметричного случая аналогично (39.70). Формально любую формулу для двухшагового выбора можно получить из формулы для трехшагового, положив $N_{ij3} = n_{ij3} = 1$ или $N_1 = n_1 = 1$. Это приводит к тому, что один индекс, а значит, и суммирование по нему становится ненужным.

39.34 Так же как при изложении расслоенного выбора, нам будет удобно провести прямое вычисление значения дисперсии нашей оценки при многошаговом выборе, а не выводить соот-

*) Скобки в индексе употребляются для того, чтобы не перепутать с используемыми ранее совместными вероятностями π_{ij}.

ветствующее выражение из общей формулы для случая неравных вероятностей. Это поможет, например, избежать употребление символов типа $\pi_{(ijk)(tuv)}$ для обозначения совместного распределения выбора двух значений y_{ijk} и y_{tuv}. Мы подробно рассмотрим два частных случая, первый из которых соответствует выбору с равными вероятностями на каждом шаге.

39.35 Рассмотрим дисперсию величины $\hat{\mu}$ в (39.69). Очевидно, что выбор на каждом шаге вносит свой вклад в значение дисперсии величины $\hat{\mu}$. Поскольку выбор при различных шагах независим, то будем проводить вычисления, используя последовательность условных дисперсий. Сначала рассмотрим дисперсию на последнем шаге при условии, что уже проведен выбор на всех предыдущих шагах. Затем будем считать переменным предпоследний шаг и т. д. вплоть до первого шага. Например, в случае трех шагов, вычислим сначала дисперсию на третьем шаге, считая, что выбор на первых двух шагах уже произведен и привел к некоторому (фиксированному) результату. Затем рассмотрим зависимость от изменения выбора на втором шаге, а потом позволим изменяться выбору на первом шаге. Символически это можно записать в виде

$$\mathsf{D}(\hat{\mu}) = \mathsf{M}(\hat{\mu} - \mu)^2 = \underset{1}{\mathsf{M}}[\underset{2}{\mathsf{M}}\{\underset{3}{\mathsf{M}}(\hat{\mu} - \mu)^2\}]. \qquad (39.71)$$

Для вычисления (39.71) мы используем общий результат о дисперсии случайной величины, частные случаи которого мы уже использовали в предыдущих главах. Докажем теперь этот результат в общем виде.

39.36 Рассмотрим случайную величину x, и пусть c — условие, осуществление которого влияет на распределение величины x. По теореме умножения вероятностей $P(x) = P_1(x|c)P_2(c)$, где P, P_1 и P_2 — вероятности, определяемые соответствующими аргументами. Тогда равенство

$$\mathsf{M}(x) = \underset{c}{\mathsf{M}}\{\mathsf{M}(x|c)\} \qquad (39.72)$$

выражает символически тот факт, что для нахождения математического ожидания величины x можно наложить сначала некоторое условие, вычислить математическое ожидание при этом условии, а затем устранить влияние условия, взяв математическое ожидание от самого условного математического ожидания.

Согласно (39.72) выражение для дисперсии получается на основе следующего тождества:

$$\underset{c}{\mathsf{M}}[\mathsf{M}(x^2|c) - \{\mathsf{M}(x|c)\}^2] + \underset{c}{\mathsf{M}}[\{\mathsf{M}(x|c)\}^2 - (\underset{c}{\mathsf{M}}\{\mathsf{M}(x|c)\})^2] \equiv$$
$$\equiv \underset{c}{\mathsf{M}}[\mathsf{M}(x^2|c) - (\underset{c}{\mathsf{M}}\{\mathsf{M}(x|c)\})^2] \equiv \mathsf{M}(x^2) - \{\mathsf{M}(x)\}^2.$$

По определению первый член в левой части есть математическое ожидание от условной дисперсии величины x при заданном c, а второй член в левой части есть дисперсия от условного математического ожидания x при условии c. Таким образом, символически имеем

$$\mathsf{D}(x) = \underset{c}{\mathsf{M}}\{\mathsf{D}(x|c)\} + \underset{c}{\mathsf{D}}\{\mathsf{M}(x|c)\}, \qquad (39.73)$$

т. е. безусловная дисперсия равна среднему от условной дисперсии плюс дисперсия от условного среднего. Заметим, что если $\mathsf{M}(x|c)$ не зависит от c, то второй член в (39.73) обращается в нуль и $\mathsf{D}(x)$ превращается просто в среднее от условной дисперсии.

Этот результат весьма общий. Например, на самом деле он использовался в **17.35** при выводе метода Рао — Блэкуэлла улучшения оценок с помощью достаточных статистик.

39.37 Используя (39.73), видим теперь, что (39.71) можно переписать в виде

$$\mathsf{D}(\hat{\mu}) = \underset{12}{\mathsf{M}}\{\underset{3}{\mathsf{D}}(\hat{\mu})\} + \underset{12}{\mathsf{D}}\{\underset{3}{\mathsf{M}}(\hat{\mu})\}, \qquad (39.74)$$

где символ «1 2» означает, что фиксируется выбор на первом и втором шагах. В соотношении (39.74) подсчитываемая дисперсия разбита на две части, каждую из которых мы вычислим отдельно.

Рассмотрим сначала значение $\underset{3}{\mathsf{D}}(\hat{\mu})$. На третьем (в более общей ситуации на последнем) шаге выбора из каждого из $\sum_{i=1}^{n_1} n_{i2}$ выбранных множеств второго шага производится выбор с равными вероятностями, при котором извлекается n_{ij3} элементов из возможного числа N_{ij3}. По существу это расслоенная выборка, где каждое множество второго порядка играет роль слоя. Определяя

$$m_{ij} = \frac{1}{n_{ij3}} \sum_{k=1}^{n_{ij3}} y_{ijk},$$

мы знаем, что

$$\mathsf{D}(m_{ij}) = \frac{\sigma_{ij}^2}{n_{ij3}}\left(1 - \frac{n_{ij3}}{N_{ij3}}\right),$$

где σ_{ij}^2 — дисперсия совокупности элементов того множества второго порядка, из которого была извлечена эта выборка

объема n_{ij3}. Из (39.69) следует, что

$$\underset{3}{\mathsf{D}}(\hat{\mu}) = \left(\frac{N_1}{Nn_1}\right)^2 \sum_{i=1}^{n_1} \left(\frac{N_{i2}}{n_{i2}}\right)^2 \sum_{j=1}^{n_{i2}} N_{ij3}^2 \frac{\sigma_{ij}^2}{n_{ij3}} \left(1 - \frac{n_{ij3}}{N_{ij3}}\right). \quad (39.75)$$

Теперь нам нужно вычислить математическое ожидание от (39.75), когда может меняться выбор на предыдущих шагах. Меняя только выбор на втором шаге, получим

$$\underset{2\,3}{\mathsf{MD}}(\hat{\mu}) = \left(\frac{N_1}{Nn_1}\right)^2 \sum_{i=1}^{n_1} \frac{N_{i2}^2}{n_{i2}} \underset{2}{\mathsf{M}} \left[\frac{1}{n_{i2}} \sum_{j=1}^{n_{i2}} N_{ij3}^2 \frac{\sigma_{ij}^2}{n_{ij3}} \left(1 - \frac{n_{ij3}}{N_{ij3}}\right)\right].$$

Но поскольку выражение в квадратных скобках есть просто выборочное среднее от соответствующих значений, то его математическое ожидание равно просто среднему по совокупности от соответствующей функции, т. е.

$$\underset{2\,3}{\mathsf{MD}}(\hat{\mu}) = \left(\frac{N_1}{Nn_1}\right)^2 \sum_{i=1}^{n_1} \frac{N_{i2}^2}{n_{i2}} \cdot \frac{1}{N_{i2}} \sum_{j=1}^{N_{i2}} N_{ij3}^2 \frac{\sigma_{ij}^2}{n_{ij3}} \left(1 - \frac{n_{ij3}}{N_{ij3}}\right).$$

Аналогично, позволив меняться выбору на первом шаге, получим

$$\underset{1\,2\,3}{\mathsf{MMD}}(\hat{\mu}) = \frac{N_1^2}{N^2 n_1} \underset{1}{\mathsf{M}} \left[\frac{1}{n_1} \sum_{i=1}^{n_1} \frac{N_{i2}}{n_{i2}} \sum_{j=1}^{N_{i2}} N_{ij3}^2 \frac{\sigma_{ij}^2}{n_{ij3}} \left(1 - \frac{n_{ij3}}{N_{ij3}}\right)\right] =$$

$$= \frac{N_1}{N^2 n_1} \sum_{i=1}^{N_1} \frac{N_{i2}}{n_{i2}} \sum_{j=1}^{N_{i2}} N_{ij3}^2 \frac{\sigma_{ij}^2}{n_{ij3}} \left(1 - \frac{n_{ij3}}{N_{ij3}}\right). \quad (38.76)$$

Таким образом, мы вычислили первый член в правой части (39.74).

39.38 Для вычисления второго члена в (39.74) найдем сначала значение $\underset{3}{\mathsf{M}}(\hat{\mu})$. Согласно (39.69) оно равно

$$\underset{3}{\mathsf{M}}(\hat{\mu}) = \frac{N_1}{Nn_1} \sum_{i=1}^{n_1} \frac{N_{i2}}{n_{i2}} \sum_{j=1}^{n_{i2}} N_{ij3} \underset{3}{\mathsf{M}}(m_{ij}) = \frac{N_1}{Nn_1} \sum_{i=1}^{n_1} \frac{N_{i2}}{n_{i2}} \sum_{j=1}^{n_{i2}} N_{ij3} \mu_{ij}, \quad (39.77)$$

где μ_{ij} — среднее по совокупности, соответствующее выборочному среднему m_{ij}. Обозначим $N_{ij3}\mu_{ij} = T_{ij}$, где T_{ij} суммарное значение игреков в этом множестве с индексами i, j. Применяя теперь снова (39.73) ко второму члену в правой части (39.74), получим

$$\mathsf{D}(\hat{\mu}) = \underset{1\,2}{\mathsf{MM}}\{\underset{3}{\mathsf{D}}(\hat{\mu})\} + \underset{1}{\mathsf{M}}[\underset{2}{\mathsf{D}}\{\underset{3}{\mathsf{M}}(\hat{\mu})\}] + \underset{1}{\mathsf{D}}[\underset{2}{\mathsf{M}}\{\underset{3}{\mathsf{M}}(\hat{\mu})\}]. \quad (39.78)$$

Последний член в правой части (39.78) можно теперь получить из (39.77) поскольку, как и раньше,

$$\underset{2}{\mathsf{M}}\{\underset{3}{\mathsf{M}}(\hat{\mu})\} = \frac{N_1}{Nn_1} \sum_{i=1}^{n_1} N_{i2} \underset{2}{\mathsf{M}} \left[\frac{1}{n_{i2}} \sum_{j=1}^{n_{i2}} T_{ij}\right] = \frac{N_1}{Nn_1} \sum_{i=1}^{n_1} \sum_{j=1}^{N_{i2}} T_{ij},$$

а значит,

$$\underset{1}{\mathsf{D}}[\underset{2}{\mathsf{M}}\{\underset{3}{\mathsf{M}}(\hat{\mu})\}] = \left(\frac{N_1}{N}\right)^2 \underset{1}{\mathsf{D}} \left[\frac{1}{n_1} \sum_{i=1}^{n_1} \left(\sum_{j=1}^{N_{i2}} T_{ij}\right)\right].$$

Но записанная здесь дисперсия является дисперсией выборочного среднего, а значит,

$$\underset{1}{\mathsf{D}}[\underset{2}{\mathsf{M}}\{\underset{3}{\mathsf{M}}(\hat{\mu})\}] = \left(\frac{N_1}{N}\right)^2 \frac{\sigma_{T_i}^2}{n_1} \left(1 - \frac{n_1}{N_1}\right), \quad (39.79)$$

где $\sigma_{T_i}^2$ — дисперсия между суммарными значениями игреков внутри множеств первого порядка, т. е.

$$\sigma_{T_i}^2 = \frac{1}{N_1 - 1} \sum_{i=1}^{N_1} \left(T_i - \frac{\sum_{i=1}^{N_1} T_i}{N_1}\right)^2,$$

где $T_i = \sum_{j=1}^{N_{i2}} T_{ij}$.

Нам осталось вычислить средний член в правой части (39.78). Согласно (39.77)

$$\underset{2}{\mathsf{D}}\{\underset{3}{\mathsf{M}}(\hat{\mu})\} = \left(\frac{N_1}{Nn_1}\right)^2 \sum_{i=1}^{n_1} N_{i2}^2 \underset{2}{\mathsf{D}} \left[\frac{1}{n_{i2}} \sum_{j=1}^{n_{i2}} T_{ij}\right] =$$

$$= \left(\frac{N_1}{Nn_1}\right)^2 \sum_{i=1}^{n_1} N_{i2}^2 \frac{\sigma_{T_{ij}}^2}{n_{i2}} \left(1 - \frac{n_{i2}}{N_{i2}}\right). \quad (39.80)$$

где $\sigma_{T_{ij}}^2$ — дисперсия между суммарными значениями для множеств второго порядка, взятая по i-му множеству первого порядка, т. е.

$$\sigma_{T_{ij}}^2 = \frac{1}{N_{i2} - 1} \sum_{j=1}^{N_{i2}} \left(T_{ij} - \frac{\sum_{j=1}^{N_{i2}} T_{ij}}{N_{i2}}\right)^2.$$

Из (39.80) теперь получаем

$$\underset{1}{\mathsf{M}}[\underset{2}{\mathsf{D}}\{\underset{3}{\mathsf{M}}(\hat{\mu})\}] = \frac{N_1^2}{N^2 n_1}\underset{1}{\mathsf{M}}\left[\frac{1}{n_1}\sum_{i=1}^{n_1} N_{i2}^2 \frac{\sigma_{T_{ij}}^2}{n_{i2}}\left(1-\frac{n_{i2}}{N_{i2}}\right)\right] =$$

$$= \frac{N_1}{N^2 n_1}\sum_{i=1}^{N_1} N_{i2}^2 \frac{\sigma_{T_{ij}}^2}{n_{i2}}\left(1-\frac{n_{i2}}{N_{i2}}\right). \qquad (39.81)$$

39.39 Подставляя теперь (39.76), (39.79) и (39.81) в (39.78), получим окончательно

$$\mathsf{D}(\hat{\mu}) = \left(\frac{N_1}{N}\right)^2 \frac{\sigma_{T_i}^2}{n_1}\left(1-\frac{n_1}{N_1}\right) + \frac{N_1}{N^2 n_1}\sum_{i=1}^{N_1} N_{i2}^2 \frac{\sigma_{T_{ij}}^2}{n_{i2}}\left(1-\frac{n_{i2}}{N_{i2}}\right) +$$

$$+ \frac{N_1}{N^2 n_1}\sum_{i=1}^{N_1}\frac{N_{i2}}{n_{i2}}\sum_{j=1}^{N_{i2}} N_{ij3}^2 \frac{\sigma_{ij}^2}{n_{ij3}}\left(1-\frac{n_{ij3}}{N_{ij3}}\right). \qquad (39.82)$$

Как это было понятно из (39.78) и из интуитивных рассуждений, (39.82) показывает, что вклад в дисперсию оценки происходит на каждом шаге выбора. Рассмотрим частный случай (39.82), когда имеется полная симметрия, т. е.

$N_{i2} = N_2$ для всех i; $N_{ij3} = N_3$ для всех i, j; $N = N_1 N_2 N_3$.

$n_{i2} = n_2$ для всех i; $n_{ij3} = n_3$ для всех i, j; $n = n_1 n_2 n_3$.

Согласно (39.70) оценка $\hat{\mu}$ сводится к полному выборочному среднему. Соотношение (39.82) сводится к

$$\mathsf{D}(\hat{\mu}) = \frac{\left(1-\frac{n_1}{N_1}\right)}{(N_2 N_3)^2}\frac{\sigma_{T_i}^2}{n_1} + \frac{\left(1-\frac{n_2}{N_2}\right)}{N_1 N_3^2 n_1 n_2}\sum_{i=1}^{N_1}\sigma_{T_{ij}}^2 +$$

$$+ \frac{\left(1-\frac{n_3}{N_3}\right)}{N_1 N_2 n_1 n_2 n_3}\sum_{i=1}^{N_1}\sum_{j=1}^{N_{i2}}\sigma_{ij}^2. \qquad (39.83)$$

Если определить теперь

$$\mu_i = T_i/(N_2 N_3),$$

$$\sigma_1^2 = \frac{1}{N_1 - 1}\sum_{i=1}^{N_1}(\mu_i - \mu)^2 \equiv \sigma_{T_i}^2/(N_2 N_3)^2,$$

$$\sigma_2^2 = \frac{1}{N_1(N_2-1)}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}(\mu_{ij} - \mu)^2 \equiv \frac{1}{N_1 N_3^2}\sum_{i=1}^{N_1}\sigma_{T_{ij}}^2,$$

$$\sigma_3^2 = \frac{1}{N_1 N_2 (N_3 - 1)}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sum_{k=1}^{N_3}(y_{ijk} - \mu_{ij})^2 \equiv \frac{1}{N_1 N_2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sigma_{ij}^2,$$

то для рассматриваемого симметричного случая (39.83) можно переписать в виде

$$\mathsf{D}(\hat{\mu}) = \frac{\sigma_1^2}{n_1}\left(1-\frac{n_1}{N_1}\right) + \frac{\sigma_2^2}{n_1 n_2}\left(1-\frac{n_2}{N_2}\right) + \frac{\sigma_3^2}{n_1 n_2 n_3}\left(1-\frac{n_3}{N_3}\right). \qquad (39.84)$$

Из (39.84) понятно, как в симметричном случае получить обобщение на большее число шагов. Общая формула для $p \geqslant 2$ шагов имеет вид

$$\mathsf{D}(\hat{\mu}) = \sum_{r=1}^{p}\frac{\sigma_r^2}{n_1 n_2 \ldots n_r}\left(1-\frac{n_r}{N_r}\right), \qquad (39.85)$$

где

$$\sigma_r^2 = \frac{1}{N_1 N_2 \ldots N_{r-1}(N_r - 1)}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\cdots\sum_{u=1}^{N_r}(y_{ij\ldots tu} - \mu_{ij\ldots t})^2.$$

39.40 Если какая-нибудь доля выбранных элементов n_r/N_r равна единице, то соответствующий член в (39.82) обращается в нуль. В частности, если $n_{ij3}/N_{ij3} = 1$ для всех i, j, то последняя сумма в правой части обращается в нуль, а оставшиеся два члена соответствуют $\mathsf{D}(\hat{\mu})$ для двухшагового выбора. Если, кроме того, $n_{i2} = N_{i2}$ для всех i, то остается только первый член в правой части (39.82), и мы получаем общее выражение для группового выбора с группами, не равными по объему. Аналогично первый член в (39.84) соответствует групповому выбору с равными объемами.

На самом деле нетрудно понять, как следует обобщить (39.82) на случай выбора с большим числом шагов. Для четвертого шага к правой части добавится член

$$\frac{N_1}{N^2 n_1}\sum_{i=1}^{N_1}\frac{N_{i2}}{n_{i2}}\sum_{j=1}^{N_{i2}}\frac{N_{ij3}}{n_{ij3}}\sum_{k=1}^{N_{ij3}} N_{ijk4}^2 \frac{\sigma_{ijk}^2}{n_{ijk4}}\left(1-\frac{n_{ijk4}}{N_{ijk4}}\right), \qquad (39.86)$$

а σ_{ij}^2 в (39.82) нужно будет заменить на $\sigma_{T_{ijk}}^2$, определение которого является естественным обобщением определения $\sigma_{T_i}^2$ и $\sigma_{T_{ij}}^2$.

Однако на практике при многошаговом выборе чрезвычайно редко используют выбор с равными вероятностями. Причина совсем проста. Дисперсия в (39.82) является дисперсией между *суммарными* значениями переменной y из различных множеств. Если объемы этих множеств существенно различаются (т. е. содержат существенно разное число множеств следующего

шага), то это делает дисперсию $\hat{\mu}$ очень большой. Как мы уже видели, этого не происходит в симметричном случае, когда на каждом шаге все множества имеют одинаковый объем и дисперсию можно переопределить как дисперсию между средними значениями. Однако в общем случае мы вынуждены искать иные схемы выбора, чтобы уменьшить величину дисперсии до приемлемого уровня. На самом деле мы достигнем желаемого, если используем выбор с переменными вероятностями. Разумеется, можно использовать произвольное множество вероятностей на каждом шаге, определить $\hat{\mu}$ по (39.68) и вычислить $\mathsf{D}(\hat{\mu})$ из (39.78). Однако в общем случае члены в (39.78) будут более сложными, поскольку они должны отражать изменение вероятностей на каждом шаге.

Вполне общее выражение для $\mathsf{D}(\hat{\mu})$ выводится формально ниже в (39.97—8) для получения оценок дисперсии. Но сначала мы рассмотрим один способ выбора, который важен на практике.

Выбор с вероятностями, пропорциональными объему

39.41 Посмотрев на (39.68), убеждаемся, что если все n_{i2} равны n_2, все n_{ij3} равны n_3, а все $\pi_{(ijk)}$ равны n/N, то оценка $\hat{\mu}$ сводится к выборочному среднему m, как и для симметричного случая с равными вероятностями в (39.70). Чтобы получить все $\pi_{(ijk)}$ равными, потребуем только, чтобы вероятность выбора i-го множества на первом шаге была равна $n_1 A_i^{(1)}/N$, вероятность на втором шаге выбрать j-е подмножество из i-го множества первого шага равнялась бы $n_2 A_{ij}^{(2)}/A_i^{(1)}$, а вероятность, извлечения на третьем шаге y_{ijk}, равнялась бы $n_3/A_{ij}^{(2)}$. Тогда

$$\pi_{(ijk)} = n_1 \frac{A_i^{(1)}}{N} \cdot n_2 \frac{A_{ij}^{(2)}}{A_i^{(1)}} \cdot \frac{n_3}{A_{ij}^{(2)}} = \frac{n}{N}.$$

Внутри каждого множества предпоследнего шага выбор на последнем шаге производится с равными вероятностями, но эти вероятности, вообще говоря, различны для разных множеств предпоследнего шага. В качестве $A_i^{(1)}$ и $A_{ij}^{(2)}$ можно взять *любое* множество чисел, соответствующих некоторой схеме выбора. Описанная схема выбора, для которой $\pi_{(ijk)} = n/N$, называется *равновзвешивающей* (self-weighting), поскольку для нее $\hat{\mu}$ является средним всех выбранных значений (с равными весами).

Просто и удобно положить

$$A_i^{(1)} = \sum_{j=1}^{N_{i2}} N_{ij3}, \qquad A_{ij}^{(2)} = N_{ij3}.$$

Тогда для каждого множества первого шага вероятность извлечения пропорциональна числу элементов, которое оно содержит, аналогично и для второго шага. Выбор на последнем шаге проводится с равными вероятностями. В этом случае мы будем говорить, что извлечение проводится с *пропорциональными объему вероятностями* на всех шагах, кроме последнего, и с равными вероятностями на последнем шаге.

Легко видеть, что при любом числе $p \geqslant 2$ шагов, если выбор проводится с пропорциональными объему вероятностями на всех шагах, за исключением последнего, где используются равные вероятности, то полная вероятность извлечения отдельного элемента равна n/N.

Впервые такой выбор исследовался Хансеном и Гурвицом (1943). Фактически это был первый случай изучения неравных вероятностей выбора.

39.42 Самый простой способ получения описанного в **39.41** равновзвешивающего выбора с пропорциональным объему вероятностями состоит в следующем. Надо выбрать с возвращением n_1 множеств первого порядка, используя при каждом извлечении вероятности $p_i^{(1)} = \sum_{j=1}^{N_{i2}} N_{ij3}/N$. Затем, используя при каждом извлечении вероятности $p_{ij}^{(2)} = N_{ij3} \Big/ \sum_{j=1}^{N_{i2}} N_{ij3}$, выбрать с возвращением по n_2 множеств второго порядка из каждого из n_1 выбранных множеств первого порядка. И наконец, с равными вероятностями $p_{ijk}^{(3)} = 1/N_{ij3}$ выбрать без возвращения по n_3 элементов из каждого из $n_1 n_2$ выбранных множеств второго порядка. Использование выбора с возвращением на первых двух шагах выбора с пропорциональными объему вероятностями позволит нам применить ниже более простую теорию из **39.12**.

Запишем оценку в виде

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} y_{ijk} = \frac{1}{n_1 n_2} \sum_i \sum_j m_{ij}.$$

Дисперсия этой оценки задается соотношением (39.78), и теперь мы вычислим соответствующие члены.

39.43 Поскольку на последнем шаге выбор проводится с равными вероятностями, то, так же как в **39.37**, имеем

$$\underset{3}{\mathsf{D}}(m_{ij}) = \frac{\sigma_{ij}^2}{n_3}\left(1 - \frac{n_3}{N_{ij3}}\right),$$

так что

$$\underset{3}{\mathsf{D}}(\hat{\mu}) = \frac{1}{(n_1 n_2)^2 n_3} \sum_i \sum_j \sigma_{ij}^2 \left(1 - \frac{n_3}{N_{ij3}}\right),$$

а значит,

$$\underset{2}{\mathsf{M}}[\underset{3}{\mathsf{D}}(\hat{\mu})] = \frac{1}{n_1^2 n_2 n_3} \sum_{i=1}^{n_1} \underset{2}{\mathsf{M}}\left[\frac{1}{n_2} \sum_{j=1}^{n_2} \sigma_{ij}^2 \left(1 - \frac{n_3}{N_{ij3}}\right)\right]. \qquad (39.87)$$

На втором шаге из i-го множества, извлеченного на первом шаге, извлекается с возвращением n_2 (из общего числа N_{i2}) множеств с вероятностями $N_{ij3} \Big/ \sum_{j=1}^{N_1} N_{ij3}$ при каждом извлечении. А значит, (39.87) принимает вид

$$\underset{2}{\mathsf{M}}[\underset{3}{\mathsf{D}}(\hat{\mu})] = \frac{1}{n_1^2 n_2 n_3} \sum_{i=1}^{n_1} \sum_{j=1}^{N_{i2}} \left(\frac{N_{ij3}}{\sum N_{ij3}}\right) \sigma_{ij}^2 \left(1 - \frac{n_3}{N_{ij3}}\right).$$

Аналогично

$$\underset{1}{\mathsf{M}}[\underset{2}{\mathsf{M}}\{\underset{3}{\mathsf{D}}(\hat{\mu})\}] = \frac{1}{n_1 n_2 n_3} \sum_{i=1}^{N_1} \left(\frac{\sum_j N_{ij3}}{N}\right) \sum_{j=1}^{N_{i2}} \left(\frac{N_{ij3}}{\sum_j N_{ij3}}\right) \sigma_{ij}^2 \left(1 - \frac{n_3}{N_{ij3}}\right) =$$

$$= \frac{1}{n_1 n_2 n_3 N} \sum_{i=1}^{N_1} \sum_{j=1}^{N_{i2}} N_{ij3} \sigma_{ij}^2 \left(1 - \frac{n_3}{N_{ij3}}\right). \qquad (39.88)$$

Это первый член в правой части (39.78). Далее

$$\underset{3}{\mathsf{M}}(\hat{\mu}) = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \mu_{ij},$$

$$\underset{2}{\mathsf{M}}[\underset{3}{\mathsf{M}}(\hat{\mu})] = \frac{1}{n_1} \sum_{i=1}^{n_1} \underset{2}{\mathsf{M}}\left[\frac{1}{n_2} \sum_{j=1}^{n_2} \mu_{ij}\right] =$$

$$= \frac{1}{n_1} \sum_{i=1}^{n_1} \sum_{j=1}^{N_{i2}} \left(\frac{N_{ij3}}{\sum_j N_{ij3}}\right) \mu_{ij} = \frac{1}{n_1} \sum_{i=1}^{n_1} \mu_i, \qquad (39.89)$$

где μ_i, как и ранее, — среднее значение y, взятое по i-му множеству первого шага. А значит,

$$\underset{1}{\mathsf{D}}[\underset{2}{\mathsf{M}}\{\underset{3}{\mathsf{M}}(\hat{\mu})\}] = \underset{1}{\mathsf{D}}\left(\frac{1}{n_1} \sum_{i=1}^{n_1} \mu_i\right) = \frac{1}{n_1} \underset{1}{\mathsf{D}}(\mu_i) =$$

$$= \frac{1}{n_1} \sum_{i=1}^{N_1} \left(\frac{\sum_{j=1}^{N_{i2}} N_{ij3}}{N}\right) \left(\mu_i - \frac{\sum_{i=1}^{N_1} \sum_{j=1}^{N_{i2}} N_{ij3} \mu_i}{N}\right)^2 =$$

$$= \frac{1}{n_1 N} \sum_{i=1}^{N_1} N_{iT} (\mu_i - \mu)^2, \qquad (39.90)$$

где суммарное число элементов в i-м множестве первого шага обозначено $N_{iT} = \sum_{j=1}^{N_{i2}} N_{ij3}$. (39.90) соответствует последнему члену в правой части (39.78). Согласно (39.89) средний член равен

$$\underset{1}{\mathsf{M}}[\underset{2}{\mathsf{D}}\{\underset{3}{\mathsf{M}}(\hat{\mu})\}] = \underset{1}{\mathsf{M}}\left[\sum_{i=1}^{n_1} \underset{2}{\mathsf{D}}\left\{\frac{1}{n_1 n_2} \sum_{j=1}^{n_2} \mu_{ij}\right\}\right] = \underset{1}{\mathsf{M}}\left[\frac{1}{n_1^2} \sum_{i=1}^{n_1} \frac{1}{n_2} \underset{2}{\mathsf{D}}(\mu_{ij})\right] =$$

$$= \underset{1}{\mathsf{M}}\left[\frac{1}{n_1^2} \sum_{i=1}^{n_1} \frac{1}{n_2} \sum_{j=1}^{N_{i2}} \left(\frac{N_{ij3}}{N_{iT}}\right) \left(\mu_{ij} - \sum_{j=1}^{N_{i2}} \frac{N_{ij3}}{N_{iT}} \mu_{ij}\right)^2\right] =$$

$$= \underset{1}{\mathsf{M}}\left[\frac{1}{n_1^2 n_2} \sum_{i=1}^{n_1} \frac{1}{N_{iT}} \sum_{j=1}^{N_{i2}} N_{ij3} (\mu_{ij} - \mu_i)^2\right] =$$

$$= \frac{1}{n_1 n_2 N} \sum_{i=1}^{N_1} \sum_{j=1}^{N_{i2}} N_{ij3} (\mu_{ij} - \mu_i)^2. \qquad (39.91)$$

Подставляя (39.88), (39.90) и (39.91) в (39.78), получим

$$\mathsf{D}(\hat{\mu}) = \frac{1}{n_1 N} \sum_{i=1}^{N_1} N_{iT} (\mu_i - \mu)^2 + \frac{1}{n_1 n_2 N} \sum_{i=1}^{N_1} \sum_{j=1}^{N_{i2}} N_{ij3} (\mu_{ij} - \mu_i)^2 +$$

$$+ \frac{1}{n_1 n_2 n_3 N} \sum_{i=1}^{N_1} \sum_{j=1}^{N_{i2}} N_{ij3} \sigma_{ij}^2 \left(1 - \frac{n_3}{N_{ij3}}\right). \qquad (39.92)$$

39.44 Следует обратить внимание на то, что (39.92) имеет почти такой же вид, как и формула (39.84) для дисперсии в симметричном случае при выборе с равными вероятностями.

И в самом деле, помимо возможного упрощения, вызванного тем, что на первых двух шагах выбор проводится с возвращением, единственным отличием является то, что каждая компонента дисперсии берется с весом N_{ij3}, как и должно было быть в силу неравного объема множеств. Вводя очевидные обозначения, соотношение (39.92) можно переписать в виде, аналогичном (39.84)

$$\mathsf{D}(\hat{\mu}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1 n_2} + \frac{\sigma_3^2}{n_1 n_2 n_3}, \tag{39.93}$$

где все дисперсии берутся между средними, а не между суммарными значениями, как это было в (39.82). Таким образом, рассмотренный выбор с пропорциональными объему вероятностями приводит к тому, что устраняется влияние переменного размера множеств на всех шагах вплоть до предпоследнего.

Очевидно, что аналогичный результат справедлив для любого числа шагов. Результат для двух шагов получается из (39.92), если считать $n_1 = N_1 = 1$, и сделать соответствующие изменения в обозначениях. В упражнении 39.24 читателю предлагается показать, что в этом случае при очевидном определении символов имеет место

$$\mathsf{D}(\hat{\mu}) = \frac{1}{n_1 N} \sum_{i=1}^{N_1} N_{i2}(\mu_i - \mu)^2 + \frac{1}{n_1 n_2 N} \sum_{i=1}^{N_1} N_{i2}\sigma_i^2 \left(1 - \frac{n_2}{N_{i2}}\right). \tag{39.94}$$

Оценки значения дисперсии в многошаговом выборе

39.45 Как указывалось в конце **39.40**, общая формула для $\mathsf{D}(\hat{\mu})$ при полностью произвольных вероятностях выбора на каждом шаге весьма громоздка и не имеет практического интереса. Однако стоит отметить, что легко получить общий метод получения несмещенной оценки величины $\mathsf{D}(\hat{\mu})$ при многошаговом выборе. Причем эта оценка имеет очень простой вид.

Предположим, что имеется произвольное число шагов выбора и что выбор на первом шаге проводится без возвращения, причем вероятность того, что i-е множество войдет в число n_1 извлеченных множеств, равна $\pi_i^{(1)}$, а совместная вероятность того, что на первом шаге окажутся извлеченными как i-е, так и j-е множество, равна $\pi_{ij}^{(1)}$. Используя (39.73) для *любой* оценки (не обязательно величины μ), получим

$$\mathsf{D}(\hat{\theta}) = \underset{1}{\mathsf{M}}\{\underset{>1}{\mathsf{D}}(\hat{\theta})\} + \underset{1}{\mathsf{D}}\{\underset{>1}{\mathsf{M}}(\hat{\theta})\}, \tag{39.95}$$

где нижний индекс «>1» указывает на все шаги выбора после первого. Предположим, что оценка $\hat{\theta}$ может быть записана в виде

$$\hat{\theta} = \sum_{i=1}^{n_1} t_i.$$

Оценка (39.68) удовлетворяет этому условию, так же как и в более общей ситуации оценка (39.20) при произвольном числе шагов. Если для выбора на первом шаге использовать альтернативную оценку (39.27), подставляя n_1 вместо n и рассматривая $y_{(r)}$ как сумму игреков, взятую по r-му множеству первого шага, то и эта оценка будет представима в указанном виде. То же относится и к улучшенной форме оценки в **39.11**.

Поскольку внутри каждого извлеченного на первом шаге множества выбор на последующих шагах проводится независимо, то

$$\underset{>1}{\mathsf{D}}(\hat{\theta}) = \sum_{i=1}^{n_1} \underset{>1}{\mathsf{D}}(t_i),$$

а значит, используя (39.19), соотношение (39.95) можно переписать в виде

$$\mathsf{D}(\hat{\theta}) = \underset{1}{\mathsf{D}}\{\underset{>1}{\mathsf{M}}(\hat{\theta})\} + \underset{1}{\mathsf{M}}\left\{\sum_{i=1}^{n_1} \underset{>1}{\mathsf{D}}(t_i)\right\} = \underset{1}{\mathsf{D}}\left\{\sum_{i=1}^{n_1} \underset{>1}{\mathsf{M}}(t_i)\right\} + \sum_{i=1}^{N_1} \pi_i^{(1)} \underset{>1}{\mathsf{D}}(t_i). \tag{39.96}$$

Применяя (39.18) к первому члену в правой части (39.96) и подставляя соответствующие выражения для дисперсий во втором члене, получим

$$\begin{aligned}\mathsf{D}(\hat{\theta}) &= \sum_{i=1}^{N_1} \pi_i^{(1)}(1-\pi_i^{(1)})\{\underset{>1}{\mathsf{M}}(t_i)\}^2 + \\ &+ \sum_{\substack{i,j=1\\ i\neq j}}^{N_1}\sum (\pi_{ij}^{(1)} - \pi_i^{(1)}\pi_j^{(1)}) \underset{>1}{\mathsf{M}}(t_i) \underset{>1}{\mathsf{M}}(t_j) + \\ &+ \sum_{i=1}^{N_1} \pi_i^{(1)}[\underset{>1}{\mathsf{M}}(t_i^2) - \{\underset{>1}{\mathsf{M}}(t_i)\}^2] = \end{aligned} \tag{39.97}$$

$$\begin{aligned} &= \sum_{i=1}^{N_1} \pi_i^{(1)}(1-\pi_i^{(1)}) \underset{>1}{\mathsf{M}}(t_i^2) + \sum_{\substack{i,j=1\\ i\neq j}}^{N_1}\sum (\pi_{ij}^{(1)} - \pi_i^{(1)}\pi_j^{(1)}) \underset{>1}{\mathsf{M}}(t_i) \underset{>1}{\mathsf{M}}(t_j) + \\ &+ \sum_{i=1}^{N_1} (\pi_i^{(1)})^2 \underset{>1}{\mathsf{D}}(t_i). \end{aligned} \tag{39.98}$$

39.46 Мы хотим теперь получить несмещенную оценку для (39.98). Поскольку t_i и t_j независимы при проведении выбора на следующих после первого шагах, то в первых двух членах правой части (39.98) нужно просто при построении оценки заменить $\underset{>1}{\mathsf{M}}(t_i^2)$ на t_i^2, а $\underset{>1}{\mathsf{M}}(t_i)\,\underset{>1}{\mathsf{M}}(t_j)$ на $t_i t_j$. Если сделать это, то первые два члена окажутся также равными первым двум членам в (39.97), где $\underset{>1}{\mathsf{M}}(t_i)$ заменено на t_i. Но поскольку первые два члена в (39.97) равны $\underset{1}{\mathsf{D}}\left\{\sum_{i=1}^{n_1}\underset{>1}{\mathsf{M}}(t_i)\right\}$ в (39.96), то, значит,

$$\sum_{i=1}^{N_1}\pi_i^{(1)}(1-\pi_i^{(1)})\,t_i^2+\sum_{\substack{i,\,j=1\\ i\neq j}}^{N_1}\sum(\pi_{ij}^{(1)}-\pi_i^{(1)}\pi_j^{(1)})\,t_i t_j=\underset{1}{\mathsf{D}}\left\{\sum_{i=1}^{n_1}t_i\right\}=\underset{1}{\mathsf{D}}(\hat\theta). \tag{39.99}$$

Последний член в правой части (39.98) также легко оценить, так как, если

$$\underset{>1}{\mathsf{M}}\{\underset{>1}{\hat{\mathsf{D}}}(t_i)\}=\underset{>1}{\mathsf{D}}(t_i),$$

то, согласно (39.19),

$$\mathsf{M}\left\{\sum_{i=1}^{n_1}\pi_i^{(1)}\underset{>1}{\hat{\mathsf{D}}}(t_i)\right\}=\underset{1}{\mathsf{M}}\left[\sum_{i=1}^{n_1}\pi_i^{(1)}\underset{>1}{\mathsf{M}}\{\underset{>1}{\hat{\mathsf{D}}}(t_i)\}\right]=$$

$$=\underset{1}{\mathsf{M}}\left[\sum_{i=1}^{n_1}\pi_i^{(1)}\underset{>1}{\mathsf{D}}(t_i)\right]=\sum_{i=1}^{N_1}(\pi_i^{(1)})^2\,\underset{>1}{\mathsf{D}}(t_i).$$

Используя (39.99), видим, что (39.98) есть математическое ожидание величины

$$\mathsf{D}^*(\hat\theta)=\underset{1}{\mathsf{D}}(\hat\theta)+\sum_{i=1}^{n_1}\pi_i^{(1)}\underset{>1}{\hat{\mathsf{D}}}(t_i). \tag{39.100}$$

Но (39.100) пока что не статистика, ибо нам нужно получить еще оценку для $\underset{1}{\mathsf{D}}(\hat\theta)$. Если $\underset{1}{\hat{\mathsf{D}}}(\hat\theta)$ — несмещенная оценка для $\underset{1}{\mathsf{D}}(\hat\theta)$, то в результате несмещенной оценкой для (39.98) будет

$$\hat{\mathsf{D}}(\hat\theta)=\underset{1}{\hat{\mathsf{D}}}(\hat\theta)+\sum_{i=1}^{n_1}\pi_i^{(1)}\underset{>1}{\hat{\mathsf{D}}}(t_i). \tag{39.101}$$

Впервые утверждение, соответствующее (39.101), сформулировал в самом общем виде Дербин (1953). До этого некоторые частные случаи рассматривал Иэйтс. Мы сформулируем это общее утверждение в виде следующего правила.

Несмещенная оценка значения дисперсии при многошаговом выборе, когда выбор на первом шаге проводится без возвращения, может быть получена в виде суммы двух компонент. Первая компонента оценивает дисперсию, как если бы имел место только выбор на первом шаге. Вторая компонента является взвешенной суммой оценок (внутри выбранных множеств первого шага) для дисперсий, соответствующих последующим шагам выбора (при фиксированном выборе множеств первого шага). При этом веса равны вероятностям извлечения соответствующих множеств первого шага.

39.47 Для того чтобы выражение (39.101) было легче использовать, процесс разбиения на компоненты можно продолжить. Если считать, что $t_i=\sum_{j=1}^{n_{i2}}t_{ij}$, то (39.101) можно применить к $\underset{>1}{\hat{\mathsf{D}}}(t_i)$ и получить

$$\underset{>1}{\hat{\mathsf{D}}}(t_i)=\underset{2}{\hat{\mathsf{D}}}(t_i)+\sum_{j=1}^{n_{i2}}\pi_{ij}^{(2)}\underset{>2}{\hat{\mathsf{D}}}(t_{ij}), \tag{39.102}$$

где $\pi_{ij}^{(2)}$ — вероятность того, что внутри выбранного на первом шаге i-го множества будет извлечено j-е множество. Подставляя (39.102) в (39.101), получим

$$\hat{\mathsf{D}}(\hat\theta)=\underset{1}{\hat{\mathsf{D}}}(\hat\theta)+\sum_{i=1}^{n_1}\pi_i^{(1)}\underset{2}{\hat{\mathsf{D}}}(t_i)+\sum_{i=1}^{n_1}\pi_i^{(1)}\sum_{j=1}^{n_{i2}}\pi_{ij}^{(2)}\underset{>2}{\hat{\mathsf{D}}}(t_{ij}). \tag{39.103}$$

Теперь понятна схема дальнейших обобщений. Для $p\geqslant 2$ шагов оценка представима в виде

$$\hat{\mathsf{D}}(\hat\theta)=\underset{1}{\hat{\mathsf{D}}}(\hat\theta)+\sum_{r=2}^{p}\left[\sum_{i=1}^{n_1}\pi_i^{(1)}\sum_{j=1}^{n_{i2}}\pi_{ij}^{(2)}\cdots\sum_{l=1}^{n_{ij\ldots r-1}}\pi_{ij\ldots l}^{(r-1)}\underset{r}{\hat{\mathsf{D}}}(t_{ij\ldots l})\right]. \tag{39.104}$$

Из (39.104) следует, что если все $\pi_i^{(1)}$ достаточно малы, чтобы ими можно было пренебречь (что может быть только в случае, когда число N_1 множеств первого шага велико), то существенный вклад в значения $\mathsf{D}(\hat\theta)$ вносит только первый член. В этом случае методы извлечения на следующих за первым шагах не влияют на форму приближенной оценки для значения дисперсии, хотя, конечно, эти методы влияют на само значение, поскольку для разных методов будут получаться разные значения.

39.48 Для полностью симметричного случая с равными вероятностями, для которого значение $\mathsf{D}(\hat\mu)$ было приведено в (39.85), оценка величины $\mathsf{D}(\hat\mu)$, соответствующая (39.104),

равна (как читателю предлагается показать в упражнении 39.25)

$$\hat{\mathsf{D}}(\hat{\mu}) = \frac{s_1^2}{n_1}\left(1 - \frac{n_1}{N_1}\right) + \sum_{r=2}^{p}\left(\frac{n_1}{N_1}\frac{n_2}{N_2}\cdots\frac{n_{r-1}}{N_{r-1}}\right)\frac{s_r^2}{n_1 n_2 \ldots n_r}\left(1 - \frac{n_r}{N_r}\right), \tag{39.105}$$

где

$$s_r^2 = \frac{1}{n_1 n_2 \ldots n_{r-1}(n_r - 1)}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\cdots\sum_{u=1}^{n_r}(y_{ij\ldots tu} - m_{ij\ldots t})^2$$

является выборочным значением, соответствующим σ_r^2, которое определяется равенством, следующим за (39.85). Соотношение (39.105) имеет ту же структуру, что и (39.85), с единственным отличием. Все члены после первого умножаются на произведение долей выбранных на предыдущих шагах элементов, т. е. $\frac{n_1}{N_1}\cdot\frac{n_2}{N_2}\cdot\ldots$ Так же как и в конце **39.47** здесь можно убедиться, что если величиной n_1/N_1 можно пренебречь, то

$$\hat{\mathsf{D}}(\hat{\mu}) \doteqdot \frac{s_1^2}{n_1} \tag{39.106}$$

вне зависимости от методов извлечения на следующих за первым шагах.

39.49 Если на первом шаге выбор проводится с возвращением, то (39.104) нужно видоизменить так, чтобы принять во внимание, что (39.19) заменяется на (39.33). Достаточно пересмотреть вывод соотношения (39.101), из которого следует (39.104). Отметим сначала, что, поскольку (39.96) основывается на том, что выбор на последующих шагах проводится независимо для различных множеств, извлеченных на первом шаге, то схема выбора должна удовлетворять следующему условию. Если какое-то множество первого шага было выбрано $r > 1$ раз, то на последующих шагах из этого множества нужно r раз провести независимый выбор.

Как уже указывалось в **39.12**, при выборе с возвращением, в соотношении (39.18) нужно в двойной сумме добавить член π_{ij} (но не $\pi_i\pi_j$) с равными индексами. При этом член с π_i^2 из первой суммы можно присоединить ко второй сумме. В результате убеждаемся, что (39.97) нужно заменить на

$$\begin{aligned}\mathsf{D}(\hat{\theta}) &= \sum_{i=1}^{N_1}\pi_i^{(1)}\{\underset{>1}{\mathsf{M}}(t_i)\}^2 + \sum_{i,j=1}^{N_1}(\pi_{ij}^{(1)} - \pi_i^{(1)}\pi_j^{(1)})\underset{>1}{\mathsf{M}}(t_i)\underset{>1}{\mathsf{M}}(t_j) + \\ &\quad + \sum_{i=1}^{N_1}\pi_i^{(1)}[\underset{>1}{\mathsf{M}}(t_i^2) - \{\underset{>1}{\mathsf{M}}(t_i)\}^2] = \\ &= \sum_{i=1}^{N_1}\pi_i^{(1)}\underset{>1}{\mathsf{M}}(t_i^2) + \sum_{i,j=1}^{N_1}(\pi_{ij}^{(1)} - \pi_i^{(1)}\pi_j^{(1)})\underset{>1}{\mathsf{M}}(t_i)\underset{>1}{\mathsf{M}}(t_j),\end{aligned} \tag{39.107}$$

что является аналогом (39.98). Как и в (39.99), видим, что это равно математическому ожиданию от

$$\underset{1}{\mathsf{D}}(\hat{\theta}) = \sum_{i=1}^{N_1}\pi_i^{(1)}t_i^2 + \sum_{i,j=1}^{N_1}(\pi_{ij}^{(1)} - \pi_i^{(1)}\pi_j^{(1)})t_i t_j. \tag{39.108}$$

А значит, в этом случае статистикой, дающей несмещенную оценку, является вместо (39.101) просто

$$\hat{\mathsf{D}}(\hat{\theta}) = \underset{1}{\hat{\mathsf{D}}}(\hat{\theta}). \tag{39.109}$$

Таким образом, последующие шаги выбора не вносят никакого вклада и влияют только на *значения* $\hat{\mathsf{D}}(\hat{\theta})$, а не на ее *форму записи*. Очевидно, что приведенное соотношение служит также для замены формулы (39.104). Таким образом, если на первом шаге выбор проводится с возвращением, то правило Иэйтса — Дербина, приведенное в конце **39.46**, упрощается. Нужно учитывать только первую компоненту, фигурирующую в этом правиле.

Если выбор на первом шаге проводится с равными вероятностями, то (39.109) сводится к (39.106), которое превращается теперь в точное равенство. И вообще, с учетом замечания в последнем абзаце **39.47** мы видим, что (39.109) можно рассматривать как предельный случай соотношения (39.104), когда все $\pi_i^{(1)}$ стремятся к нулю. Это соответствует тому, что оценка дисперсии при одношаговом простом случайном выборе с возвращением может быть получена из соответствующей формулы для выбора без возвращения, если положить $n/N \to 0$.

39.50 Если при каждом извлечении первого шага вероятности выбора одни и те же, то общая формула (39.109) может быть на самом деле выписана явно, поскольку в этом случае можно использовать выражение для оценки дисперсии в одношаговом выборе с возвращением с неравными вероятностями, полученное в **39.12**. При этом вместо (39.34) нужно взять оценку $\hat{\theta} = \sum_{i=1}^{n_1} t_i$, так что взамен (39.36) мы будем иметь

$$\hat{\mathsf{D}}(\hat{\theta}) = \underset{1}{\mathsf{D}}(\hat{\theta}) = \frac{n_1}{n_1 - 1}\sum_{i=1}^{n_1}\left(t_i - \frac{\hat{\theta}}{n_1}\right)^2. \tag{39.110}$$

Таким образом, для многошагового выбора с произвольным числом шагов в случае, когда на первом шаге выбор проводится с возвращением и с неизменными вероятностями при каждом извлечении, оценка значения дисперсии имеет очень простой вид. При этом никаких условий на последующие шаги (за исключением упомянутых в первом абзаце **39.49**) не накладывается.

Размещение с минимальной дисперсией при многошаговом выборе

39.51 Ограничимся сначала изучением случая, когда на каждом шаге выбора из каждого множества предыдущего шага извлекается одно и то же число элементов. (Для случая трех шагов это означает, что $n_{i2} = n_2$, $n_{ij3} = n_3$.) В этом случае как общая формула (39.82) для равных вероятностей, так и результат (39.92) для пропорциональных объему вероятностей имеет вид

$$\mathsf{D}(\hat{\mu}) = v_0 + \frac{v_1}{n_1} + \frac{v_2}{n_1 n_2} + \frac{v_3}{n_1 n_2 n_3}, \qquad (39.111)$$

где v_l — функции только от величин, определяемых совокупностью (но не выборкой). Во многих ситуациях весьма правдоподобно, что функция затрат при трехшаговом выборе имеет вид

$$C = c_0 + n_1 c_1 + n_1 n_2 c_2 + n_1 n_2 n_3 c_3, \qquad (39.112)$$

где c_0 — накладные расходы, а c_l — стоимость извлечения одного множества l-го шага. Соотношения (39.111—12) имеют тот же вид, что и (39.50—1) с $w_l = n_1 \ldots n_l$; из (39.53) следует, что при фиксированном C значение $\mathsf{D}(\hat{\mu})$ будет минимальным (или при фиксированном $\mathsf{D}(\hat{\mu})$ величина C), если взять

$$(n_1 \ldots n_l)^2 \propto v_l / c_l, \qquad l = 1, 2, 3. \qquad (39.113)$$

Рассматривая отношения (39.113) при последовательных значениях l, получим

$$n_3^2 = \frac{v_3}{c_3} \cdot \frac{c_2}{v_2}, \qquad n_2^2 = \frac{v_2}{c_2} \cdot \frac{c_1}{v_1}; \qquad (39.114)$$

при этом n_1 определяется из (39.114) и того из выражений (39.111—12), которое фиксируется. На этот результат стоит обратить внимание. Из него следует, что объемы выборок на последующих шагах определяются только стоимостью и величинами, зависящими только от совокупности, *вне зависимости от полного объема выборки*, равного $n_1 n_2 n_3$. Так что, если меняется количество отпущенных средств (или желаемая точность оценки), то это приводит к изменению только n_1. Очевидно, что аналогичное утверждение справедливо при произвольном числе шагов $p \geqslant 2$. При этом (39.113) выполняется тогда для $l = 1, \ldots, p$, а $p - 1$ отношений с последовательными значениями l, так же как в (39.114), определяют $n_2, n_3, \ldots, n_p$, а n_1, как и выше, определяется по величине полных затрат или по желаемой точности выводов.

39.52 В **39.51** рассматривался случай, когда объем выборки зависит только от номера шага, и изучался вопрос о наилучших объемах выборки на разных шагах при заданных вероятностях извлечения. Так что схема выбора фиксировалась, а определению подлежал только объем выборок. Теперь, следуя Хансену и Гурвицу (1949), можно поставить более трудный вопрос. Какими должны быть сами *вероятности извлечения*, чтобы при фиксированных затратах значение дисперсии было минимальным? В **39.8** мы видели, что если бы в одношаговом случае вероятности π_i можно было выбрать пропорциональными значениям y_i, то дисперсия тождественно равнялась бы нулю. Как указывает общая формула для дисперсии (39.98), в многошаговом случае ситуация более сложная, поскольку на результат влияют вероятности последующих шагов. Однако, если на первом шаге выбор проводится с возвращением, то (39.97) заменяется на (39.107). Более того, если (как читателю предлагается показать в упражнении 39.27) при каждом извлечении на первом шаге используются одни и те же вероятности, то (39.107) сводится с применением (39.32) к

$$\mathsf{D}(\hat{\theta}) = \sum_{i=1}^{N_1} \pi_i^{(1)} \{\underset{>1}{\mathsf{M}}(t_i)\}^2 - \frac{1}{n_1}\left\{\sum_{i=1}^{N_1} \pi_i^{(1)} \underset{>1}{\mathsf{M}}(t_i)\right\}^2 + \sum_{i=1}^{N_1} \pi_i^{(1)} [\underset{>1}{\mathsf{M}}(t_i^2) - \{\underset{>1}{\mathsf{M}}(t_i)\}^2] = \sum_{i=1}^{N_1} \pi_i^{(1)} \underset{>1}{\mathsf{M}}(t_i^2) - \frac{1}{n_1}\left\{\sum_{i=1}^{N_1} \pi_i^{(1)} \underset{>1}{\mathsf{M}}(t_i)\right\}^2, \qquad (39.115)$$

а это выражение зависит только от $\pi_i^{(1)}$ на первом шаге.

39.53 Ограничимся теперь двухшаговым выбором, когда на первом шаге проводится выбор с возвращением и с одними и теми же вероятностями извлечения (так что справедливо (39.115)), а на втором шаге проводится выбор с равными вероятностями и с равновзвешивающей схемой (см. **39.41**), так что для всех элементов совокупности полная вероятность извлечения одна и та же. Как и в **39.42**, это выбор с пропорциональными объему вероятностями, и $\pi_i^{(1)} \cdot \frac{n_{i2}}{N_{i2}} = \frac{n}{N}$, так что

$$n_{i2} = \frac{N_{i2}}{\pi_i^{(1)}} \cdot \frac{n}{N}. \qquad (39.116)$$

Рассмотрим теперь несколько более общий вид функции затрат, чем двухшаговый эквивалент (39.112), а именно:

$$C = c_0 + n_1 c_1 + \left(\sum_{i=1}^{n_1} N_{i2}\right) c_2 + \left(\sum_{i=1}^{n_1} n_{i2}\right) c_2'. \qquad (39.117)$$

Здесь второму шагу соответствуют две компоненты. Одна пропорциональна полному объему извлеченных множеств первого шага, а вторая — объему выборки. ($N_{i2}c_2$ можно рассматривать как затраты на подготовку i-го множества первого шага к проведению последующих извлечений.) Если $c_2 = 0$ и $n_{i2} = n_2$ для всех i, то (39.117) снова превращается в (39.112). В соответствии с (39.116) соотношение (39.117) можно переписать в виде

$$C = c_0 + n_1 c_1 + \sum_{i=1}^{n_2} N_{i2}\left(c_2 + \frac{n}{N}\frac{c_2'}{\pi_i^{(1)}}\right). \qquad (39.118)$$

Но (39.118) представляет собой случайную величину, значение которой нельзя фиксировать заранее, поэтому вместо (39.118) мы будем рассматривать ее математическое ожидание, равное в соответствии с (39.33)

$$\mathsf{M}(C) = c_0 + n_1 c_1 + c_2 \sum_{i=1}^{N_1} \pi_i^{(1)} N_{i2} + nc_2'.$$

Поскольку, согласно (39.32), $\sum_{i=1}^{N_1} \pi_i^{(1)} = n_1$, перепишем это в виде

$$\mathsf{M}(C) - c_0 - nc_2' = \sum_{i=1}^{N_1} \pi_i^{(1)} (c_1 + N_{i2}c_2). \qquad (39.119)$$

Так как выбор равновзвешивающий, то, согласно **39.41**,

$$\hat{\mu} = \sum_{i=1}^{n_1} t_i = \frac{1}{n}\sum_{i=1}^{n_1}\sum_{j=1}^{n_{i2}} y_{ij} = m,$$

а значит, в соответствии с (39.116) можно положить

$$t_i = \frac{1}{n}\sum_{j=1}^{n_{i2}} y_{ij} = \frac{N_{i2}}{N\pi_i^{(1)}} \cdot \frac{1}{n_{i2}}\sum_{j=1}^{n_{i2}} y_{ij} = \frac{N_{i2}}{N}\frac{m_i}{\pi_i^{(1)}}.$$

Таким образом, в рассматриваемом случае (39.115) принимает вид

$$N^2\mathsf{D}(\hat{\mu}) + \frac{1}{n_1}\left\{\sum_{i=1}^{N_1} \underset{>1}{\mathsf{M}}(N_{i2}m_i)\right\}^2 = \sum_{i=1}^{N_1} \frac{1}{\pi_i^{(1)}} \underset{>1}{\mathsf{M}}\{(N_{i2}m_i)^2\}. \qquad (39.120)$$

39.54 Выше мы видели, что математическое ожидание функции затрат — линейная функция от $\pi_i^{(1)}$, а дисперсия — линейная функция от обратных к $\pi_i^{(1)}$ величин. Таким образом, мы находимся в ситуации (39.50—1), и справедливы рассуждения из **39.20** с заменой w_i на $\pi_i^{(1)}$, c_i на $(c_1 + N_{i2}c_2)$, а v_i — на $\{\underset{>1}{\mathsf{M}}(N_{i2}m_i)^2\}$. Из (39.53) сразу следует, что $\pi_i^{(1)}$, минимизирующие $\mathsf{D}(\hat{\mu})$ при фиксированном $\mathsf{M}(C)$, определяются соотношениями

$$(\pi_i^{(1)})^2 \propto \frac{\{\underset{>1}{\mathsf{M}}(N_{i2}m_i)^2\}}{c_1 + N_{i2}c_2}, \qquad i = 1, \ldots, N_1. \qquad (39.121)$$

Знаменатель в правой части равен полным затратам на извлечение i-го множества первого шага и на его подготовку к проведению извлечений второго шага. Числитель соответствует отклонениям внутри i-го множества первого шага. Как и следовало ожидать, результат носит общий характер.

Если $n_{i2} = N_{i2}$, то $N_{i2}m_i \equiv \sum_{j=1}^{N_{i2}} y_{ij}$. Если, кроме того, $c_2 = 0$, то (39.121) совпадает с одношаговым результатом: $\pi_i^{(1)}$ должны быть пропорциональны суммарным значениям игреков в i-м множестве. И вообще, если значением $N_{i2}c_2$ можно пренебречь по сравнению с c_1, а m_i меняются мало, то приближенно мы должны иметь $\pi_i^{(1)} \propto N_{i2}$, т. е. выбор проводится почти с пропорциональными объему вероятностями. В противоположном случае, когда можно пренебречь c_1 по сравнению с $N_{i2}c_2$, то при мало изменяющихся m_i получим приближенно $\pi_i^{(1)} \propto N_{i2}^{1/2}$. Для большинства практических ситуаций эти случаи являются крайними.

39.55 Вычисление относительной эффективности одношагового и многошагового выбора и даже просто получение оценки эффективности многошагового выбора, вообще говоря, существенно сложнее, чем аналогичная задача для расслоенной выборки, которая рассматривалась в **39.27**. Иэйтс (1960) рассматривает несколько частных случаев. При многошаговом выборе значение дисперсии редко уменьшается, на самом деле причиной применения многошагового выбора является желание уменьшить затраты, а не значение дисперсии, при этом дополнительные ресурсы можно конечно использовать для увеличения объема выборки. Результат **39.51** делает это утверждение понятным. Там показано, что одношаговый случайный выбор наиболее эффективен в том случае, когда решение (39.114) таково, что n_2 и n_3 принимают максимально возможные значения, т. е. $n_2 = N_{i2}$, $n_3 = N_{ij3}$. Поскольку N_{i2} и N_{ij3} обычно сами велики, то такое решение требует очень больших затрат

и отношений (39.114), которые почти никогда на практике не достигаются.

39.56 Отметим, наконец, кратко, что та выгода от использования расслоения, которую мы обсуждали в **39.13—27**, имеет место и для каждого шага многошагового выбора. При этом на практике часто схема многошагового выбора объединяется с расслоением, особенно на первом шаге, который часто дает наибольший вклад в величину дисперсии. Вся приведенная выше теория применяется отдельно внутри каждого слоя, включая правило Иэйтса — Дербина из **39.46** и **39.49** для оценки дисперсии.

УПРАЖНЕНИЯ

39.1 Проводится n извлечений при случайном выборе с возвращением из совокупности, состоящей из N элементов. В выборке оказалось d *различных* элементов. Доказав неравенство

$$\mathsf{M}\left(\frac{1}{d}\right) \leqslant \frac{1}{N}\left(1+\frac{N-1}{n}\right),$$

показать, что среднее соответствующих d значений является несмещенной оценкой для среднего совокупности и что дисперсия этой оценки при $n > 2$ меньше (а при $n = 2$ равна) дисперсии для полного выборочного среднего.

(Радж и Хамис, 1958. Результат остается верным, если d фиксировано, а n случайно.)

39.2 Из совокупности, состоящей из N элементов, проводится выбор без возвращения двух элементов. Вероятность, что при первом извлечении выбирается i-й элемент, равна p_i, где $\sum_{i=1}^{N} p_i = 1$. Если при первом извлечении был выбран i-й элемент, то вероятности того, что при втором извлечении будет выбран j-й элемент ($j \neq i$), пропорциональны

$${}_{(2)}p_j = p_j\left(\frac{1}{1-2p_j}+\frac{1}{1-2p_i}\right).$$

Показать, что

$$\sum_{\substack{j=1\\ j\neq i}}^{N} {}_{(2)}p_j = 1+\sum_{k=1}^{N}\frac{p_k}{1-2p_k},$$

что ${}_{(2)}p_j p_i$ симметрично по i и j, а значит, p_i равно также безусловной вероятности того, что при втором извлечении будет выбран i-й элемент. А значит, в **39.6** $\pi_i = 2p_i$ и

$$\pi_{ij} = 2p_i p_j\left(\frac{1}{1-2p_j}+\frac{1}{1-2p_i}\right)\left(1+\sum_k \frac{p_k}{1-2p_k}\right)^{-1}.$$

(Дербин, 1965. Сэмпфорд (1967) обобщил метод на случай $n > 2$).

39.3 Показать, что если в (39.23) вероятности π_i пропорциональны значениям y_i (которые считаются положительными), то $\mathsf{D}(\hat{\mu}) = 0$. Показать также, что в этом случае оценка дисперсии (39.24) также равняется нулю, а оценка дисперсии (39.22) принимает вид

$$\hat{\mathsf{D}}_1(\hat{\mu}) = \mu^2\left\{1-\frac{1}{n^2}\sum_{\substack{i,\,j=1\\ i\neq j}}^{n}\sum \frac{\pi_i\pi_j}{\pi_{ij}}-\frac{m}{N\mu}\right\},$$

где m — выборочное среднее. Вывести отсюда, что $\hat{\mathsf{D}}_1(\hat{\mu})$ может принимать отрицательные значения.

39.4 В (39.24) показать, что если $\pi_{1j} = \pi_{2j} = \dfrac{\delta}{N-2}$ для всех $j \neq 1, 2$ и в результате двух наблюдений получены y_1 и y_2, то

$$\hat{\mathsf{D}}_2(\hat{\mu}) = \frac{(\pi_{12}+\delta)^2-\pi_{12}}{\pi_{12}(\pi_{12}+\delta)^2}(y_1-y_2)^2.$$

Показать отсюда, что $\hat{\mathsf{D}}_2(\hat{\mu})$ может принимать отрицательные значения.

(Дербин, 1953.)

39.5 Показать, что если при первом извлечении проводится выбор без возвращения с неравными вероятностями, так что вероятность извлечения i-го элемента равна $p_i > 0$, $\sum_i p_i = 1$, а при всех последующих извлечениях вероятности выбора равны, то в обозначениях **39.6** получим

$$\pi_{ij} = \frac{n-1}{(N-1)(N-2)}\{(N-n)(p_i+p_j)+n-2\},$$

$$\pi_i = \frac{1}{N-1}\{(N-n)p_i+(n-1)\},$$

а значит, оценка дисперсии (39.24) всегда положительная для этой схемы выбора.

(А. Сен, 1953.)

39.6 Пусть $N \geqslant 3$, $n = 2$ и первое извлечение, так же как в примере 39.5, проводится с вероятностями p_i. Пусть для оставшихся $(N-1)$ вероятностей при втором извлечении выполняются те же пропорции, что и при первом извлечении. Показать, что

$$\pi_{ij} = p_i p_j\left(\frac{1}{1-p_i}+\frac{1}{1-p_j}\right),$$

$$\pi_i = p_i\left(\frac{1}{1-p_j}+\sum_{\substack{r=1\\ r\neq i,\,j}}^{N}\frac{p_r}{1-p_r}\right),$$

и что при фиксированных i и j сумма в выражении для π_i минимальна, когда

$$\frac{p_r}{1-p_r} = \frac{(1-p_i-p_j)}{(N-2)-(1-p_i-p_j)}.$$

Вывести отсюда, что для этой схемы выбора оценка дисперсии (39.24) всегда положительна.

(А. Сен, 1953.)

39.7 Показать, что для схемы выбора из упражнения 39.5 при любом множестве p_i должны выполняться неравенства $\pi_i > \frac{n-1}{N-1}$ для всех i. Используя (39.16), показать что не более чем одно π_i может равняться единице.

39.8 Показать, что для случайного выбора без возвращения с равными вероятностями, как (39.22), так и (39.24) сводятся к оценке дисперсии $\hat{\mathbf{D}}(m)$, задаваемой (39.12).

39.9 Показать, что z_u, определенное в (39.25), имеет дисперсию

$$\mathbf{D}(z_u) = \sum_{(1)} p_{(1)} \sum_{(2)} p_{(2)} \cdots \sum_{(u-1)} p_{(u-1)} \sum_{(u)} \frac{y_{(u)}^2}{p_{(u)}} - \\ - \sum_{(1)} p_{(1)} \sum_{(2)} p_{(2)} \cdots \sum_{(u-1)} \left\{ N\mu - \sum_{r=1}^{u-1} y_r^{(2)} \right\}, \qquad u \geqslant 2,$$

где каждая сумма берется по всем возможным при указанном извлечении элементам. Используя этот результат, показать, что при $n = 2$ статистика (39.27) имеет дисперсию

$$\mathbf{D}(\bar{z}) = \frac{1}{4}\left[\left\{\sum_{(1)} \frac{y_{(1)}^2}{p_{(1)}} - N^2\mu^2\right\} + \left\{\sum_{(1)} p_{(1)} \sum_{(2)} \frac{y_{(2)}^2}{p_{(2)}} - \sum_{(1)} p_{(1)} (N\mu - y_{(1)})^2\right\}\right].$$

(Радж, 1956.)

39.10 Показать, что если при выборе с равными вероятностями z, определяемое (39.26), нужно свести к выборочному среднему m, помноженному на N, то веса должны удовлетворять соотношениям

$$\frac{c_u}{c_1} = \frac{N^{(u-1)}}{(N-2)^{(u-1)}}, \qquad 2 \leqslant u \leqslant n.$$

Эти $n-1$ условий, наряду с равенством $\Sigma c_u = 1$ однозначно определяют веса.

(См. Радж, 1956.)

39.11 Рассмотрим следующую процедуру отбора n элементов из совокупности, содержащей N элементов y_s с использованием вероятностей π_s ($s = 1, \ldots, N$):

(1) Возьмем число M, которое больше или равно максимальному из значений π_s, скажем $\pi_{\max}$. Случайным образом с равными вероятностями выберем число r между 0 и M. Аналогично выберем целое число s_1 среди целых чисел от 1 до N. Если $r \leqslant \pi_{s_1}$, то включим y_{s_1} в выборку. Если $r > \pi_{s_1}$, то повторим процедуру заново.

(2) Выберем далее значения s_t последовательно без возвращения из совокупности целых чисел от 1 до N. Для этой последовательности включим в выборку те значения y_{s_t}, для которых последовательные суммы $r + \sum_{u=2}^{t} \pi_u$ впервые превзойдут одно из значений $M, 2M, \ldots, (n-1)M$.

Показать, что если $M \leqslant \frac{1-\pi_{\max}}{n-1}$, то описанная процедура приводит к выбору без возвращения с требуемыми вероятностями. Показать, что повторение n раз операции (1) приводит к выборке с возвращением с заданными π_s.

(Лахири, 1951; Гранди, 1954.)

39.12 Пусть $x_1, \ldots, x_n$ — некоррелированные случайные величины с одинаковым средним μ и не обязательно равными дисперсиями. Показать, что $\bar{x} = \sum_{i=1}^{n} x_i/n$ является несмещенной оценкой для μ и что $\sum_i (x_i - \bar{x})^2/\{n(n-1)\}$ является несмещенной оценкой дисперсии величины $\bar{x}$.

39.13 Показать, что обобщения биномиального и пуассоновского выбора, рассмотренные в **5.10**, являются частными случаями расслоенного выбора. Получить отсюда выражения для дисперсий (5.26) и формулу, следующую за (5.27).

39.14 В **39.15**—18 показать, что

$$D = \mathbf{D}(m_R) - \mathbf{D}(\hat{\mu}_{\text{МД}}) - \frac{(N-n)}{nN(N-1)} \sum_i N_i (\mu_i - \mu)^2 = \\ = \frac{1}{nN^2(N-1)}(P - Q - R),$$

где

$$P = N\left\{N \sum N_i\sigma_i^2 - \left(\sum N_i\sigma_i^2\right)^2\right\} \geqslant 0,$$
$$Q = n\left(\sum N_i\sigma_i^2 - N \sum \sigma_i^2\right) < 0,$$
$$R = N^2 \sum \sigma_i^2 - \left(\sum N_i\sigma_i\right)^2 > 0.$$

Отсюда $P = 0$ тогда и только тогда, когда все σ_i равны. Используя (39.46), показать теперь, что если $N \to \infty$, а N_i/N фиксированы, то выполняется (39.49).

Показать далее, что если $P = 0$, а n достаточно мало, то D — отрицательно. Если $N - n$ достаточно мало, то также

$$\mathbf{D}(m_R) < \mathbf{D}(\hat{\mu}_{\text{МД}}).$$

(Армитедж, 1947.)

39.15 В упражнении 39.14 показать, что если σ_i достаточно различаются, то $P - R$, как правило, будет положительным и что тогда D — убывающая функция n, так что уменьшение дисперсии, вызванное расслоением, убывает с возрастанием n. Вывести отсюда, что если какое-нибудь n_i, отвечающее размещению с МД, превосходит соответствующее ему N_i, то мы увеличим выгоду от стратификации, если положим $n_i = N_i$, а оставшиеся $n - N_i$ наблюдений будем проводить в соответствии с размещением с МД.

(Армитедж, 1947.)

39.16 Показать, что если планирование выборки таково, что выборочное среднее m является несмещенной оценкой для среднего совокупности μ, то

$$\mathbf{M}\left\{\frac{1}{n}\sum_{i=1}^{n}(y_i - m)^2\right\} = \sigma_y^2 - \mathbf{D}(m),$$

где $\sigma_y^2 = M(y - \mu)^2$. Проверить утверждение для случайного выбора без возвращения с равными вероятностями.

(Этот результат получил Киш.)

39.17 Показать, что если в 39.24 имеется $k = 2$ слоя, причем второй слой извлекается полностью ($n_2 = N_2$), то лучшим выбором точки разбиения c_1 области изменения y для минимизации дисперсии (39.30) является

$$c_1 = \mu_1 + \left(\frac{N_1}{n_1}\right)^{1/2} \sigma_1$$

(См. Далэниус, 1952.)

39.18 Показать, что если объем выборки n_l в каждом слое один и тот же, то точки разбиения области изменения y, минимизирующие (39.39), удовлетворяют соотношениям (см. 39.25)

$$N_l \{\sigma_l^2 + (c_l - \mu_l)^2\} = N_{l+1} \{\sigma_{l+1}^2 + (c_l - \mu_{l+1})^2\}.$$

(Кокрэн, 1961.)

39.19 *После* извлечения большой случайной выборки элементы совокупности разбиваются на k слоев. Показать, что если использовать в этих обстоятельствах оценку среднего совокупности, основанную на расслоении (эта оценка задается (39.38)), то ожидаемое значение дисперсии этой оценки (принимая во внимание случайные отклонения объемов выборок внутри слоев) приблизительно равно $\mathsf{D}(\hat{\mu}_{\text{ВРД}})$, которое определялось в (39.45).

39.20 Показать, что если объем выборок в слоях выбрать так, чтобы минимизировать дисперсию оценки величины $\mathsf{D}(\hat{\mu})$, т. е. (39.40), и если при этом значениями n_l/N_l можно пренебречь, то формула (39.42), описывающая распределение с МД заменяется на

$$\frac{n_l}{N_l} = \frac{\sigma_l (\beta_{2l} - 1)^{1/4}}{\sum_l N_l \sigma_l (\beta_{2l} - 1)^{1/4}/n},$$

где β_{2l} — отношение моментов μ_4/μ_2^2 для l-го слоя. Таким образом, если β_{2l} не постоянно для всех l, то это распределение отличается от (39.42).

(Росс, 1961.)

39.21 Показать, что если объемы выборок в слоях отличаются от объемов при распределении по слоям с МД (определяемых (39.42)) на Δn_l, то $\mathsf{D}(\hat{\mu})$ возрастает по сравнению с $\mathsf{D}(\hat{\mu}_{\text{МД}})$, так что их отношение равно приблизительно

$$1 + \frac{1}{n} \sum_l \{(\Delta n_l)^2/n_l\}.$$

39.22 Используя 39.2, показать, что если при групповом выборе с группами равного объема выбрать одну группу объема n, то (39.7) задает значение дисперсии выборочного среднего, причем ρ — внутригрупповой коэффициент корреляции для групп (см. 26.25—6, т. 2).

39.23 Показать, что обобщенные биномиальный и пуассоновский выбор, рассмотренные в 5.11, являются частным случаем двухшагового выбора. Вывести отсюда формулы для дисперсий (5.29) и (5.30).

39.24 Вывести (39.94) из (39.92).

39.25 Показать, что (39.105) является частным случаем (39.104).

39.26 Используя упражнение 39.12, получить (39.110) для оценки дисперсии при многошаговом выборе, для которого на первом шаге проводится выбор с возвращением с одним и тем же множеством неравных вероятностей при каждом извлечении.

(Дербин, 1953.)

39.27 Используя (39.32), показать, что если при каждом извлечении первого шага применять выбор с возвращением с одними и теми же не обязательно равными вероятностями, то средний член в правой части (39.107) равен

$$-\frac{1}{n_1}\left\{\sum_{i=1}^{N_1} \pi_i^{(1)} \underset{>1}{\mathsf{M}}(t_i)\right\}^2,$$

а значит, разность между дисперсией (39.107) для выбора с возвращением и дисперсией (39.97) для выбора без возвращения можно записать в виде

$$\mathsf{D} = \frac{n_1 - 1}{n_1}\left\{\sum_{i=1}^{n_1} \pi_i^{(1)} \underset{>1}{\mathsf{M}}(t_i)\right\}^2 - \sum_{\substack{i,\, j=1 \\ i \neq j}}^{N_1}\sum \pi_{ij}^{(1)} \underset{>1}{\mathsf{M}}(t_i) \underset{>1}{\mathsf{M}}(t_j).$$

Показать, что это выражение может быть как положительным, так и отрицательным, но если на первом шаге выбор проводится с равными вероятностями, то $D \geqslant 0$.

Показать, что если оценку дисперсия (39.110), справедливую для выбора с возвращением, применить в случае, когда на самом деле выбор проводится без возвращения, то у этой оценки будет смещение, равное $\frac{n_1}{n_1 - 1} D$. Так что, если при выборе без возвращения получается меньшая дисперсия ($D > 0$), то применение оценки, справедливой для выбора с возвращением завышает значение дисперсии.

(Дербин, 1953.)

39.28 Пусть задана многошаговая схема выбора такая, что выбор на s-м шаге ($s \geqslant 1$) проводится с возвращением, а выбор на шаге с номером ($s + 1$) проводится с равными вероятностями. Показать (см. упражнение 39.27), что если для множества, выбранного на s-м шаге r раз, просто увеличивать объем выборки на шаге с номером $s + 1$ в r раз, то дисперсия оценки (39.68) для среднего совокупности будет меньше, чем если на шаге с номером $s + 1$ сделать r независимых выборок заданного объема внутри этого множества.

39.29 Пусть задана многошаговая схема выбора такая, что выбор на s-м шаге ($s \geqslant 1$) проводится с возвращением. Показать, что если для множества, выбранного на s-м шаге r раз, проводить на шаге с номером $s + 1$ выборку заданного объема один раз, а результат взять с весом r, то дисперсия оценки (39.68) для среднего совокупности будет больше, чем если на шаге с номером $s + 1$ сделать r независимых выборок заданного объема внутри этого множества.

39.30 Пусть из совокупности, содержащей N элементов, проводится выбор n элементов без возвращения с неравными вероятностями. Вероятность того, что фиксированные n элементов будут извлечены, причем в фиксированном порядке, обозначим $p_{(s)}$. Пусть p_s — вероятность того, что эти же n элементов будут извлечены в произвольном порядке. Тогда $p_s = \sum_{(s)} p_{(s)}$, где сумма берется по всем $n!$ возможным перестановкам этих элементов. При этом $\sum_s p_s = 1$, где сумма берется по всем $\binom{N}{n}$ возможным выборкам объема n. Если $z_{(s)}$ — такая статистика, в которой принимается во внимание порядок выбора, а $z_s = \sum_{(s)} p_{(s)} z_{(s)}^2/p_s$, то используя (39.72—3), показать, что

$$\mathsf{M}(z_s) = \mathsf{M}(z_{(s)}), \quad \mathsf{D}(z_s) \leqslant \mathsf{D}(z_{(s)}),$$

причем равенство возможно только тогда, когда все значения $z_{(s)}$ совпадают. Используя этот результат, показать, что оценка $\bar{z}_{(s)}$, определенная в (39.27), может быть улучшена и что $\hat{\mathrm{D}}(\hat{z}_{(s)})$, определенная в (39.31), аналогично может быть улучшена.

(См. Мёрфи (1957) и Патхак (1961а).)

39.31 Проводится выбор без возвращения с неравными вероятностями. При первом извлечении вероятности выбора для N элементов совокупности равны ${}_{(1)}p_i$, $i = 1, \ldots, N$. При последнем извлечении для вероятностей выбора оставшихся к этому моменту элементов остаются те же пропорции, что и при первом извлечении. Показать, что в этом случае (39.25) можно записать в виде

$$z_1 = \frac{y_{(1)}}{{}_{(1)}p_{(1)}},$$

$$z_u = \sum_{r=1}^{u-1} y_{(r)} + \frac{y_u}{{}_{(1)}p_u}\left(1 - \sum_{r=1}^{u-1} {}_{(1)}p_{(r)}\right), \qquad u = 2, \ldots, n,$$

а значит, улучшенный вариант оценки $\bar{z}_{(s)}$, описанный в упражнении 39.30, можно записать в виде

$$\bar{z}_s = \sum_{i=1}^{n} y_i p_{s\,|\,i}/p_s,$$

где $p_{s\,|\,i}$ условная вероятность выбора наблюденных значений при условии, что при первом извлечении отбирается y_i.

(Мёрфи, 1957.)

ГЛАВА 40

ТЕОРИЯ ВЫБОРОЧНЫХ ОБСЛЕДОВАНИЙ: ДОПОЛНИТЕЛЬНАЯ ИНФОРМАЦИЯ

40.1 В главе 39 мы изучали задачу планирования выборки. Теперь мы обратимся к вопросу, который возникает при любой схеме выбора, а именно к улучшению эффективности оценки.

В **39.8** и **39.13** мы столкнулись с тем фактом, что знание переменной, сильно коррелированной с изучаемой, может помочь нам назначить вероятности выбора или сконструировать слои, чтобы уменьшить значение дисперсии оценки. Такая *дополнительная информация* о вспомогательной переменной может быть использована непосредственно для изменения вида оценки с целью улучшения ее эффективности *).

«Оценки-отношения» и их модификации

40.2 Предположим, как и в **39.2**, что проводится выбор без возвращения с равными вероятностями из конечной совокупности. Пусть нам нужно оценить среднее значение по совокупности переменной y, которое мы обозначим теперь μ_y, а мы знаем значение среднего по совокупности для переменной x, скажем μ_x, причем при извлечении мы узнаем как значение x, так и значение y. Желательно, чтобы мы имели возможность использовать это дополнительное знание. Мы предполагаем, что $\mu_x \neq 0$ $\mu_y \neq 0$.

Двумя интуитивно обоснованными оценками величины μ_y являются

$$\bar{\mu}_y = \mu_x m_y / m_x \tag{40.1}$$

и

$$\hat{\mu}_y = \mu_x m_{y/x}, \tag{40.2}$$

где m с соответствующим индексом обозначает выборочное среднее той переменной, которая указана в индексе. В (40.1) исполь-

*) Не следует смешивать с использованием дополнительной информации (инструментальных переменных) для получения возможности раздельной оценки параметров, как это было в **29.33—46** т. 2. Это скорее аналогично использованию сопутствующих переменных в ковариационном анализе (см. **35.67—8**), поскольку последние уменьшают остаточную дисперсию.

причем равенство возможно только тогда, когда все значения $z_{(s)}$ совпадают. Используя этот результат, показать, что оценка $\bar{z}_{(s)}$, определенная в (39.27), может быть улучшена и что $\hat{\mathrm{D}}(\hat{z}_{(s)})$, определенная в (39.31), аналогично может быть улучшена.

(См. Мёрфи (1957) и Патхак (1961а).)

39.31 Проводится выбор без возвращения с неравными вероятностями. При первом извлечении вероятности выбора для N элементов совокупности равны ${}_{(1)}p_i$, $i = 1, \ldots, N$. При последнем извлечении для вероятностей выбора оставшихся к этому моменту элементов остаются те же пропорции, что и при первом извлечении. Показать, что в этом случае (39.25) можно записать в виде

$$z_1 = \frac{y_{(1)}}{{}_{(1)}p_{(1)}},$$

$$z_u = \sum_{r=1}^{u-1} y_{(r)} + \frac{y_u}{{}_{(1)}p_u}\left(1 - \sum_{r=1}^{u-1} {}_{(1)}p_{(r)}\right), \qquad u = 2, \ldots, n,$$

а значит, улучшенный вариант оценки $\bar{z}_{(s)}$, описанный в упражнении 39.30, можно записать в виде

$$\bar{z}_s = \sum_{i=1}^{n} y_i p_{s\,|\,i}/p_s,$$

где $p_{s\,|\,i}$ условная вероятность выбора наблюденных значений при условии, что при первом извлечении отбирается y_i.

(Мёрфи, 1957.)

ГЛАВА 40

ТЕОРИЯ ВЫБОРОЧНЫХ ОБСЛЕДОВАНИЙ: ДОПОЛНИТЕЛЬНАЯ ИНФОРМАЦИЯ

40.1 В главе 39 мы изучали задачу планирования выборки. Теперь мы обратимся к вопросу, который возникает при любой схеме выбора, а именно к улучшению эффективности оценки.

В **39.8** и **39.13** мы столкнулись с тем фактом, что знание переменной, сильно коррелированной с изучаемой, может помочь нам назначить вероятности выбора или сконструировать слои, чтобы уменьшить значение дисперсии оценки. Такая *дополнительная информация* о вспомогательной переменной может быть использована непосредственно для изменения вида оценки с целью улучшения ее эффективности *).

«Оценки-отношения» и их модификации

40.2 Предположим, как и в **39.2**, что проводится выбор без возвращения с равными вероятностями из конечной совокупности. Пусть нам нужно оценить среднее значение по совокупности переменной y, которое мы обозначим теперь μ_y, а мы знаем значение среднего по совокупности для переменной x, скажем μ_x, причем при извлечении мы узнаем как значение x, так и значение y. Желательно, чтобы мы имели возможность использовать это дополнительное знание. Мы предполагаем, что $\mu_x \neq 0$ $\mu_y \neq 0$.

Двумя интуитивно обоснованными оценками величины μ_y являются

$$\tilde{\mu}_y = \mu_x m_y / m_x \tag{40.1}$$

и

$$\hat{\mu}_y = \mu_x m_{y/x}, \tag{40.2}$$

где m с соответствующим индексом обозначает выборочное среднее той переменной, которая указана в индексе. В (40.1) исполь-

*) Не следует смешивать с использованием дополнительной информации (инструментальных переменных) для получения возможности раздельной оценки параметров, как это было в **29.33—46** т. 2. Это скорее аналогично использованию сопутствующих переменных в ковариационном анализе (см. **35.67—8**), поскольку последние уменьшают остаточную дисперсию.

зуется *отношение* выборочных средних, а в (40.2) среднее выборочных отношений величин y и x в качестве «корректирующего множителя» к известному значению μ_x.

Математические ожидания оценок (40.1—2) получаются сразу, если заметить, что, согласно определению ковариации C,

$$C\left(\frac{m_y}{m_x},\ m_x\right)=\mathsf{M}(m_y)-\mathsf{M}\left(\frac{m_y}{m_x}\right)\mathsf{M}(m_x)=\mu_y-\frac{1}{\mu_x}\mathsf{M}(\tilde{\mu}_y)\mu_x,$$

а значит,

$$\mathsf{M}(\tilde{\mu}_y)=\mu_y-C\left(\frac{m_y}{m_x},\ m_x\right). \tag{40.3}$$

Аналогично

$$C\left(\frac{y}{x},\ x\right)=\mu_y-\mathsf{M}\left(\frac{y}{x}\right)\mu_x,$$

так что

$$\mathsf{M}(\hat{\mu}_y)=\mu_x\mathsf{M}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{x_i}\right)=\mu_x\mathsf{M}\left(\frac{y}{x}\right)=\mu_y-C\left(\frac{y}{x},\ x\right). \tag{40.4}$$

Из (40.3—4) следует, что, вообще говоря, обе оценки являются смещенными. Далее, поскольку (согласно неравенству Коши — Буняковского *)) ковариация между двумя величинами не может быть больше чем произведение их стандартных отклонений, то

$$\begin{aligned}|\mathsf{M}(\tilde{\mu}_y)-\mu_y|&\leqslant\left\{\mathsf{D}\left(\frac{m_y}{m_x}\right)\mathsf{D}(m_x)\right\}^{1/2},\\ |\mathsf{M}(\hat{\mu}_y)-\mu_y|&\leqslant\left\{\mathsf{D}\left(\frac{y}{x}\right)\mathsf{D}(x)\right\}^{1/2}.\end{aligned} \tag{40.5}$$

Из (40.5) следует, что между оценками имеется существенное различие. Если объем выборки n стремится к бесконечности, то $\mathsf{D}(m_y/m_x)$ и $\mathsf{D}(m_x)$ будут дисперсиями выборочных средних, имеют порядок n^{-1}, а значит, такой же порядок имеет и смещение оценки $\tilde{\mu}_y$. Для $\hat{\mu}_y$ это несправедливо, поскольку $\mathsf{D}(y/x)$ и $\mathsf{D}(x)$ вообще не зависят от n. На самом деле из самих определений в (40.1—2) легко видеть, что $\tilde{\mu}_y$ является состоятельной оценкой, поскольку $m_y\to\mu_y$, $m_x\to\mu_x$. В то же время $\hat{\mu}_y\to\mu_x\mu_{y/x}$, что, вообще говоря, не совпадает с μ_y. Смещение оценки $\tilde{\mu}_y$ будет подробно рассмотрено ниже в **40.9**. Но сначала мы посмотрим, как можно избавиться от смещения у оценки $\hat{\mu}_y$.

40.3 Из (40.4) понятно, что для устранения смещения оценки $\hat{\mu}_y$ нужно получить несмещенную оценку величины $C(y/x,\ x)$. Поскольку y/x наблюдается при каждом извлечении, можно вычислить выборочную ковариацию величин y/x и x. Рассматривая аналог (12.109) для двух переменных, убеждаемся, что несмещенной оценкой для K_{11} в случае конечной совокупности является k-статистика k_{11}, а значит, несмещенной оценкой для значения ковариации по совокупности является

*) См. сноску на стр. 67 (*Прим. ред.*)

$$\begin{aligned}\hat{C}\left(\frac{y}{x},\ x\right)&=\frac{N-1}{N}\cdot\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{y_i}{x_i}-m_{y/x}\right)(x_i-m_x)=\\ &=\frac{N-1}{N}\cdot\frac{n}{n-1}(m_y-m_{y/x}m_x).\end{aligned} \tag{40.6}$$

В результате из (40.2) и (40.4) получаем, что несмещенной оценкой μ_y является оценка

$$\hat{\mu}'_y=\mu_x m_{y/x}+\frac{N-1}{N}\cdot\frac{n}{n-1}(m_y-m_{y/x}m_x), \tag{40.7}$$

впервые данная Хартли и Россом (1954). Если $\frac{N-1}{N}\cdot\frac{n}{n-1}$ близко к единице, то (40.7) сводится к

$$\hat{\mu}'_y\doteqdot m_y-m_{y/x}(m_x-\mu_x). \tag{40.8}$$

40.4 Устранение смещения для $\tilde{\mu}_y$ менее важно, поскольку в **40.2** мы видели, что эта оценка состоятельна. Тем не менее этот вопрос стоит рассмотреть, поскольку оценки-отношения иногда используются при малых значениях n. Тот прием, которым мы воспользовались в **40.3**, теперь применить нельзя, поскольку подлежащей оценке ковариацией в (40.3) является $C(m_y/m_x,\ m_x)$, а выборка доставляет нам только одно значение m_y и m_x. Однако легко получить простое приближение.

Если $n=1$, то $C(m_y/m_x,\ m_x)$ совпадает с $C(y/x,\ x)$. Кроме того, ковариация между выборочными средними для двух случайных величин обратно пропорциональна объему выборки (это следует, например, из правила 10 для k-статистик в **12.14** или может быть легко доказано непосредственно). То же самое приближенно справедливо и в рассматриваемом случае, когда ищется ковариация одного выборочного среднего и отношения другого среднего к первому. Таким образом,

$$C\left(\frac{m_y}{m_x},\ m_x\right)\doteqdot\frac{1}{n}C\left(\frac{y}{x},\ x\right),$$

и, используя (40.6) из (40.3), получим приблизительно несмещенную оценку

$$\tilde{\mu}'_y=\tilde{\mu}_y+\frac{1}{n}\hat{C}\left(\frac{y}{x},\ x\right)=\mu_x\frac{m_y}{m_x}+\frac{N-1}{N}\cdot\frac{1}{n-1}(m_y-m_{y/x}m_x). \tag{40.9}$$

Этот результат другим методом получил Нието де'Паскуаль (1961). Отсутствие множителя n во втором члене (40.9) по сравнению с (40.7) снова иллюстрирует разницу порядков в величинах смещения в (40.1—2).

40.5 Теперь нам нужно рассмотреть дисперсии альтернативных модифицированных оценок-отношений (40.7) и (40.9), чтобы знать, какую из них использовать в разных обстоятельствах.

Рассмотрим только случай, когда $N \to \infty$, т. е. выбор практически является простым случайным выбором. Используя (40.6) перепишем (40.7) в виде

$$\hat{\mu}'_y = \mu_x m_{y/x} + k_{11}, \tag{40.10}$$

где k_{11} представляет собой k-статистику величин y/x и x. А значит,

$$\mathsf{D}(\hat{\mu}'_y) = \mu_x^2 \mathsf{D}(m_{y/x}) + 2\mu_x C(m_{y/x}, k_{11}) + \mathsf{D}(k_{11}), \tag{40.11}$$

что в обозначениях **13.2** т. 1 переписывается в виде

$$\mathsf{D}(\hat{\mu}'_y) = \mu_x^2 \mathsf{D}(m_{y/x}) + 2\mu_x \varkappa \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} + \varkappa \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}. \tag{40.12}$$

Далее, используя (3.80) т. 1, получим, согласно **12.14**,

$$\varkappa \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \frac{\varkappa_{21}}{n} \equiv \frac{\mu_{21}}{n}, \tag{40.13}$$

в то время как (13.7) и (3.81) дают

$$\varkappa \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} = \frac{\mu_{22}}{n} + \frac{\mu_{20}\mu_{02}}{n(n-1)} - \frac{(n-2)\mu_{11}^2}{n(n-1)}, \tag{40.14}$$

где моменты и семиинварианты в (40.13—14) относятся к совместному распределению y/x и x. Таким образом, в этих обозначениях (40.12) можно переписать в виде

$$\mathsf{D}(\hat{\mu}'_y) = \mu_{01}^2\mu_{20}/n + 2\mu_{01}\mu_{21}/n + \\ + \frac{1}{n}\left\{\mu_{22} + \frac{1}{n-1}(\mu_{20}\mu_{02} - (n-2)\mu_{11}^2)\right\}. \tag{40.15}$$

Формула (40.15) допускает полезное упрощение. Заметим сначала, что по определению

$$\mathsf{D}\left\{\left(\frac{y}{x} - \mu_{y/x}\right)(x - \mu_x)\right\} \equiv \mathsf{M}\left\{\left(\frac{y}{x} - \mu_{y/x}\right)^2 (x - \mu_x)^2\right\} - \\ - \left\{\mathsf{M}\left[\left(\frac{y}{x} - \mu_{y/x}\right)(x - \mu_x)\right]\right\}^2 = \mu_{22} - \mu_{11}^2$$

в обозначениях (40.15), которое, таким образом, эквивалентно

$$n\mathsf{D}(\hat{\mu}'_y) = \mu_{01}^2\mu_{20} + 2\mu_{01}\mu_{21} + \mathsf{D}\left\{\left(\frac{y}{x} - \mu_{y/x}\right)(x - \mu_x)\right\} + \\ + \frac{1}{n-1}(\mu_{20}\mu_{02} + \mu_{11}^2). \tag{40.16}$$

Рассмотрим теперь тождество

$$\mu_x\left(\frac{y}{x} - \mu_{y/x}\right) + \left(\frac{y}{x} - \mu_{y/x}\right)(x - \mu_x) = y - \mu_{y/x}x. \tag{40.17}$$

Если взять дисперсию левой части (40.17), то легко убедиться, что она в точности совпадает с тремя первыми членами в правой части (40.16). Таким образом, эти члены можно заменить дисперсией правой части (40.17). В результате

$$n\mathsf{D}(\hat{\mu}'_y) = \mathsf{D}(y - \mu_{y/x}x) + \frac{1}{n-1}(\mu_{02}\mu_{20} + \mu_{11}^2)$$

и, возвращаясь к нашим исходным обозначениям,

$$n\mathsf{D}(\hat{\mu}'_y) = \mathsf{D}(y) + \mu_{y/x}^2\mathsf{D}(x) - 2\mu_{y/x}C(y, x) + \\ + \frac{1}{n-1}\left\{\mathsf{D}\left(\frac{y}{x}\right)\mathsf{D}(x) + C^2\left(\frac{y}{x}, x\right)\right\}. \tag{40.18}$$

Этот результат получили Гудмэн и Хартли (1958). При $n \to \infty$ членом в фигурных скобках в (40.18) можно пренебречь, а значит,

$$n\mathsf{D}(\hat{\mu}'_y) \sim \mathsf{D}(y) + \mu_{y/x}^2\mathsf{D}(x) - 2\mu_{y/x}C(y, x). \tag{40.19}$$

Самый простой способ получения оценки для (40.18) состоит в следующем. Надо переписать (40.12) в терминах семиинвариантов и использовать k-статистики для их оценки. Гудмэн и Хартли (1958) приводят формулы для вычислений.

Робсон (1957) приводит обобщение (40.18) и ее несмещенной оценки, принимая во внимание конечность совокупности.

40.6 Подобным же образом можно получить дисперсию величины (40.9), которую по аналогии с (40.10) мы перепишем в виде

$$\tilde{\mu}'_y = \mu_x \frac{m_y}{m_x} + \frac{k_{11}}{n}. \tag{40.20}$$

Отсюда

$$\mathsf{D}(\tilde{\mu}'_y) = \mu_x^2\mathsf{D}\left(\frac{m_y}{m_x}\right) + \frac{2}{n}\mu_x C\left(\frac{m_y}{m_x}, k_{11}\right) + \frac{1}{n^2}\mathsf{D}(k_{11}).$$

Так же, как это делалось при выводе оценки $\tilde{\mu}'_y$ в **40.4**, воспользуемся следующим приближением:

$$C\left(\frac{m_y}{m_x},\ k_{11}\right) \doteqdot \frac{1}{n} C\left(\frac{y}{x},\ k_{11}\right),$$

так что, согласно (40.11),

$$\mathsf{D}(\tilde{\mu}'_y) \doteqdot \mu_x^2 \mathsf{D}\left(\frac{m_y}{m_x}\right) + \frac{1}{n^2}\left\{2\mu_x C\left(\frac{y}{x},\ k_{11}\right) + \mathsf{D}(k_{11})\right\} \doteqdot$$

$$\doteqdot \mu_x^2 \mathsf{D}\left(\frac{m_y}{m_x}\right) + \frac{1}{n^2}\{\mathsf{D}(\hat{\mu}'_y) - \mu_x^2 \mathsf{D}(m_{y/x})\}. \qquad (40.21)$$

Все три дисперсии в правой части (40.21) имеют порядок n^{-1}, а значит, второй член по сравнению с первым членом в правой части (40.21) имеет относительный порядок n^{-2}. Таким образом, наше приближение можно записать в виде

$$\mathsf{D}(\tilde{\mu}'_y) = \mu_x^2 \mathsf{D}\left(\frac{m_y}{m_x}\right)\left\{1 + O\left(\frac{1}{n^2}\right)\right\}. \qquad (40.22)$$

Поскольку смещение у исходной оценки $\tilde{\mu}_y$ имеет порядок n^{-1} (это было показано в **40.2**), то смещение у $\tilde{\mu}'_y$ не больше по порядку, а значит, квадрат смещения по порядку не больше чем n^{-2}. А значит, средняя квадратичная ошибка (которая, учитывая смещенность $\tilde{\mu}'_y$, больше, чем дисперсия, подходит для выводов) может быть записана в виде

$$\mathsf{M}\{(\tilde{\mu}'_y - \mu_y)^2\} = \mu_x^2 \mathsf{D}\left(\frac{m_y}{m_x}\right)\left\{1 + O\left(\frac{1}{n}\right)\right\}. \qquad (40.23)$$

Нието де'Паскуаль (1961) приводит более точное приближение. Главный член в (40.22) и (40.23) представляет собой дисперсию немодифицированной оценки $\tilde{\mu}_y$. Ее легко оценить до порядка n^{-1}, используя соотношение (10.17) т. 1, которое в нашем случае дает

$$\mathsf{D}\left(\frac{m_y}{m_x}\right) = \frac{1}{n}\frac{\mu_y^2}{\mu_x^2}\left\{\frac{\mathsf{D}(y)}{\mu_y^2} + \frac{\mathsf{D}(x)}{\mu_x^2} - \frac{2C(y, x)}{\mu_y \mu_x}\right\}. \qquad (40.24)$$

Отсюда (40.23) принимает вид

$$n\mathsf{M}\{(\tilde{\mu}'_y - \mu_y)^2\} \sim \mathsf{D}(y) + \frac{\mu_y^2}{\mu_x^2}\mathsf{D}(x) - 2\frac{\mu_y}{\mu_x}C(y, x). \qquad (40.25)$$

Но (40.25) можно оценить с небольшим смещением, заменяя μ_y/μ_x на m_y/m_x, а дисперсию и ковариацию их несмещенными оценками. В результате получаем следующую оценку средней квадратичной ошибки:

$$\hat{\mathsf{M}}\{(\tilde{\mu}'_y - \mu_y)^2\} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(y_i - \frac{m_y}{m_x}x_i\right)^2.$$

40.7 Сравним теперь (40.19) и (40.25). Следуя Гудмэну и Хартли (1958), разность можно записать в виде

$$n\mathsf{M}\{(\tilde{\mu}'_y - \mu_y)^2\} - n\mathsf{D}(\hat{\mu}'_y) =$$

$$= \mathsf{D}(x)\left\{\left(\frac{\mu_y}{\mu_x} - \frac{C(y, x)}{\mathsf{D}(x)}\right)^2 - \left(\mu_{y/x} - \frac{C(y, x)}{\mathsf{D}(x)}\right)^2\right\}. \qquad (40.26)$$

Из (40.26) понятно, что модифицированная оценка отношения средних $\tilde{\mu}'_y$ является более эффективной, чем модифицированная оценка среднего отношений $\hat{\mu}'_y$ или менее эффективной в зависимости от того, будет ли коэффициент линейной регрессии y на x, равный $\beta_{yx} = C(y, x)/\mathsf{D}(x)$, ближе к отношению средних, взятых по всей совокупности, или к взятому по совокупности среднему отношений. На практике первая ситуация является, по-видимому, более типичной, так что имеющая небольшое смещение оценка (40.9) является более предпочтительной, чем несмещенная оценка (40.7).

Оценка (40.9) более эффективна, чем обычное выборочное среднее m_y тогда и только тогда, когда правая часть в (40.25) меньше своего первого члена, т. е. если

$$\frac{\mu_y}{\mu_x}\left(\frac{\mu_y}{\mu_x} - 2\beta_{yx}\right) < 0, \qquad (40.27)$$

т. е.

$$\beta_{yx} > \frac{1}{2}\frac{\mu_y}{\mu_x} > 0$$

или

$$\beta_{yx} < \frac{1}{2}\frac{\mu_y}{\mu_x} < 0.$$

Таким образом, мы охарактеризовали эффективность оценки $\tilde{\mu}'_y$ по сравнению как с $\hat{\mu}'_y$, так и с m_y в терминах относительной величины регрессионного коэффициента β_{yx} и отношения взятых по совокупности средних μ_y/μ_x. Это совпадает с тем, что и следовало ожидать, поскольку существенным элементом оценки $\tilde{\mu}'_y$ является отношение выборочных средних m_y/m_x.

Олкин (1958) обобщил результаты о немодифицированной оценке $\tilde{\mu}_y$ на случай, когда x — вектор. См. также П. Рао и Мудхолкар. (1967).

40.8 Приближенно несмещенная оценка (40.9) была получена прямым оцениванием смещения $\tilde{\mu}_y$, вычисленного в (40.3). С другой стороны, можно уменьшить порядок величины смещения, используя метод Кенуя, описанный в **17.10** т. 2. Этот метод состоит в подсчете $\tilde{\mu}_y$ для каждой из n различных выборок объема $(n-1)$, получающихся исключением одного наблюдения, усреднении этих значений и применении (17.10) для получения модифицированной оценки. Тогда смещение имеет порядок n^{-2}, а дисперсия не меняется вплоть до порядка n^{-1} (см. упражнение 17.18). Как мы убедились, то же самое выполняется для $\tilde{\mu}_y'$ в (40.23).

Дербин (1959а) применил упрощенный вид метода Кенуя для модификации оценки-отношения общего типа, имеющей вид $r = t_y/t_x$ (частным случаем которого является (40.1)). Предполагалось, что смещение при оценке $\mathsf{M}(t_y)/\mathsf{M}(t_x)$ имеет порядок n^{-1}. Если ту же самую статистику r вычислить для первых $n/2$ и следующих $n/2$ наблюдений (n четное) и обозначить соответственно r_1 и r_2, то рассматривалась следующая модифицированная оценка:

$$t(r) = 2r - \frac{1}{2}(r_1 + r_2). \tag{40.28}$$

Если регрессия t_y на t_x линейна с постоянной дисперсией порядка n^{-1}, а само t_x распределено нормально с дисперсией порядка n^{-1}, то показано, что смещение оценки (40.28) имеет порядок n^{-2}, а дисперсия с точностью до n^{-1} совпадает с дисперсией r, но асимптотически *меньше* последней, если принять во внимание члены порядка n^{-2} и меньше. Аналогичный результат справедлив, когда t_x имеет гамма-распределение.

Дж. Рао (1965) и Дж. Рао и Уэбстер (1966) показали, что применение исходного метода Кенуя дает даже меньшее смещение и среднеквадратичную ошибку, как в случае нормального, так и в случае гамма-распределения. Дж. Рао (1967) показал, что эта оценка всегда имеет меньшую среднеквадратичную ошибку, чем (40.7), (40.9), (40.28) и другие оценки.

Приведенный выше результат является более общим, чем это кажется с первого взгляда, поскольку t_y и t_x в силу центральной предельной теоремы обычно имеют асимптотически двумерное нормальное распределение с дисперсиями порядка n^{-1}. По крайней мере это выполняется для m_y и m_x из (40.1), а значит, выполняются предположения линейной регрессии. Отсюда следует, что ничего не теряется при использовании смещенной оценки (40.28).

Простой числовой пример с $n = 2$, $N = 4$, рассматриваемый Гудмэном и Хартли (1958), Дербиным (1959а) и Нито де Паскуалем (1961), дает следующие результаты

Значения (x, y) в совокупности: (2,2), (2,6), (4,6) и (6,10)

Оценка	Уравнение	Среднеквадратичная ошибка
$t(\tilde{\mu}_y)$	(40.28), (40.1)	0,38
$\tilde{\mu}_y'$	(40.9)	0,44
$\hat{\mu}_y'$	(40.7)	0,56
$\tilde{\mu}_y$	(40.1)	0,92
$\hat{\mu}_y$	(40.2)	2,41
m_y	(выборочное среднее)	2,67

(40.29)

В упражнении 40.17 читателю предлагается проверить правильность приведенных значений.

40.9 Если переписать (40.1) в виде

$$\frac{\tilde{\mu}_y}{\mu_x} = \frac{m_y}{m_x} = \frac{\mu_y}{\mu_x}\left(1 + \frac{m_y - \mu_y}{\mu_y}\right)\left(1 + \frac{m_x - \mu_x}{\mu_x}\right)^{-1} \tag{40.30}$$

и разложить $\left(1 + \frac{m_x - \mu_x}{\mu_x}\right)^{-1}$ в ряд Тейлора, сходящийся с вероятностью единица при N, $n \to \infty$, то, взяв математическое ожидание, получим

$$\mathsf{M}\left(\frac{m_y}{m_x}\right) = \frac{\mu_y}{\mu_x}\left\{1 + \left(\frac{1}{n} - \frac{1}{N}\right)\left(\frac{\mathsf{D}(x)}{\mu_x^2} - \frac{C(y, x)}{\mu_y \mu_x}\right) + O\left(\frac{1}{M^2}\right)\right., \tag{40.31}$$

где M может относиться как к N, так и к n. А значит, оценка

$$u = \frac{m_y}{m_x}\left\{1 - \left(\frac{1}{n} - \frac{1}{N}\right)\left(\frac{s_x^2}{m_x^2} - \frac{s_{xy}}{m_y m_x}\right)\right\} \tag{40.32}$$

с точностью до первого порядка устраняет смещение оценки m_y/m_x. В (40.32) выборочные дисперсия и ковариация, как обычно, имеют делителем $(n-1)$.

Теперь можно непосредственно, хотя это и несколько громоздко, вычислить среднее значение и дисперсию оценки u, используя следующие соотношения (первые три из которых приведены в (12.117), (12.119) и (12.121), т. 1, а остальные выво-

дятся методами примера 13.2):

$$\left.\begin{aligned}
\mathbf{M}(m_x-\mu_x)^3&=\left(\alpha_2-\frac{3\alpha_1}{N}\right)K_{20},\\
\mathbf{M}(m_x-\mu_x)^4&=3\alpha_1^2K_{20}^2+O(n^{-3}),\\
\mathbf{M}[(m_x-\mu_x)(s_x^2-K_{20})]&=\alpha_1K_{30},\\
\mathbf{M}[(m_x-\mu_x)^2(m_y-\mu_y)]&=\left(\alpha_2-\frac{3\alpha_1}{N}\right)K_{21},\\
\mathbf{M}[(m_x-\mu_x)^2(m_y-\mu_y)^2]&=\alpha_1^2(2K_{11}^2+K_{20}K_{02})+O(n^{-3}),\\
\mathbf{M}[(m_x-\mu_x)^3(m_y-\mu_y)]&=3\alpha_1^2K_{20}K_{11}+O(n^{-3}),\\
\mathbf{M}[(s_x^2-K_{20})(m_y-\mu_y)]&=\alpha_1K_{21},\\
\mathbf{M}[(m_x-\mu_x)(s_{xy}-K_{11})]&=\alpha_1K_{21},\\
\mathbf{M}[(s_{xy}-K_{11})(m_y-\mu_y)]&=\alpha_1K_{12}.
\end{aligned}\right\}\qquad(40.33)$$

Здесь $\alpha_r=(n^{-r}-N^{-r})$, как и в (12.116), причем как в этой, так и в предыдущей главе мы опускаем индекс N у $\mathbf{M}$. Тин (1965) приводит соответствующие значения с точностью до n^{-2}:

$$\mathbf{M}(u)=\left(\frac{\mu_y}{\mu_x}\right)\left\{1-\left(2\alpha_2-\frac{3\alpha_1}{N}\right)(C_{21}-C_{30})-3\alpha_1^2C_{20}(C_{20}-C_{11})\right\},\qquad(40.34)$$

$$\mathbf{D}(u)=\left(\frac{\mu_y}{\mu_x}\right)^2\left\{\alpha_1(C_{20}+C_{02}-2C_{11})+\right.$$
$$\left.+\alpha_1^2(2C_{20}^2-4C_{20}C_{11}+C_{11}^2+C_{20}C_{02})+\frac{2\alpha_1}{N}(C_{30}-2C_{21}+C_{12})\right\},\qquad(40.35)$$

где

$$C_{rs}=K_{rs}/(\mu_x^r\mu_y^s).$$

Точно таким же образом Тин (1965) получает результаты для исходной оценки $\tilde{\mu}_y/\mu_x=m_y/m_x$:

$$\mathbf{M}\left(\frac{m_y}{m_x}\right)=\left(\frac{\mu_y}{\mu_x}\right)\left\{1+\alpha_1(C_{20}-C_{11})+\left(\alpha_2-\frac{3\alpha_1}{N}\right)(C_{21}-C_{30})+\right.$$
$$\left.+3\alpha_1^2C_{20}(C_{20}-C_{11})\right\},\qquad(40.36)$$

$$\mathbf{D}\left(\frac{m_y}{m_x}\right)=\left(\frac{\mu_y}{\mu_x}\right)^2\left\{\alpha_1(C_{20}+C_{02}-2C_{11})+\right.$$
$$+\alpha_1^2(8C_{20}^2-16C_{20}C_{11}+5C_{11}^2+3C_{20}C_{02})-$$
$$\left.-2\left(\alpha_2-\frac{3\alpha_1}{N}\right)(C_{30}-2C_{21}+C_{12})\right\}\qquad(40.37)$$

в то время как для $t\left(\frac{m_y}{m_x}\right)$, определяемого (40.28), он находит

$$\mathbf{M}\left\{t\left(\frac{m_y}{m_x}\right)\right\}=\left(\frac{\mu_y}{\mu_x}\right)\left\{1-(C_{20}-C_{11})/N-2\alpha_2(C_{21}-C_{30})+\right.$$
$$\left.+3\left(\frac{2}{n^2}-\frac{1}{N^2}\right)C_{20}(C_{20}-C_{11})\right\},\qquad(40.38)$$

$$\mathbf{D}\left\{t\left(\frac{m_y}{m_x}\right)\right\}=\left(\frac{\mu_y}{\mu_x}\right)^2\left\{\alpha_1(C_{20}+C_{02}-2C_{11})+\right.$$
$$+2\left(\frac{2}{n^2}-\frac{5}{nN}+\frac{4}{N^2}\right)C_{20}(C_{20}-2C_{11})+\left(\frac{2}{n^2}-\frac{6}{nN}+\frac{5}{N^2}\right)C_{11}^2+$$
$$\left.+\left(\frac{2}{n^2}-\frac{4}{nN}+\frac{3}{N^2}\right)C_{20}C_{02}+\frac{4\alpha_1}{N}(C_{30}-2C_{21}+C_{12})\right\}.\qquad(40.39)$$

Из этих результатов понятно, что смещения оценок u и $t(m_y/m_x)$ очень малы и не содержат члена, имеющего порядок n^{-1}, в отличие от смещения оценки m_y/m_x. Все три дисперсии имеют один и тот же главный член, который мы уже подсчитали в (40.24). При этом мы видели, что он является главным членом и для среднеквадратичной ошибки оценки $\tilde{\mu}_y'/\mu_x$, определенной в (40.9). В упражнении 40.13 читателю предлагается показать, что для приближений следующего порядка выполняется

$$\mathbf{D}(u)<\mathbf{D}\left\{t\left(\frac{m_y}{m_x}\right)\right\}<\mathbf{D}\left(\frac{m_y}{m_x}\right),\qquad(40.40)$$

а значит, u, определяемая (40.32), оказывается более предпочтительной, чем другие рассмотренные здесь оценки, как из соображений смещения, так и из соображений среднеквадратичной ошибки.

Тин (1965) рассматривает также другие оценки, близко связанные с u (см. упражнение 40.14), и проводит сравнение этих оценок в случае двумерного нормального распределения и в других случаях.

Регрессионные оценки

40.10 При условии, что мы знаем среднее по совокупности μ_x дополнительной переменной, как это было в **40.2**, естественно рассмотреть применение теории регрессии для повышения эффективности оценивания μ_y. Простейшей является оценка с линейной регрессией

$$\check{\mu}_y=m_y+b(\mu_x-m_x).\qquad(40.41)$$

Эта оценка обобщает (40.1). Она сводится к последней, если положить $b=m_y/m_x$. Однако обычно b выбирается как коэффициент регрессии y по x по методу НК.

Если бы зависимость между y и x подчинялась предположениям линейной модели (19.8), то можно было бы применять теорию метода НК из главы 19 т. 2 и из пунктов, следующих за 28.12. Однако в изучаемой ситуации мы имеем дело с конечной совокупностью и во всяком случае во многих применениях x является случайной величиной, так что применение метода НК не является строго обоснованным. Вместо этого заметим, что совместное распределение m_y и m_x имеет средние μ_y, μ_x, дисперсии $\mathsf{D}(x)/n$, $\mathsf{D}(y)/n$ и ковариацию $C(y, x)/n$. А значит, из (40.41) следует, что если не учитывать обусловленную выборкой ошибку в значении b вне зависимости от того, каким способом мы получили это b, то

$$n\mathsf{D}(\check{\mu}_y) = \mathsf{D}(y) + b^2\mathsf{D}(x) - 2bC(y, x). \qquad (40.42)$$

Соответствующая несмещенная оценка имеет вид

$$\hat{\mathsf{D}}(\check{\mu}_y) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\{(y_i - m_y) - b(x_i - m_x)\}^2. \qquad (40.43)$$

Приведенная ниже соотношения (40.25) асимптотическая формула для оценки среднеквадратичной ошибки $\tilde{\mu}'_y$ (совпадающей асимптотически с дисперсией $\tilde{\mu}'_y$ и с дисперсией немодифицированной оценки $\tilde{\mu}_y$), получается из (40.43), если, как и выше, положить $b = m_y/m_x$. В силу центральной предельной теоремы m_y и m_x обычно асимптотически нормальны, а значит, $\check{\mu}_y$ также асимптотически нормальна.

40.11 Соотношение (40.42) представляет собой только асимптотическое равенство, поскольку только тогда можно пренебречь обусловленной выборкой ошибкой в значении b. Мы видим, что регрессионная оценка $\check{\mu}_y$ более эффективна, чем оценка m_y, равная выборочному среднему, если правая часть в (40.42) меньше, чем ее первый член, т. е. если $2bC(y, x) > b^2\mathsf{D}(x)$.

Если выбрать b по методу НК, то при $n \to \infty$ b стремится к $C(y, x)/\mathsf{D}(x)$, а значит, приведенное условие всегда выполняется, если $C(y, x) \neq 0$. Аналогично, сравнивая (40.42) с (40.25), убеждаемся, что условием, гарантирующим большую эффективность $\check{\mu}_y$ по сравнению с $\tilde{\mu}'_y$, является

$$\mathsf{D}(x)\left\{b^2 - \left(\frac{\mu_y}{\mu_x}\right)^2\right\} < 2C(y, x)\left(b - \frac{\mu_y}{\mu_x}\right). \qquad (40.44)$$

Если b получается по методу НК и стремится к $C(y, x)/\mathsf{D}(x)$, то (40.44) сводится асимптотически к

$$\mathsf{D}(x)\left(b - \frac{\mu_y}{\mu_x}\right)^2 > 0. \qquad (40.45)$$

Но это не выполняется только при $b = \mu_y/\mu_x$, когда (40.41) асимптотически сводится к (40.1). Таким образом, по крайней мере асимптотически, мы ничего не теряем, используя регрессионную оценку, полученную методом НК.

40.12 Регрессионная оценка (40.41), разумеется, является смещенной. Для устранения этого смещения рассмотрим общие методы построения несмещенных оценок, предложенные Микки (1959) и У. Уилльямсом (1961, 1962). Эти методы сделают также более понятными рассмотренные нами ранее вопросы с оценками-отношениями.

Несмещенные оценки с дополнительными переменными

40.13 Для начала заметим, что для любой константы a оценка

$$m_y - a(m_x - \mu_x) \qquad (40.46)$$

является несмещенной оценкой μ_y, однако это, вообще говоря, не так, если a — статистика, вычисленная по той же самой выборке, что и m_x. Предположим теперь, что выборка из n наблюдений разбита случайным образом на подвыборку из p наблюдений и «оставшуюся» выборку из $n-p$ наблюдений. (Чтобы быть более точным, можно в качестве подвыборки взять первые p наблюдений в порядке их извлечения). Используем теперь подвыборку для определения a в (40.46), а средние m_y и m_x вычислим только для оставшейся выборки. Поскольку оставшаяся выборка является случайной выборкой из совокупности, содержащей $N-p$ членов, не вошедших в подвыборку, то в результате мы получим несмещенную оценку для среднего «оставшейся» совокупности. Более того, средние значения по оставшейся выборке и по оставшейся совокупности мы можем выразить в терминах средних по всей выборке и по всей совокупности и средних по подвыборке. Для различения мы будем писать последние с аргументом (p). Таким образом, (40.46) приводит к оценке

$$u_p = \frac{nm_y - pm_y(p)}{n-p} - a(p)\left\{\frac{nm_x - pm_x(p)}{n-p} - \frac{N\mu_x - pm_x(p)}{N-p}\right\},$$

которая является несмещенной оценкой для $\{N\mu_y - pm_y(p)\}/(N-p)$. В результате несмещенной оценкой для самого μ_y будет $\{(N-p)u_p + pm_y(p)\}/N$. Эту оценку мы запишем в виде

$$t_p = \frac{N-p}{N}\frac{n}{n-p}\{m_y - a(p)(m_x - \mu_x)\} - \\ - \frac{N-n}{N}\frac{p}{n-p}\{m_y(p) - a(p)(m_x(p) - \mu_x)\}. \qquad (40.47)$$

Мы можем выбрать произвольное число p такое, что $1 \leqslant p \leqslant n-1$. При данном p функция $a(p)$, определяемая по подвыборке, также произвольна. В результате мы получили большой класс несмещенных оценок параметра μ_y, основанных на нашем знании параметра μ_x.

Те же самые рассуждения справедливы и в многомерном случае, когда x — вектор.

40.14 Общий класс оценок (40.47) обладает следующим не очень удобным свойством: они зависят от порядка, в котором проводятся извлечения в выборке. Можно избавиться от этого, рассматривая t_p для каждой из $n!$ возможных упорядочений выборки и усредняя их. Получающаяся в результате усредненная оценка $\bar{t}_p$ иногда имеет простой вид и не требует больших выкладок при ее вычислении по выборочным значениям. (Этот процесс усреднения в точности совпадает с проведенным нами в **39.11** по тем же причинам. Однако там результаты не получались простыми для вычислений, поскольку выбор проводился с неравными вероятностями). Упражнение 39.30, доказательство которого в рассматриваемых условиях упрощается, поскольку выбор проводится с равными вероятностями, показывает, что усредненная оценка $\bar{t}_p$ имеет дисперсию не большую, чем любая отдельная оценка t_p.

40.15 Если в (40.47) положить $a(p) = m_{y/x}(p) = \frac{1}{p}\sum_{l=1}^{p}\frac{y_l}{x_l}$ и $p=1$, то оценка сводится к

$$t_1 = \mu_x \frac{y_1}{x_1} + \frac{N-1}{N}\cdot\frac{n}{n-1}\left(m_y - \frac{y_1}{x_1}m_x\right),$$

где y_1 и x_1 — значения, соответствующие первому наблюдению. Усредняя по всем $n!$ возможным упорядочениям выборки, получим оценку

$$\bar{t}_1 = \mu_x m_{y/x} + \frac{N-1}{N}\frac{n}{n-1}(m_y - m_{y/x}m_x), \qquad (40.48)$$

которая совпадает с $\hat{\mu}'_y$, определенной в (40.7). Более того, если мы возьмем любое другое значение p и то же самое $a(p)$, что и выше, то усредненное значение $\bar{t}_p$ будет тем же самым, что и в (40.48), хотя, конечно, само t_p будет отличаться. Таким образом, если $a(p) = m_{y/x}(p)$, то мы получаем несмещенную оценку, основанную на $m_{y/x}$.

Эти результаты подсказывают нам, что точный несмещенный вариант оценки типа (40.1) разумно искать, положив $a(p) = m_y(p)/m_x(p)$. В этом случае (40.47) принимает вид

$$t_p = \mu_x \frac{m_y(p)}{m_x(p)} + \frac{(N-p)}{N}\frac{n}{(n-p)}\left(m_y - \frac{m_y(p)}{m_x(p)}m_x\right),$$

что после усреднения имеет тот же вид с $m_y(p)/m_x(p)$, замененным на его среднее значение. При $p=1$ это, конечно, совпадает с (40.48), поскольку в этом случае $a(p)$ совпадает с рассмотренным ранее. Другим простейшим выбором является $p=n-1$, когда среднее значение величины $m_y(p)/m_x(p)$, взятое по всем перестановкам, сводится к $\frac{1}{n}\sum_{i=1}^{n}\left(\frac{nm_y - y_i}{nm_x - x_i}\right) = R$. В этом случае

$$\bar{t}_{n-1} = \mu_x R + \frac{(N-p)}{N}\frac{n}{(n-p)}(m_y - Rm_x) \qquad (40.49)$$

является несмещенной оценкой μ_y.

40.16 Обращаясь теперь к регрессионной оценке, естественно рассмотреть $a(p) = b_{yx}(p)$. Тогда (40.47) сводится к

$$t_p = m_y - b_{yx}(p)(m_x - \mu_x) - \\ - \frac{(N-n)}{N}\frac{p}{(n-p)}\{(m_y(p) - m_y) - b_{yx}(p)(m_x(p) - m_x)\}. \qquad (40.50)$$

Усреднение приводит просто к замене $b_{yx}(p)$ на соответствующее среднее значение. Значение $p=1$ теперь невозможно, поскольку тогда b_{yx} не имеет смысла. Как и ранее, другим простейшим случаем является $p=n-1$. В этом случае нужно n раз вычислить коэффициент регрессии, каждый раз не принимая во внимание последовательно одно из наблюдений, а затем усреднить, чтобы получить $\bar{b}_{yx}(n-1)$. Тогда (40.50) сведется к

$$t_{n-1} = m_y - \bar{b}_{yx}(n-1)(m_x - \mu_x) - \\ - \frac{N-n}{Nn}\left\{\sum_{i=1}^{n} b_{yx}(n-1)x_i - n\bar{b}_{yx}(n-1)m_x\right\}, \qquad (40.51)$$

где x_i под знаком суммы соответствует значению наблюдения, не принятого во внимание при подсчете того $b_{yx}(n-1)$, на которое x_i множится. Оценка (40.51) эквивалентна обычной регрессионной оценке (40.41), если все $b_{yx}(n-1)$ совпадают. Но в общем случае это не так. Однако если n велико, то $b_{yx}(n-1)$ может меняться очень мало и соответственно оценки мало отличаются.

Оценивание дисперсии

40.17 Для несмещенной оценки (40.47), вообще говоря, нельзя выписать значение дисперсии, поскольку все зависит от выбора $a(p)$. Однако если несколько видоизменить схему оценивания, то можно получить такую оценку, дисперсия которой тоже может быть оценена.

Предположим, что n наблюдений разбиты на k подвыборок в порядке их извлечения, так что r-я подвыборка содержит n_r наблюдений и $\sum_{r=1}^{k} n_r = n$. Обозначим частные суммы $\sum_{r=1}^{q} n_r = n_{+q}$, так что $n_{+1} = n_1$, а $n_{+k} = n$. Оценку (40.47) будем обозначать теперь $t(p, n)$, чтобы подчеркнуть, что подвыборка объема p взята из выборки объема n. Рассмотрим последовательность, состоящую из $k-1$ оценок $t(n_{+1}, n_{+1}), t(n_{+2}, n_{+3}), \ldots, t(n_{+(k-1)}, n_{+k})$. В каждой из этих оценок в качестве подвыборки берется вся выборка предыдущей оценки. Все эти оценки не коррелированы, поскольку математическое ожидание может быть представлено в виде последовательности условных математических ожиданий (см. **39.35**) в соответствии с разбиением на k подгрупп и для $r < s$

$$\mathbf{M}\{t(n_{+r}, n_{+(r+1)})\, t(n_{+s}, n_{+(s+1)})\} =$$
$$= \underset{1}{\mathbf{M}} \ldots \underset{r+1}{\mathbf{M}} \{t(n_{+r}, n_{+(r+1)}) \underset{r+2}{\mathbf{M}} \ldots \underset{k}{\mathbf{M}} [t(n_{+s}, n_{+(s+1)})]\} =$$
$$= \underset{1}{\mathbf{M}} \ldots \underset{r+1}{\mathbf{M}} \{t(n_{+r}, n_{+(r+1)})\, \mu_y\} = \mu_y^2.$$

Теперь, чтобы получить оценку дисперсии среднего значения последовательности несмещенных некоррелированных оценок $t(n_{+r}, n_{+(r+1)})$, можно воспользоваться результатом упражнения 39.12. При этом сами оценки не обязаны иметь такой же вид.

40.18 В **40.17** мы не требовали, чтобы k подвыборок были одного объема. Предположим теперь, что это выполняется, так что $n_r = n/k$. Если каждую из подвыборок использовать последовательно для вычисления $a(p) = a(n/k)$ в (40.47) и каждый раз вычислять соответствующее $t(n/k, n)$, то эти k значений не будут теперь некоррелированными, как это было в **40.17**. Их среднее значение равно

$$T\left(\frac{n}{k}, n\right) = \frac{1}{k}\sum^{k} t\left(\frac{n}{k}, n\right) = m_y - \bar{a}\left(\frac{n}{k}\right)(m_x - \mu_x) +$$
$$+ \frac{N-n}{Nk(k-1)}\sum^{k}\left\{a\left(\frac{n}{k}\right) - \bar{a}\left(\frac{n}{k}\right)\right\}\left\{m_x\left(\frac{n}{k}\right) - m_x\right\}, \qquad (40.52)$$

где черта над буквой означает усреднение по k полученным значениям. Первые два члена в правой части (40.52) имеют в точности такой вид, как и (40.46), но теперь они не являются несмещенной оценкой, поскольку $\bar{a}(n/k)$ вычислено на основе той же самой выборки, что и m_x. Математическое ожидание этих двух членов равно $\mu_y - C\{\bar{a}(n/k), m_x\}$. Последний член в (40.52), очевидно, является несмещенной оценкой для этой ковариации,

если рассматривать совокупность как состоящую из $\frac{N}{(n/k)}$ групп объема n/k, и считать, что случайным образом проводится выборка k таких групп.

Модификация схемы выбора для устранения смещения

40.19 До сих пор в этой главе все наши рассуждения о регрессионных оценках и оценках-отношениях были направлены на то, чтобы видоизменить вид оценок для устранения или уменьшения смещения, при выборе без возвращения с равными вероятностями. Другой способ достижения той же цели состоит в изменении схемы выбора так, чтобы сами исходные оценки становились несмещенными.

Из (39.20) следует, что при любых вероятностях извлечения π_i выражение $\frac{1}{N}\sum_{i=1}^{n}\frac{y_i}{\pi_i}$ задает несмещенную оценку μ_y. Предположим теперь, что мы взяли

$$\pi_i = nx_i \Big/ \sum_{i=1}^{N} x_i, \qquad (40.53)$$

причем теперь нужно предполагать, что вспомогательная переменная x положительна, чтобы (40.53) задавало соответствующее множество вероятностей. В этом случае сразу получим, что $\mu_x m_{y/x}$ будет несмещенной оценкой μ_y. Иными словами, для такой схемы выбора оценка $\hat{\mu}_y$, определенная в (40.2), является в точности несмещенной. С другой стороны, можно рассматривать множество из $\binom{N}{n}$ возможных выборок как совокупность, из которой должен быть извлечен один элемент. Пусть $(m_y)_i$ и $(m_x)_i$, $i = 1, \ldots, \binom{N}{n}$, средние для соответствующих выборок. Если упомянутое одно извлечение провести с вероятностями

$$\pi_i = (m_x)_i \Big/ \sum_{j=1}^{\binom{N}{n}} (m_x)_j = (m_x)_i \Big/ \left\{\binom{N}{n}\mu_x\right\}, \qquad (40.54)$$

то рассуждения, аналогичные вышеприведенным, показывают, что оценка (39.20) принимает вид

$$\frac{1}{\binom{N}{n}} \frac{m_y}{m_x \Big/ \left\{\binom{N}{n}\mu_x\right\}} = \mu_x m_y / m_x,$$

так что, как впервые отметил Лахири (1951), оценка $\tilde{\mu}_y$, определенная в (40.1), становится в точности несмещенной.

Дисперсии этих оценок и несмещенные оценки этих дисперсий получаются, как обычно, из (39.23—4). Разумеется, для того чтобы в случае $\tilde{\mu}_y$ можно было оценить дисперсии, нужны по крайней мере две выборки (или случайная подвыборка из одной выборки).

Нанджамма, Мёрфи и Сэти (1959) рассматривают общую задачу модификации схемы выбора, чтобы сделать оценки-отношения несмещенными с применением к некоторым типам выборочных обследований. См. также Патхак (1964а).

Расслоенный и многошаговый выбор

40.20 Как оценки-отношения, так и регрессионные оценки можно применять отдельно к каждому из некоторого числа слоев при условии, что для каждого слоя известно среднее по совокупности переменной x. С другой стороны, конкретную оценку-отношение, или регрессионную оценку, можно применить, используя объединение результатов для всех слоев. Следует ожидать, что первая из упомянутых процедур будет, вообще говоря, более эффективной. Кокрэн (1963) рассматривает более подробно смещенные оценки-отношения и регрессионные оценки.

Несмещенные расслоенные оценки-отношения рассматривали в одномерном случае Нито де'Паскуаль (1961) и У. Уилльямс (1961), а в многомерном — Олкин (1958). Робсон и Витхаясаи (1961) рассматривали ситуацию, аналогичную расслоенной, когда y и x могут быть представлены в виде суммы k соответствующих компонент. Киш и Гесс (1959) получили асимптотическую формулу для дисперсии смещенной комбинированной оценки-отношения при расслоенном многошаговом выборе.

40.21 Дербин (1953) указал, что поскольку из (40.30)

$$\frac{m_y}{m_x} = \frac{\mu_y}{\mu_x} + \frac{m_y - \left(\frac{\mu_y}{\mu_x}\right) m_x}{\mu_x} + o(n^{-1/2}), \tag{40.55}$$

то отношения выборочных средних асимптотически линейны по $z = y - (\mu_y/\mu_x)x$, так что

$$\frac{m_y}{m_x} - \frac{\mu_y}{\mu_x} \doteqdot \frac{1}{n}\sum_{i=1}^{n} z_i/\mu_x. \tag{40.56}$$

Отсюда следует, что для оценки $\mathbf{D}\left(\frac{m_y}{m_x}\right)$ в случае многошагового выбора асимптотически применимы рассуждения **39.45—50** и в случае больших выборок можно применять правила Иэйтса — Дербина из **39.46** и **39.49**. Это справедливо также внутри каждого слоя многошаговой схемы. То же самое можно применить к регрессионным оценкам.

Двухэтапная выборка

40.22 Как при обсуждении расслоенного выбора в **39.15—27**, так и при использовании вспомогательной переменной для улучшения эффективности оценки в этой главе предполагалось, что мы обладаем некоторыми знаниями о совокупности, позволяющими нам строить несмещенные оценки. В последнем случае мы предполагали, что нам известно μ_x, а в первом — относительный объем слоев N_l/N, который нужен для вычисления оценки (39.38). Если у нас нет соответствующей информации, то в практических ситуациях сама собой напрашивается следующая процедура. Провести предварительный случайный выбор с равными вероятностями для получения такой информации, а затем использовать ее при проведении основной выборки, направленной на достижение исходной цели — оценивание среднего совокупности. Очевидно, что такая процедура будет экономически приемлемой только в том случае, когда затраты на предварительную выборку малы по сравнению с результирующей выгодой в эффективности, получающейся в основной выборке. Ниже мы сформулируем это утверждение более точно.

Схема выбора такого вида называется двухэтапной выборкой (two phase sampling). (Мы не употребляем применявшегося ранее термина «двойная выборка» («double sampling»), поскольку этот термин мы уже использовали в главе 34 для обозначения последовательного метода, целью которого было получение доверительного интервала заданной длины. Сейчас у нас иная цель. Мы хотим в первую очередь улучшить эффективность оценки на втором этапе, получив на первом этапе нужную информацию). Двухэтапная выборка отличается от двухшагового выбора тем, что в нем на каждом этапе извлекаются одни и те же объекты.

40.23 Следуя Нейману (1938), который впервые решил соответствующую задачу, рассмотрим сначала задачу расслоения. Мы хотим провести разбиение на фиксированное число k заданных слоев, но не знаем пропорции $N_l/N = W_l$, в которых элементы совокупности входят в эти слои, а значит, не можем использовать оценку (39.38). В соответствии с этим мы делаем предварительную выборку объема n_1 с равными вероятностями Пусть частоты попадания в соответствующие k слоев оказались $n_{11}, \ldots, n_{1k}$, где $\sum_{l=1}^{k} n_{1l} = n_1$. Тогда пропорции $w_{1l} = n_{1l}/n_1$

являются, разумеется, несмещенными оценками пропорций W_l в совокупности. Поэтому в качестве оценки параметра μ естественно взять

$$\hat{\mu}_{12} = \sum_{l=1}^{k} w_{1l} m_{2l}, \tag{40.57}$$

где m_{2l} — выборочные средние слоев во второй (основной) выборке, имеющей объем n_2 и n_{2l} наблюдений в l-м слое. Теперь возникает вопрос, должны ли n_{2l} быть подвыборками n_{1l} или они должны извлекаться независимо. Для практических ситуаций больше подходит первый случай, поскольку, если N_l неизвестно, то никакого полного описания слоев быть не может и выбор на втором этапе проводится на основе случайной выборки из каждого слоя, полученной на первом этапе. Более того, хотя первый этап логически предшествует второму, но если на втором этапе используются подвыборки первого, то оба этапа могут иногда проводиться одновременно.

Предположим, во всяком случае, что n_{1l} настолько велики по сравнению с предполагаемыми значениями n_{2l}, что наблюденные значения n_{1l} заведомо больше, чем фиксированные заранее значения n_{2l}.

40.24 Используем теперь (см. **39.35**) символ $\underset{2}{\mathsf{M}}$ для обозначения математического ожидания по второму этапу при условии, что результат первого этапа фиксирован, а через $\underset{1}{\mathsf{M}}$ обозначим математическое ожидание по первому этапу. Используя (39.72) получим, что математическое ожидание оценки (40.57) равно

$$\mathsf{M}(\hat{\mu}_{12}) = \underset{1}{\mathsf{M}}\{\underset{2}{\mathsf{M}}(\hat{\mu}_{12})\} = \sum_l \underset{1}{\mathsf{M}}\{w_{1l}\underset{2}{\mathsf{M}}(m_{2l})\} =$$
$$= \sum_l \underset{1}{\mathsf{M}}\{w_{1l}\}\mu_l = \sum_l W_l\mu_l = \mu, \tag{40.58}$$

так что $\hat{\mu}_{12}$ — несмещенная оценка. Согласно (39.73) ее дисперсия равна

$$\mathsf{D}(\hat{\mu}_{12}) = \underset{1}{\mathsf{M}}\{\underset{2}{\mathsf{D}}(\hat{\mu}_{12})\} + \underset{1}{\mathsf{D}}\{\underset{2}{\mathsf{M}}(\hat{\mu}_{12})\}. \tag{40.59}$$

Первый член в правой части (40.59) можно вычислить, заметив, что так же, как в (39.39),

$$\underset{2}{\mathsf{D}}(\hat{\mu}_{12}) = \underset{2}{\mathsf{D}}\left(\sum_l w_{1l}m_{2l}\right) = \sum_l w_{1l}^2 \frac{\sigma_l^2}{n_{2l}}\left(1 - \frac{n_{2l}}{N_l}\right),$$

а значит,

$$\underset{1}{\mathsf{M}}\{\underset{2}{\mathsf{D}}(\hat{\mu}_{12})\} = \sum_l \underset{1}{\mathsf{M}}(w_{1l}^2)\frac{\sigma_l^2}{n_{2l}}\left(1 - \frac{n_{2l}}{N_l}\right). \tag{40.60}$$

Если предположить теперь, что все N_l очень велики, то последним множителем в правой части (40.60) можно пренебречь. Более того, выборка на первом этапе дает оценки вероятностей в мультиномиальной схеме, так что применимо (5.80), и (40.60) принимает вид

$$\underset{1}{\mathsf{M}}\{\underset{2}{\mathsf{D}}(\hat{\mu}_{12})\} = \sum_l \left\{\frac{W_l(1-W_l)}{n_1} + W_l^2\right\}\frac{\sigma_l^2}{n_{2l}}. \tag{40.61}$$

Используя (10.16) и (5.80), получим, что второй член в правой части (40.59) равен

$$\underset{1}{\mathsf{D}}\{\underset{2}{\mathsf{M}}(\hat{\mu}_{12})\} = \underset{1}{\mathsf{D}}\left\{\sum_l w_{1l}\mu_l\right\} =$$
$$= \sum_l \mu_l^2 \underset{1}{\mathsf{D}}(w_{1l}) + \sum_{\substack{l,p\\ l\neq p}}\sum \mu_l\mu_p \underset{1}{C}(w_{1l}, w_{1p}) =$$
$$= \sum_l \mu_l^2 \frac{W_l(1-W_l)}{n_1} - \sum_{\substack{l,p\\ l\neq p}}\sum \mu_l\mu_p \frac{W_lW_p}{n_1} =$$
$$= \frac{1}{n_1}\left\{\sum_l W_l\mu_l^2 - \left(\sum_l W_l\mu_l\right)^2\right\}. \tag{40.62}$$

Подставляя (40.61—2) в (40.59) и используя равенство $\sum_l W_l\mu_l = \mu$, получим

$$\mathsf{D}(\hat{\mu}_{12}) = \sum_l \left\{\frac{W_l(1-W_l)}{n_1} + W_l^2\right\}\frac{\sigma_l^2}{n_{2l}} + \frac{1}{n_1}\sum_l W_l(\mu_l - \mu)^2. \tag{40.63}$$

Напомним, что согласно (39.46) и **39.19**, последний член в правой части (40.63) выражает выигрыш в точности расслоенного выбора при ВРД по сравнению с нерасслоенным выбором, когда объемы всех слоев велики и объем выборки равен n_1.

40.25 Если $n_1 \to \infty$, так что W_l, по существу, можно считать известными, то (40.63) сводится к обычной формуле (39.39) для дисперсии при расслоенном выборе. Даже при не очень больших n_1 членом $W_l(1-W_l)/n_1$, который не превосходит $1/(4n_1)$, можно обычно пренебречь по сравнению с W_l^2. Вспоминая из **39.18**, что при больших N_l справедливо $\sum_l W_l(\mu_l - \mu)^2 \doteq \sigma^2 - \sum_l W_l\sigma_l^2$, можно (40.63) приближенно переписать в виде

$$\mathsf{D}(\hat{\mu}_{12}) \doteq \frac{\sigma^2 - \sum_l W_l\sigma_l^2}{n_1} + \sum_l W_l^2\frac{\sigma_l^2}{n_{2l}}. \tag{40.64}$$

Почти несмещенная оценка для (40.64) получается, если подставить w_{1l} вместо W_l, s^2 вместо σ^2 и s_l^2 вместо σ_l^2.

40.26 Предположим теперь, что функция затрат при двухэтапном выборе имеет вид

$$C = c_0 + n_1 c_1 + \sum_{l=1}^{k} n_{2l} c_{2l}. \qquad (40.65)$$

Соотношения (40.64—5) имеют тот же самый вид, что и (39.50—1). Но тогда из (39.53) следует, что для объемов выборок, минимизирующих $\mathsf{D}(\hat{\mu}_{12})$ при фиксированном C (или наоборот), имеет место

$$n_1^2 \propto \frac{\sigma^2 - \sum W_l \sigma_l^2}{c_1}, \quad n_{2l}^2 \propto \frac{W_l^2 \sigma_l^2}{c_{2l}}; \qquad (40.66)$$

при этом коэффициент пропорциональности получается из (40.65) или (40.64) в зависимости от того, какая из этих величин фиксируется.

Из (40.66) следует, что на втором этапе наблюдения должны быть распределены между слоями так же, как и при обычном расслоении в (39.55) (при этом следует помнить, что этот простой результат получился в результате пренебрежения членами порядка n_1^{-1}). Объем выборки на первом этапе прямо пропорционален числителю первого члена в правой части (40.64) (который соответствует добавке к дисперсии, обусловленной оцениванием на первом этапе) и обратно пропорционален затратам на проведение выбора. Оба эти утверждения согласуются с интуицией.

40.27 Двухэтапная выборка проводилась для улучшения эффективности оценки с помощью расслоения. Напомним, однако, что, согласно **39.18**, если даже W_l в точности известны, наилучшее расслоение все равно может привести к потере эффективности. Впрочем, как мы видели в **39.19**, это не так, если все N_l достаточно велики. Однако дополнительная компонента к значению дисперсии, обусловленная оцениванием W_l на первом этапе выборки, может привести теперь к потере эффективности даже при больших значениях N_l.

Пусть после извлечения с равными вероятностями нерасслоенной выборки объема n оказалось, что n_l наблюдений принадлежит l-му слою. Естественно предположить, что накладные расходы при проведении такой выборки равны тому же самому c_0, что и в (40.65), и что затраты на проведение наблюдения в l-м слое также остаются неизменными и равными c_{2l}. В этом случае функция затрат равна

$$C_R = c_0 + \sum_l n_l c_{2l},$$

причем теперь n_l — случайные величины с математическим ожиданием nW_l, так что нужно рассматривать математическое ожидание затрат

$$\mathsf{M}(C_R) = c_0 + n \sum W_l c_{2l}. \qquad (40.67)$$

Если $\mathsf{M}(C_R)$ должно равняться значению C из (40.65), то должно выполняться равенство

$$n = \frac{n_1 c_1 + \sum_l n_{2l} c_{2l}}{\sum_l W_l c_{2l}},$$

а значит, дисперсия нерасслоенной оценки будет (при больших N)

$$\mathsf{D}(m_R) = \frac{\sigma^2}{n} = \frac{\sigma^2 \sum_l W_l c_{2l}}{n_1 c_1 + \sum_l n_{2l} c_{2l}}. \qquad (40.68)$$

Из (40.64) и (40.68) следует, что отношение дисперсий при двухэтапной расслоенной и при нерасслоенной выборке равно

$$\frac{\mathsf{D}(\hat{\mu}_{12})}{\mathsf{D}(m_R)} = \frac{\left(\dfrac{\sigma^2 - \sum_l W_l \sigma_l^2}{n_1} + \sum W_l^2 \dfrac{\sigma_l^2}{n_{2l}}\right)\left(n_1 c_1 + \sum n_{2l} c_{2l}\right)}{\sigma^2 \sum W_l c_{2l}}. \qquad (40.69)$$

Числитель в (40.69) равен произведению $\mathsf{D}(\hat{\mu}_{12})(C - c_0)$, которое минимально, если n_1 и n_{2l} удовлетворяют (40.66). Но, согласно (39.52), для этого минимально возможного значения имеет место

$$\frac{\min \mathsf{D}(\hat{\mu}_{12})}{\mathsf{D}(m_R)} = \frac{\left[\{(\sigma^2 - \sum W_l \sigma_l^2) c_1\}^{1/2} + \sum W_l \sigma_l c_{2l}^{1/2}\right]^2}{\sigma^2 \sum W_l c_{2l}}. \qquad (40.70)$$

По-видимому это наиболее полезная форма записи отношения дисперсий. Рассматривая снова числитель (40.70), убеждаемся, что, согласно неравенству Коши, он не больше чем

$$\{(\sigma^2 - \sum W_l \sigma_l^2) + \sum W_l \sigma_l^2\}\{c_1 + \sum W_l c_{2l}\},$$

а значит, из (40.70)

$$\frac{\min \mathsf{D}(\hat{\mu}_{12})}{\mathsf{D}(m_R)} \leqslant 1 + \frac{c_1}{\sum W_l c_{2l}}. \qquad (40.71)$$

Таким образом, если $c_1 = 0$, то двухэтапный расслоенный выбор с распределением объемов выборок с МД никогда не

хуже, чем нерасслоенный выбор с теми же самыми ожидаемыми затратами. Но если $c_1 = 0$, то W_l можно вычислить точно с нулевыми затратами, так что, по существу, мы получим обычный расслоенный выбор. Таким образом, мы получили перефразировку результата **39.19**, рассмотрев дополнительно случай разных затрат на наблюдения в разных слоях.

Если $c_1 > 0$ в (40.70), то может оказаться, что нерасслоенная выборка более эффективна, но (40.70), очевидно, — возрастающая функция от c_1, и если c_1 мало по сравнению со взвешенным средним $\sum W_l c_{2l}$, то правая часть (40.71) не может быть намного больше единицы. Так что в самом плохом случае двухэтапная расслоенная выборка с соответствующим распределением по слоям может привести лишь к небольшой потере в эффективности. Простой числовой пример неблагоприятной ситуации получается, если положить $c_1 = 1$, $c_{2l} = 6$ для всех l, $\sigma^2 = 10$, $\sigma_l^2 = 6$ для всех l. Тогда (40.70) дает значение $(2+6)^2/(10 \times 6) = 1{,}07$. Если же $c_{2l} = 9$, то (40.70) дает $(1+7{,}35)^2/(10 \times 9) = 0{,}77$.

40.28 Если первый этап выбора проводится для оценки среднего вспомогательной переменной μ_x, чтобы на втором этапе построить регрессионную оценку, или оценку-отношение, то, так же как в **40.24**, при вычислении дисперсии используется (39.73). Рассмотрим только простейший случай, когда применяется смещенная оценка-отношение (40.1). При двухэтапной выборке получим

$$\tilde{\mu}_{12} = m_x^{(1)} \cdot m_y^{(2)}/m_x^{(2)},$$

где верхний индекс указывает на этап. Если два этапа представляют собой *независимый* выбор с равными вероятностями, то, используя **40.2**, получим

$$\mathbf{M}(\tilde{\mu}_{12}) = \underset{1}{\mathbf{M}}\{m_x^{(1)} \underset{2}{\mathbf{M}}(m_y^{(2)}/m_x^{(2)})\} = \\ = \underset{1}{\mathbf{M}}\{m_x^{(1)}[\mu_y/\mu_x + O(n_2^{-1})]\} = \mu_y + O(n_2^{-1}).$$

Пренебрегая членами порядка n_2^{-1} и выше получим, согласно (39.73),

$$\mathbf{D}(\tilde{\mu}_{12}) = \underset{1}{\mathbf{M}}\{\underset{2}{\mathbf{D}}(\tilde{\mu}_{12})\} + \underset{1}{\mathbf{D}}\{\underset{2}{\mathbf{M}}(\tilde{\mu}_{12})\} = \\ = \underset{1}{\mathbf{M}}\{(m_x^{(1)})^2 \underset{2}{\mathbf{D}}(m_y^2/m_x^2)\} + \underset{1}{\mathbf{D}}\{m_x^{(1)}\mu_y/\mu_x\}. \qquad (40.72)$$

Таким образом,

$$\mathbf{D}(\tilde{\mu}_{12}) \doteqdot \underset{2}{\mathbf{D}}(m_y^{(2)}/m_x^{(2)})\{\underset{1}{\mathbf{D}}(m_x^{(1)}) + \mu_x^2\} + \left(\frac{\mu_y}{\mu_x}\right)^2 \mathbf{D}(m_x^{(1)}).$$

Используя теперь (40.24), получим

$$\mathbf{D}(\tilde{\mu}_{12}) \doteqdot \frac{1}{n_2}\left\{\mathbf{D}(y) + \left(\frac{\mu_y}{\mu_x}\right)^2 \mathbf{D}(x) - 2\frac{\mu_y}{\mu_x} C(y, x)\right\}\left(1 + \frac{\mathbf{D}(x)}{n_1\mu_x^2}\right) + \\ + \left(\frac{\mu_y}{\mu_x}\right)^2 \frac{\mathbf{D}(x)}{n_1}. \qquad (40.73)$$

Первый множитель при $1/n_2$ в правой части (40.73) есть просто (40.25), примененное к выборке второго этапа. Как и в (40.63), выборка первого этапа приводит к появлению второго множителя, у которого член при $1/n_1$ дает увеличение дисперсии, и к появлению еще одного слагаемого порядка $1/n_1$. Поскольку мы уже пренебрегли членами, меньшими по порядку чем $1/n_2$, то (поскольку предполагается, что $n_1 > n_2$), пренебрегая членом, имеющим порядок $1/n_1 n_2$, получим приближение

$$\mathbf{D}(\tilde{\mu}_{12}) \doteqdot \frac{1}{n_2}\left\{\mathbf{D}(y) + \left(\frac{\mu_y}{\mu_x}\right)^2 \mathbf{D}(x) - 2\frac{\mu_y}{\mu_x} C(y, x)\right\} + \left(\frac{\mu_y}{\mu_x}\right)^2 \frac{\mathbf{D}(x)}{n_1}. \qquad (40.74)$$

Если выборка второго этапа не независима, а является подвыборкой первого этапа, то в (40.72) нужно μ_y/μ_x заменить на $m_y^{(1)}/m_x^{(1)}$, так что второй член в правой части (40.72) становится просто $\underset{1}{\mathbf{D}}(m_y^{(1)}) = \mathbf{D}(y)/n_1$. Если n_1 очень велико по сравнению с n_2, то первый член (40.72) в нашем приближении имеет тот же вид, что и раньше. Если же n_2/n_1 не очень мало, то приближение улучшается введением корректирующего множителя $\left(1 - \frac{n_2}{n_1}\right)$ к первому члену (40.74), так что в результате получается

$$\mathbf{D}(\tilde{\mu}_{12}) \doteqdot \left(\frac{1}{n_2} - \frac{1}{n_1}\right)\left\{\mathbf{D}(y) + \left(\frac{\mu_y}{\mu_x}\right)^2 \mathbf{D}(x) + 2\frac{\mu_y}{\mu_x} C(y, x)\right\} + \frac{\mathbf{D}(y)}{n_1}. \qquad (40.75)$$

40.29 Кокрэн (1963) и Иэйтс (1960) приводят подробности применения двуэтапных методов для регрессионных оценок. При этом для получения полезных формул для дисперсии необходимы некоторые ограничительные предположения.

Двухэтапная выборка естественным образом обобщается на *многоэтапную выборку*. Но для этой более общей процедуры имеется очень мало как теоретических, так и прикладных работ.

С. Гхош (1963а) рассматривает тип двухэтапной выборки, когда целью первого этапа является образование групп для выбора на втором этапе.

Радж (1964) рассматривает случай, когда первый этап служит для определения вероятностей извлечения, используемых на втором этапе. Сингх (1967) показывает, что в этом случае оценка (39.27) предпочтительнее, чем некоторые другие оценки.

Области изучения

40.30 При рассмотрении оценок-отношений мы считали, что вспомогательная переменная x имеет самый общий вид. На практике одним из самых важных является случай, когда x принимает только значения 0 или 1, указывая на то, принадлежит ли соответствующий элемент совокупности к некоторой подгруппе. Этому случаю соответствуют, например, следующие ситуации:

(а) Проводится выбор из всей совокупности, а нас интересует, исходя из некоторых посторонних соображений, только часть совокупности. Например, выбор проводится из совокупности людей в возрасте от 21 года и выше, а нас интересуют только люди в возрасте от 21 до 65. В этом случае объем выборки n из интересующей нас совокупности является, очевидно, случайной величиной и выборочное среднее для любой измеряемой в этой совокупности переменной имеет вид $\sum_{i=1}^{n} y_i x_i \Big/ \sum_{i=1}^{n} x_i$, где x_i, как и выше, принимает значения 0 или 1. Если известно среднее значение для совокупности переменной x (т. е. доля людей в возрасте от 21 до 65), то можно применить всю приведенную выше теорию.

(б) Нас интересует вся совокупность, из которой проводится выборка, но некоторые выбранные элементы не доставляют информации из-за утери записей, неполного обследования или просто уклонения от ответа (особенно в совокупности людей). Снова n — случайная величина и применимо замечание в конце (а)*).

(в) Мы имеем результаты наблюдений для всей выборки из интересующей нас совокупности, но хотим вычислить параметры для подгрупп этой совокупности. Например, имеется выборка из совокупности людей, а мы хотим вычислить значение некоторой статистики только для мужчин или только для женщин. Если слои в выборке соответствуют мужчинам и женщинам, то никаких новых моментов не возникает, ибо объемы выборок, соответствующих мужчинам и женщинам, будут фиксированными. Однако обычно такое расслоение невозможно, так что эти объемы выборок являются случайными величинами (хотя в рассматриваемом простом случае их сумма не случайна). И вообще, объем

*) В случае уклонения от ответа существует осложнение, поскольку уклонение от ответа может быть коррелированным со значением y, так что ответившая группа не может доставить несмещенную оценку среднего значения y для всей совокупности в целом.

выборки для любой не заданной заранее подгруппы должен быть случайной величиной.

40.31 Интересующие нас подгруппы называются *областями* (изучения). Разумеется, сами слои могут быть областями, но тогда не нужно никакой новой теории. Мы будем говорить об областях, подразумевая, что выборочные частоты соответствующих подгрупп случайны вне зависимости от причин, приводящих к этой случайности.

Часто области пересекаются со слоями и множествами различных шагов в выборке. И именно в этих случаях возникают новые вопросы. Иэйтс (1960) (как и в более раннем издании его книги) приводит несколько формул для областей, пересекающих слои. Кокрэн (1963) дает вывод некоторых из этих формул. Дербин (1958) рассматривает как указанную, так и несколько многошаговых ситуаций. Хартли (1959) также выводит некоторые результаты Иэйтса для ковариаций средних по областям и приводит некоторые дальнейшие результаты. Наше изложение следует Дербину (1958).

Области пересекающие слои

40.32. Как и в главе 39, предположим, что мы имеем k слоев с частотами N_l, $l = 1, \ldots, k$ и $\sum_l N_l = N$, в то время как частоты слоев в выборке равны n_l, а $\sum_l n_l = n$. Рассмотрим некоторую область, скажем d. Обозначим через $N_l^{(d)}$ частоту, l-го слоя в области, так что $\sum_l N_l^{(d)} = N^{(d)}$, где $N^{(d)}$ частота области во всей совокупности. Пусть $n_l^{(d)}$ и $n^{(d)} = \sum_l n_l^{(d)}$ — определенные аналогичным образом величины для выборки. Заметим, что, в то время как частоты слоев в совокупности известны, частоты слоев в области, вообще говоря, неизвестны. Конечно, нерасслоенному случайному выбору соответствует $k = 1$.

Пусть $y^{(d)}$ равно наблюденным значениям y для элементов, принадлежащих области, и равно нулю для других элементов. Тогда

$$y_{lj}^{(d)} = h_{lj} y_{lj}, \tag{40.76}$$

где

$$h_{lj} = \begin{cases} +1 & \text{внутри области,} \\ 0 & \text{вне области.} \end{cases} \tag{40.77}$$

В результате имеем

$$N_l^{(d)} = \sum_{j=1}^{N_l} h_{lj}, \qquad N^{(d)} = \sum_{l=1}^{k} N_l^{(d)} = \sum_{l=1}^{k}\sum_{j=1}^{N_l} h_{lj},$$
$$n_l^{(d)} = \sum_{j=1}^{n_l} h_{lj}, \qquad n^{(d)} = \sum_{l=1}^{k} n_l^{(d)} = \sum_{l=1}^{k}\sum_{j=1}^{n_l} h_{lj}. \tag{40.78}$$

Определим далее средние значения по области внутри слоев

$$\mu_l^{(d)} = \sum_{j=1}^{N_l} y_{lj}^{(d)} / N_l^{(d)} \tag{40.79}$$

и среднее значение по всей области

$$\mu^{(d)} = \sum_{l=1}^{k}\sum_{j=1}^{N_l} y_{lj}^{(d)} / N^{(d)} = \sum_l N_l^{(d)} \mu_l^{(d)} / N^{(d)}. \tag{40.80}$$

40.33 Мы хотим теперь найти оценку величины $\mu^{(d)}$ в (40.80). Рассмотрим сначала случай, когда выбор проводится с равными вероятностями с использованием ВРД, скажем $f = n/N$, как в главе 39. Оценка-отношение

$$m^{(d)} = \sum_l \sum_{j=1}^{n_l} y_{lj}^{(d)} / n^{(d)} = \sum_l \sum_j y_{lj}^{(d)} \Big/ \sum_l \sum_j h_{lj} \tag{40.81}$$

является выборочным аналогом $\mu^{(d)}$. По существу, это оценка (40.1), где числитель и знаменатель суммируются отдельно внутри слоев, т. е. мы получили пример «комбинированной» оценки-отношения, упомянутой в **40.20**. Используя аналоги (40.3) и (40.5), убеждаемся, что приближение первого порядка дает с использованием (40.78) и (40.80)

$$\mathbf{M}(m^{(d)}) = \mathbf{M}\left(\sum_l \sum_{j=1}^{n_l} y_{lj}^{(d)}\right) \Big/ \mathbf{M}\left(\sum_l \sum_{j=1}^{n_l} h_{lj}\right) =$$
$$= \left(\sum_l f \sum_{j=1}^{N_l} y_{lj}^{(d)}\right) \Big/ \left(\sum_l f \sum_{j=1}^{N_l} h_{lj}\right) = \mu^{(d)}. \tag{40.82}$$

40.34 Для нахождения дисперсии $m^{(d)}$ положим

$$z_{lj} = h_{lj}(y_{lj} - \mu^{(d)}) = y_{lj}^{(d)} - h_{lj}\mu^{(d)}, \tag{40.83}$$

так что, применяя (40.78) и (40.81), получим

$$m^{(d)} - \mu^{(d)} = \sum_l \sum_{j=1}^{n_l} z_{lj} / n^{(d)}. \tag{40.84}$$

Но (40.84) асимптотически сводится к

$$m^{(d)} - \mu^{(d)} \sim \sum_l \sum_j z_{lj} / \mathbf{M}(n^{(d)}) = \sum_l \sum_j z_{lj} \Big/ \left(n \frac{N^{(d)}}{N}\right). \tag{40.85}$$

Таким образом,

$$\mathbf{D}(m^{(d)}) \sim \left(\frac{N}{N^{(d)}}\right)^2 \mathbf{D}\left(\frac{1}{n} \sum_l \sum_{j=1}^{n_l} z_{lj}\right). \tag{40.86}$$

Дисперсия в правой части (40.86) совпадает с дисперсией среднего значения при расслоенной выборке с использованием ВРД. А значит, согласно (39.45), имеем

$$\mathbf{D}(m^{(d)}) \sim \left(\frac{N}{N^{(d)}}\right)^2 \frac{N-n}{nN^2} \sum_l N_l \sigma_l^2(z) =$$
$$= \frac{N-n}{n(N^{(d)})^2} \sum_l \left[\sum_{j=1}^{N_l} z_{lj}^2 - \frac{\left(\sum_{j=1}^{N_l} z_{lj}\right)^2}{N_l}\right]; \tag{40.87}$$

при этом мы пренебрегли тем, что в определении $\sigma_l^2(z)$ в качестве делителя фигурирует $N_l - 1$, а не N_l. Согласно (40.83) соотношение (40.87) можно, с использованием (40.78—9) и того факта, что $h_{lj}^2 = h_{lj}$, переписать в виде

$$\frac{n(N^{(d)})^2}{N-n} \mathbf{D}(m^{(d)}) \sim \sum_l \left[\sum_{j=1}^{N_l} \{h_{lj}(y_{lj} - \mu^{(d)})\}^2 - \frac{\left\{\sum_{j=1}^{N_l} h_{lj}(y_{lj} - \mu^{(d)})\right\}^2}{N_l}\right] =$$
$$= \sum_l \left[\sum_{j=1}^{N_l} h_{lj}(y_{lj} - \mu^{(d)})^2 - \frac{\{N_l^{(d)}(\mu_l^{(d)} - \mu^{(d)})\}^2}{N_l}\right]. \tag{40.88}$$

Сумма по j в правой части (40.88) может быть переписана в виде

$$\sum_{j=1}^{N_l} h_{lj}(y_{lj} - \mu^{(d)})^2 = \sum_j (y_{lj}^{(d)} - \mu^{(d)})^2 h_{lj} =$$
$$= \sum_j (y_{lj}^{(d)} - \mu_l^{(d)})^2 h_{lj} + N_l^{(d)}(\mu_l^{(d)} - \mu^{(d)})^2. \tag{40.89}$$

Подставляя (40.89) в (40.88), получим

$$\mathbf{D}(m^{(d)}) \sim \frac{N-n}{n(N^{(d)})^2} \sum_l \left[\sum_{j=1}^{N_l} (y_{lj}^{(d)} - \mu_l^{(d)})^2 h_{lj} + \right.$$
$$\left. + N_l^{(d)}\left(1 - \frac{N_l^{(d)}}{N_l}\right)(\mu_l^{(d)} - \mu^{(d)})^2\right]. \tag{40.90}$$

40.35 Ограничившись рассматриваемой нами точностью приближения, можно написать $n/N = n^{(d)}/N^{(d)}$, а тогда первый член в правой части (40.90) перепишется в виде

$$\frac{N^{(d)} - n^{(d)}}{n^{(d)}(N^{(d)})^2} \sum_l \sum_j (y_{lj}^{(d)} - \mu_l^{(d)})^2 h_{lj}. \qquad (40.91)$$

Как следует из нашего вывода (40.87) из (39.45), соотношение (40.91) в точности совпадает с «дисперсией», которую мы бы получили, если бы считали выборочные частоты слоев в области фиксированными (на самом деле это не так). А значит, второй член в (40.90) указывает на увеличение дисперсии, соответствующее тому, что $n_l^{(d)}$ — случайные величины. Этот член велик, если средние по области внутри слоев $\mu_l^{(d)}$ различаются существенно и в особенности, если доли слоев, принадлежащие области, $N_l^{(d)}/N_l$, малы.

Если убрать множитель $\left(1 - \frac{N_l^{(d)}}{N_l}\right)$ во втором члене правой части (40.90) и воспользоваться приближением (40.91), то все выражение (40.90) в рассматриваемом приближении будет совпадать с первой формулой в **39.18**, которая, как мы видели, задает дисперсию для *нерасслоенного выбора*. Таким образом, если все доли $N_l^{(d)}/N_l$ малы, то существование у $n_l^{(d)}$ ненулевой дисперсии, по существу, сводит на нет всю выгоду в эффективности оценки, получающуюся в результате расслоения. Только в том случае, когда по крайней мере для некоторых слоев объемы области в них велики, сохраняется большая часть выгоды.

Оценку дисперсии (40.90) можно получить точно так же, как была получена и сама формула (40.90). В упражнении 40.15 читателю предлагается проделать соответствующие выкладки.

40.36 Если выбор проводится не с равными долями, то оценку (40.81) нужно видоизменить, чтобы должным образом учесть вклады слоев. Вместо (40.81) положим

$$r = \sum_l \frac{N_l}{n_l} \sum_{j=1}^{n_l} y_{lj}^{(d)} \Big/ \sum_l \frac{N_l}{n_l} n_l^{(d)}. \qquad (40.92)$$

В упражнении 40.16 читателю предлагается показать, что r — асимптотически несмещенная оценка $\mu^{(d)}$ с дисперсией

$$\mathsf{D}(r) \sim \frac{1}{(N^{(d)})^2} \sum_l \frac{N_l}{n_l}\left(1 - \frac{n_l}{N_l}\right) \sum_{j=1}^{n_l} \left(z_{lj} - \frac{\sum_{j=1}^{N_l} z_{lj}}{N_l}\right)^2, \qquad (40.93)$$

что обобщает (40.87). После подстановки значений z_{lj} это сводится к

$$\mathsf{D}(r) = \left(\sum_l \frac{N_l}{n_l} n_l^{(d)}\right)^{-2} \sum_l \frac{N_l}{n_l}\left(1 - \frac{n_l}{N_l}\right)\left[\sum_{j=1}^{N_{lj}} (y_{lj}^{(d)} - \mu_l^{(d)})^2 h_{lj} + \right.$$
$$\left. + N_l^{(d)}\left(1 - \frac{N_l^{(d)}}{N_l}\right)(\mu_l^{(d)} - \mu^{(d)})^2\right], \qquad (40.94)$$

что обобщает (40.90). Читателю снова предлагается вывести в упражнении 40.16, что оценкой (40.94) служит

$$\hat{\mathsf{D}}(r) = \left(\sum_l \frac{N_l}{n_l} n_l^{(d)}\right)^{-2} \sum_l \frac{N_l^2}{n_l(h_l - 1)}\left(1 - \frac{n_l}{N_l}\right)\left[\sum_{j=1}^{n_l} (y_{lj}^{(d)} - m_l^{(d)})^2 h_{lj} + \right.$$
$$\left. + n_l^{(d)}\left(1 - \frac{n_l^{(d)}}{n_l}\right)(m_l^{(d)} - m^{(d)})^2\right]. \qquad (40.95)$$

Области при многошаговом выборе

40.37 Мы ограничимся разбором случая, когда на первом шаге из общего числа S множеств извлекается с возвращением s множеств. При этом разрешается любое число шагов (включая нуль), однако мы ограничимся далее равновзвешивающей схемой (см. **39.41**), для которой выборочное среднее является оценкой соответствующего среднего по совокупности μ.

Мы хотим оценить среднее по всей области, обозначенное, как и ранее, $\mu^{(d)}$, где

$$\mu^{(d)} = \sum_{i=1}^{S} \sum_j \cdots \sum_p y_{ij\ldots p}^{(d)} / N^{(d)}, \qquad (40.96)$$

а $N^{(d)}$ — полная частота области

$$N^{(d)} = \sum_{i=1}^{S} \sum_j \cdots \sum_p h_{ij\ldots p} = \sum_{i=1}^{S} N_i^{(d)}, \qquad (40.97)$$

а $h_{ij\ldots p}$, как и в (40.77), принимают значения 0,1.

40.38 Оценкой является выборочный аналог $\mu^{(d)}$

$$m^{(d)} = \sum_{i=1}^{s} \sum_j \cdots \sum_p y_{ij\ldots p}^{(d)} / n^{(d)} \equiv \sum_{i=1}^{s} y_i^{(d)} / n^{(d)}, \qquad (40.98)$$

где

$$n^{(d)} = \sum_{i=1}^{s} \sum_j \cdots \sum_p h_{ij\ldots p} \equiv \sum_{i=1}^{s} n_i^{(d)}. \qquad (40.99)$$

Как и в (40.82), приближение первого порядка дает

$$\mathsf{M}(m^{(d)}) \sim \mathsf{M}\left(\sum_{i=1}^{s}\sum_{j}\cdots\sum_{p} y_{ij\ldots p}^{(d)}\right)\Big/\mathsf{M}(n^{(d)}) =$$

$$= f\sum_{i=1}^{S}\sum_{j}\cdots\sum_{p} y_{ij\ldots p}^{(d)}\Big/ f\sum_{i=1}^{S}\sum_{j}\cdots\sum_{p} h_{ij\ldots p} = \mu^{(d)}, \quad (40.100)$$

где общий множитель f у числителя и знаменателя равен полной вероятности извлечения для каждого значения y из совокупности.

40.39 Так же как и в (40.83), определим

$$z_{ij\ldots p} = y_{ij\ldots p}^{(d)} - h_{ij\ldots p}\mu^{(d)}, \quad (40.101)$$

и, как и в (40.84), получим, что

$$m^{(d)} - \mu^{(d)} = \frac{1}{n^{(d)}}\sum_{i=1}^{s}\sum_{j}\cdots\sum_{p} z_{ij\ldots p} =$$

$$= \frac{s}{n^{(d)}}\cdot\frac{1}{s}\sum_{i=1}^{s}\left(\sum_{j}\cdots\sum_{p} z_{ij\ldots p}\right). \quad (40.102)$$

Отсюда, переходя сразу к оценке дисперсии величины $m^{(d)}$, получим

$$\hat{\mathsf{D}}(m^{(d)}) \sim \left\{\frac{s}{\mathsf{M}(n^{(d)})}\right\}^2 \hat{\mathsf{D}}(\bar{z}), \quad (40.103)$$

где $\bar{z}$ — среднее от s значений $z_i = \sum_j \cdots \sum_p z_{ij\ldots p}$. Далее из (39.109)

$$\hat{\mathsf{D}}(\bar{z}) = \underset{1}{\hat{\mathsf{D}}}(\bar{z}),$$

и, в частности, если вероятности выбора одни и те же для каждого извлечения первого шага, то из (39.110) получим в этом случае с $t_i = z_i/s$

$$\hat{\mathsf{D}}(\bar{z}) = \underset{1}{\hat{\mathsf{D}}}(z) = \frac{1}{s(s-1)}\sum_{i=1}^{s}(z_i - \bar{z})^2.$$

Отсюда (40.103) принимает вид

$$\hat{\mathsf{D}}(m^{(d)}) \sim \frac{s}{(s-1)\{\mathsf{M}(n^{(d)})\}^2}\sum_{i=1}^{s}(z_i - \bar{z})^2. \quad (40.104)$$

Так же как в (40.87—8), используя (40.101) и определения $y_i^{(d)}$ и $n_i^{(d)}$ в (40.98—9), преобразуем сумму квадратов в (40.104)

к виду

$$\sum_{i=1}^{s}(z_i - \bar{z})^2 = \sum_i z_i^2 - s\bar{z}^2 =$$

$$= \sum_i \left(y_i^{(d)} - n_i^{(d)}\mu^{(d)}\right)^2 - \frac{(n^{(d)})^2}{s}(m^{(d)} - \mu^{(d)})^2. \quad (40.105)$$

Предполагая, что $n^{(d)} \sim \mathsf{M}(n^{(d)})$, из (40.104—5) получаем

$$\hat{\mathsf{D}}(m^{(d)}) \sim \frac{s}{(s-1)(n^{(d)})^2}\sum_{i=1}^{s}\left(y_i^{(d)} - n_i^{(d)}\mu^{(d)}\right)^2 - \frac{1}{s-1}(m^{(d)} - \mu^{(d)})^2, \quad (40.106)$$

и в рассматриваемом приближении математическое ожидание второго члена в правой части (40.106) равно $-\frac{1}{s-1}\mathsf{D}(m^{(d)})$. Перенося его в левую часть, найдем

$$\hat{\mathsf{D}}(m^{(d)}) \sim \frac{1}{(n^{(d)})^2}\sum_{i=1}^{s}\left(y_i^{(d)} - n_i^{(d)}\mu^{(d)}\right)^2. \quad (40.107)$$

Соотношение (40.107) все еще не статистика, поскольку оно зависит от $\mu^{(d)}$. Но в рассматриваемом приближении $\mu^{(d)}$ можно заменить на $m^{(d)}$, так что окончательно получим оценку

$$\hat{\mathsf{D}}(m^{(d)}) = \frac{1}{(n^{(d)})^2}\sum_{i=1}^{s}\left(y_i^{(d)} - n_i^{(d)}m^{(d)}\right)^2. \quad (40.108)$$

Если после первого шага не проводится никакого выбора, то в рассматриваемом приближении (40.108) согласуется с результатом упражнения 40.15.

40.40 Если бы мы считали $n_i^{(d)}$ фиксированными (что не так), то для оцениваемой дисперсии величины $m^{(d)}$ мы должны были бы, согласно (39.110) с $t_i = y_i^{(d)}/n^{(d)}$, положить

$$\frac{1}{(n^{(d)})^2}\frac{s}{s-1}\sum_{i=1}^{s}\left(y_i^{(d)} - \frac{n^{(d)}}{s}m^{(d)}\right)^2. \quad (40.109)$$

Сравнение (40.108) и (40.109) показывает, что случайность $n_i^{(d)}$ приводит к тому, что в выражении для дисперсии средняя частота области $n^{(d)}/s$ в множестве первого шага заменяется на конкретные частоты области в множествах первого шага $n_i^{(d)}$. Возрастание дисперсии велико только в том случае, если $n_i^{(d)}$ очень значительны и если они отрицательно коррелированы с $y_i^{(d)}$, что неправдоподобно в практических ситуациях.

УПРАЖНЕНИЯ

40.1 Из (40.7) и (40.9) показать, что

$$\frac{\mu_x}{n-1}(nm_y/m_x - m_{y/x})$$

является приближенно несмещенной оценкой для μ_y.

(См. Мёрфи и Нанджамма, 1959.)

40.2 Показать, что *оценка-произведение*

$$\bar{\mu}_y = m_y m_x/\mu_x$$

являющаяся аналогом оценки-отношения (40.1), имеет смещение $C(m_y, m_x)/\mu_x$, а значит оценка

$$\bar{\mu}_y' = \left\{ n(N-1)\, m_y m_x - (N-n) \sum_{i=1}^{n} y_i x/n \right\} \Big/ \{N(n-1)\,\mu_x\}$$

несмещенная. Показать далее, что при $N \to \infty$

$$\mathrm{D}(\bar{\mu}_y') = \frac{\mu_y^2}{n}\left[\left(\frac{\mathrm{D}(y)}{\mu_y^2} + \frac{\mathrm{D}(x)}{\mu_x^2} + \frac{2C(y, x)}{\mu_y \mu_x}\right) + \right.$$
$$\left. + \frac{1}{n-1}\left(\frac{\mathrm{D}(y)\,\mathrm{D}(x) + \{C(y, x)\}^2}{\mu_y^2 \mu_x^2}\right)\right].$$

Вывести отсюда, что если $C(y, x) < 0$, то $\bar{\mu}_y'$ более эффективна, чем $\bar{\mu}_y'$, чья средняя квадратичная ошибка задается (40.25). Показать также, что (см. (40.27)) $\bar{\mu}_y'$ более эффективна, чем выборочное среднее m_y, если $\beta_{yx} < -\frac{1}{2}\frac{\mu_y}{\mu_x} < 0$ или $\beta_{yx} > -\frac{1}{2}\frac{\mu_y}{\mu_x} > 0$.

(См. Робсон, 1957.)

40.3 Проводится обратный выбор (см. примеры 9.13 и 34.1) с неравными вероятностями и с возвращением из совокупности, состоящей из N элементов, причем извлечения продолжаются до тех пор, пока не будут извлечены $(r+1)$ различных элементов. Пусть на это потребовалось $(n+1)$ извлечений $(n \geqslant r)$. Последнее извлечение не учитывается и в результате получается r различных значений y_i, наблюденных соответственно n_i раз, так что $\sum_{i=1}^{r} n_i = n$.

Показать, что $t = \frac{1}{n}\sum_{i=1}^{r} n_i y_i$ является несмещенной оценкой среднего совокупности и что (см. упражнение 39.12)

$$\hat{\mathrm{D}}(t) = \frac{1}{n(n-1)}\sum_{i=1}^{r} n_i (y_i - t)^2.$$

является несмещенной оценкой дисперсии оценки t.

(Сэмпфорд (1962); Патхак (1964b) улучшил оценку — см. 39.5.)

40.4 Показать, что если k слоев фиксированных объемов N_i образованы случайным делением совокупности объема N, а затем проводятся n извлечений с равными вероятностями без возвращения с использованием ВРД, то дисперсия оценки (39.38) для указанной процедуры в точности совпадает с дисперсией для среднего при нерасслоенной выборке того же объема и из той же совокупности.

Показать далее, что если использовать не ВРД, а любое другое распределение по слоям, то дисперсия (39.38) увеличивается.

40.5 Рассматривается следующая схема выбора, состоящая из одного или большего числа шагов. Выбирается с возвращением n множеств первого шага из одного слоя, состоящего из N множеств с использованием неравных вероятностей $p_r \left(\sum_{r=1}^{N} p_r = 1\right)$ при каждом извлечении. На последующих шагах выбор проводится независимо внутри каждого извлеченного множества первого шага. Эта схема видоизменяется предварительным случайным разбиением N множеств первого шага на n групп, содержащих $N_1, \ldots, N_n$ множеств и выбором по одному элементу из каждой группы с теми же самыми относительными вероятностями, что и в первой схеме, т. е. $p_r/p_{i.}$, где $p_{i.}$ равно сумме p_r по i-й группе. Последующие шаги не меняются. Показать, что $t_M(z) = \sum_{i=1}^{n} p_{i.} z_i$ для видоизмененной схемы имеет то же самое математическое ожидание, что и $t_0(z) = \frac{1}{n}\sum_{i=1}^{n} z_i$ для исходной схемы, и что

$$\mathrm{D}\{t_M(z)\} = n\,\frac{\left(\sum_{i=1}^{n} N_i^2 - N\right)}{N(N-1)}\,\mathrm{D}\{t_0(z)\} + n\,\frac{\left(N^2 - \sum_{i=1}^{n} N_i^2\right)}{N(N-1)}\,D,$$

где D — компонента $\mathrm{D}\{t_0(z)\}$, обусловленная последующими шагами.

40.6 Показать в упражнении 40.5, что если N_i выбрать равными, то для одношагового выбора $t_M(z)$ никогда не менее эффективна, чем $t_0(z)$. В то же время такой выбор N_i соответствует наилучшей возможности, чтобы при многошаговой выборке получить не меньшую эффективность (см. упражнение 40.4).

(Результаты этого и предыдущих двух упражнений получены Стьюартом (1964). Он рассматривал обобщения модифицированной схемы выбора. См. также Дж. Рао, Хартли и Кокрэн (1962).)

40.7 Совокупность состоит из N элементов. Случайной выборкой с равными вероятностями извлекается n элементов, причем только n_1 из них дают ответ о значении переменной y. Для оставшихся $n - n_1 = n_2$ элементов, не давших ответа, проводится подвыборка по одному на каждые k элементов, так что объем подвыборки $r_2 = n_2/k$. Затраты на всю процедуру равны

$$C = c_0 n + c_1 n_1 + c_2 n_2.$$

Оценкой μ_y является

$$\hat{\mu}_y = \frac{1}{n}(n_1 m_1 + n_2 m_2),$$

где m_1 и m_2 — средние значения для давших ответ в исходной выборке и соответственно для подвыборки из не давших ответ в первый раз. Используя (39.72—3), показать, что $\hat{\mu}_y$ — несмещенная оценка с дисперсией

$$\mathbf{D}(\hat{\mu}_y) = \left(1 + \frac{n}{N}\right)\frac{\sigma^2}{n} + W_2(k-1)\frac{\sigma_2^2}{n},$$

где σ^2 — дисперсия y, взятая по всей совокупности, а σ_2^2 — дисперсия, взятая по потенциально не дающим ответ элементам совокупности, доля которых в совокупности равна W_2. Используя 39.20, показать, что $\mathbf{D}(\hat{\mu}_y)$ минимально при фиксированном математическом ожидании полных затрат, если k удовлетворяет соотношению

$$k^2_{\text{МД}} = \frac{c_2(\sigma^2 - W_2\sigma_2^2)}{\sigma_2^2\{c_0 + c_1(1 - W_2)\}}.$$

(См. Хансен и Гурвиц, 1946.)

40.8 Пусть в упражнении 40.7 $k = 1$, т. е. все не давшие ответ извлекаются во второй раз, и пусть $\sigma^2 = \sigma_2^2$, $N \to \infty$ и математическое ожидание полных затрат то же самое, что и для выборки с МД, т. е. при $k = k_{\text{МД}}$. Показать, что эффективность в этом случае задается соотношением

$$\frac{\mathbf{D}(\hat{\mu}_y \mid k_{\text{МД}})}{\mathbf{D}(\hat{\mu}_y \mid k = 1)} = 1 - \frac{W_2(1 - W_2)\left(1 - \frac{1}{k_{\text{МД}}}\right)^2}{W_2 + (1 - W_2)/k^2_{\text{МД}}}.$$

Показать, что если $W_2 = 0{,}6$ и $k^2_{\text{МД}} \leqslant 3$, то эффективность для $k = 1$ составляет более 94%. Это иллюстрирует относительную нечувствительность эффективности к отклонению от $k_{\text{МД}}$.

(Дербин, 1954.)

40.9 Из очень большой совокупности последовательно извлекаются две выборки с равными вероятностями, причем наблюдение ведется за разными переменными. Дисперсия обеих переменных одна и та же и равна σ^2, а корреляция между ними равна ρ. Первая выборка имеет объем n и дает среднее первой переменной m_1. Во второй выборке сохраняется доля f от первой выборки (среднее первой переменной оказывается равным m_1', а среднее второй переменной m_2'), а оставшиеся $n(1-f)$ членов первой выборки (со средним первой переменной, равным m_1^0) заменяются на новую выборку, дающую среднее значение m_2'' второй переменной. Для среднего значения второй переменной имеются две независимые оценки. Одна из них m_2'', а вторая

$$\hat{\mu}_{12} = m_2' + b_{21}(m_1 - m_1')$$

— двухэтапная регрессионная оценка, основанная на наблюденном коэффициенте регрессии b_{21} значения второй переменной на значения первой переменной для nf сохраненных элементов. Показать, что

$$\mathbf{D}(\hat{\mu}_{12}) \doteqdot \frac{\sigma^2}{nf}\{1 - (1 - f)\rho^2\},$$

а значит, линейная комбинация $\hat{\mu}_{12}$ и m_2'' с МД равна

$$\hat{\mu} = \frac{f}{\{1 - (1 - f)^2\rho^2\}}\hat{\mu}_{12} + \frac{(1 - f)\{1 - (1 - f)\rho^2\}}{\{1 - (1 - f)^2\rho^2\}}m_2''$$

с дисперсией

$$\mathbf{D}(\hat{\mu}) \doteqdot \frac{\sigma^2}{n}\left\{\frac{1 - (1 - f)\rho^2}{1 - (1 - f)^2\rho^2}\right\}.$$

40.10 В упражнении 40.9 показать далее, что независимыми оценками разности между средними значениями двух переменных служат $d_1 = m_2' - m_1'$ или $d_2 = m_1'' - m_1^0$. Показать, что линейная комбинация d_1 и d_2 с МД имеет вид

$$d = \frac{f}{\{1 - \rho(1 - f)\}}d_1 + \frac{(1 - f)(1 - \rho)}{\{1 - \rho(1 - f)\}}d_2,$$

а ее дисперсия равна

$$\mathbf{D}(d) \doteqdot \frac{2\sigma^2}{n}\frac{(1 - \rho)}{\{1 - \rho(1 - f)\}}.$$

(Иэйтс (1960); Паттерсон (1950) излагает теорию для нескольких переменных; Вос (1964) приводит формулу для дисперсии при выборе, одновременном по времени и пространству.)

40.11 Простым случайным выбором извлекается n элементов из бесконечной совокупности, состоящей из k слоев. Доля l-го слоя в совокупности равна p_l, а в результате выбора из l-го слоя оказываются извлеченными n_l элементов, $\sum_l n_l = n$. До выбора фиксируются *намеченные* объемы выборок m_l из l-го слоя. Для тех слоев, для которых $m_l > n_l$, внутри этих слоев независимо проводится простой случайный выбор объема $m_l - n_l$. Пусть затраты на отдельные извлечения в исходной выборке равны c, в дополнительной выборке внутри l-го слоя $c_l > c$, а затраты на лишние члены в l-м слое (в намеченной выборке) равны $c_l' < c$. Показать, что математическое ожидание затрат, необходимых для получения намеченной расслоенной выборки, равно

$$\mathbf{M}(C) = nc + \sum_l \mathbf{P}\{n_l < m_l\}\,\mathbf{M}(m_l - n_l \mid n_l < m_l)\,c_l + \\ + \mathbf{P}\{n_l \geqslant m_l\}\,\mathbf{M}(m_l - n_l \mid n_l \geqslant m_l)\,c_l'],$$

и что если n_l достаточно велики, то это приближенно равно

$$\mathbf{M}(C) \doteqdot nc + \sum_l (c_l - c_l')\left[(m_l - np_l)\,G\left\{\frac{m_l - np_l}{[np_l(1 - p_l)]^{1/2}}\right\} + \right. \\ \left. + [n_l p_l(1 - p_l)]^{1/2}\,g\left\{\frac{m_l - np_l}{[np_l(1 - p_l)]^{1/2}}\right\}\right] + \sum_l (m_l - np_l)\,c_l',$$

где G — стандартное нормальное распределение, а g — его плотность. Показать, что если n увеличивается на единицу, то изменение $\mathsf{M}(C)$ равно

$$\Delta\mathsf{M}(C) = c - \sum_l \left[\mathsf{P}\{n_l < m_l\}(c_l - c_l') + c_l'\right] p_l,$$

и что для наименьшего n, для которого $\Delta\mathsf{M}(C) > 0$, приближенно выполняется равенство

$$\sum_l \left\{ \frac{(c_l - c_l')\, p_l}{c - \sum_l c_l' p_l} \right\} \mathsf{P}\{n_l < m_l\} = 1.$$

В случае $c_l' = 0$ (т. е. когда затраты на лишние наблюдения во всех слоях равны нулю) это равенство сводится к

$$\sum c_l p_l \mathsf{P}\{n_l < m_l\} = c,$$

а если все c_l равны, все c_l' равны и все $p_l = \frac{1}{k}$, то к

$$\mathsf{P}\{n_l < m_l\} = \frac{c - c_l'}{c_l - c_l'}.$$

(Н. Джонсон (1957); Янг (1961) рассмотрел аналогичную последовательную схему, в которой выбор проводится из всей совокупности до тех пор, пока из каждого слоя не будет извлечено нужное число m_l элементов.)

40.12 Используя (40.33), проверить выражения (40.44—9) для математических ожиданий и дисперсий трех статистик, рассмотренных в **40.9**.

40.13 Выписать разности $\mathsf{D}\left(\frac{m_y}{m_x}\right) - \mathsf{D}\left\{t\left(\frac{m_y}{m_x}\right)\right\}$ и $\mathsf{D}\left\{t\left(\frac{m_y}{m_x}\right)\right\} - \mathsf{D}(u)$ в **40.9** и показать, что они положительны.

(Тин, 1965.)

40.14 В **40.9** показать, что оценка

$$b = \frac{m_y}{m_x} \frac{\left(1 + \alpha_1 \frac{s_{xy}}{m_x m_y}\right)}{\left(1 + \alpha_1 \frac{s_x^2}{m_x^2}\right)}$$

с точностью до n^{-2} имеет математическое ожидание

$$\mathsf{M}(b) = \frac{\mu_y}{\mu_x}\left\{1 - \left(2\alpha_2 - \frac{3\alpha_1}{N}\right)(C_{21} - C_{30}) - 2\alpha_1^2 C_{20}(C_{20} - C_{11})\right\}$$

и ту же самую дисперсию, что и u, определенное в (40.32), так что фактически b и u эквивалентны.

(Тин (1965) — оценка принадлежит Билу.)

40.15 Показать, что оценкой для (40.90) служит

$$\hat{\mathsf{D}}(m^{(d)}) = \frac{(N-n)}{N(n^{(d)})^2} \sum_{l=1}^{k} \frac{n_l}{n_l - 1} \left[\sum_{j=1}^{n_l} h_{lj}\left(y_{lj}^{(d)} - m_l^{(d)}\right)^2 + n_l^{(d)}\left(1 - \frac{n_l^{(d)}}{n_l}\right)\left(m_l^{(d)} - m^{(d)}\right)^2\right],$$

где $m_e^{(d)}$ — выборочный аналог $\mu_l^{(d)}$.

(Дербин, 1958.)

40.16 В **40.36** вывести (40.94), обобщающее (40.90), и (40.95), обобщающее упражнение 40.15.

(Дербин, 1958.)

40.17 Проверить численные значения среднеквадратичных ошибок для шести оценок из (40.29).

ГЛАВА 41

ТЕОРИЯ МНОГОМЕРНЫХ РАСПРЕДЕЛЕНИЙ

41.1 С определенной точки зрения почти весь этот том посвящен многомерному анализу, т. е. анализу систем, каждый элемент которых характеризуется значениями многих подлежащих измерению или классификации величин. Однако пока нам обычно удавалось упрощать задачи: или (как при выборочном обследовании) мы начинали с того, что интересовались оценкой какого-либо одного параметра, например среднего, или (как при планировании эксперимента) мы упорядочивали переменные в уравнениях регрессии так, чтобы оценки коэффициентов регрессии становились ортогональными, что приводило к тому, что при классификации отдельные эффекты удавалось индивидуализировать.

Теперь мы пойдем дальше и изучим системы бо́льшей общности, характеризуемые зависимыми между собой величинами. В этой главе мы обсудим несколько характерных задач о распределениях. Если не оговорено противное, первоначальные (или основные) распределения мы будем считать многомерными нормальными. К сожалению, отличительной чертой обсуждаемой тематики является то, что в других случаях найти точные распределения интересующих нас статистик очень трудно.

41.2 В т. 1, главе 15 мы приводили две формы записи p-мерного нормального распределения

$$dF \propto \exp\left\{-\frac{1}{2}\sum_{j=1}^{p}\sum_{k=1}^{p}\alpha_{jk}^{0}\left(\frac{x_j-\mu_j}{\sigma_j}\right)\left(\frac{x_k-\mu_k}{\sigma_k}\right)\right\}\prod_{j=1}^{p}\frac{dx_j}{\sigma_j}, \quad (41.1)$$

$$-\infty \leqslant x_j \leqslant \infty, \quad j=1, 2, \ldots, p;$$

$$dF = \frac{|\alpha|^{1/2}}{(2\pi)^{p/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})\boldsymbol{\alpha}(\boldsymbol{x}-\boldsymbol{\mu})\right\}\prod_j dx_j, \quad (41.2)$$

где μ_j, σ_j^2 — математическое ожидание и дисперсия j-й величины, $\boldsymbol{\alpha}$ — матрица, обратная матрице рассеяния, а α_{jk}^0 — элементы матрицы $\boldsymbol{\alpha}^0$, обратной к корреляционной матрице (ρ_{jk}).

В главе 15 мы мало касались проблемы выборок. Теперь мы будем различать между собой теоретические и выборочные характеристики, т. е. параметры и их оценки. Пусть a_{jk} — выборочное значение α_{jk}, c_{jk} — выборочная ковариация, соответствующая теоретическому значению γ_{jk}. Тогда

$$\gamma_{jk} = \rho_{jk}\sigma_j\sigma_k, \quad (41.3)$$

$$c_{jk} = r_{jk}s_js_k. \quad (41.4)$$

Матрицу рассеяния, которую мы в главе 15 обозначали $\boldsymbol{V}$, здесь обозначим символом $\boldsymbol{\gamma}$, так что

$$\boldsymbol{\alpha} = \boldsymbol{\gamma}^{-1}. \quad (41.5)$$

Из (15.15) (т. 1) следует, что характеристическая функция распределения (41.2) равна

$$\varphi(t) = \exp\{-\boldsymbol{t}'\boldsymbol{\gamma}\boldsymbol{t}/2\}\exp\{i\boldsymbol{t}'\boldsymbol{\mu}\}. \quad (41.6)$$

41.3 Выборочные значения x_{jl}, $l=1, 2, \ldots, n$, дают нам значение функции правдоподобия, которая равна произведению n сомножителей типа (41.1). Логарифм этой функции будет суммой n слагаемых. Символом S обозначим суммирование по элементам выборки, символом $\sum$ — суммирование компонент многомерных величин. Тогда

$$\frac{\partial}{\partial\mu_j}\log L = \mathrm{S}\sum_{k=1}\frac{\alpha_{jk}^0}{\sigma_j}\left(\frac{x_k-\mu_k}{\sigma_k}\right) = 0, \quad (41.7)$$

что приводит к уравнению

$$\sum\frac{\alpha_{jk}^0}{\sigma_j\sigma_k}(\bar{x}_k-\hat{\mu}_k) = 0. \quad (41.8)$$

Так как $\boldsymbol{\alpha}^0$ не вырождена, p таких равенств эквивалентны тому, что

$$\hat{\mu}_k = \bar{x}_k, \qquad k=1, 2, \ldots, p. \quad (41.9)$$

В том, что выборочное среднее является оценкой МП теоретического среднего, нет ничего удивительного.

41.4 Для параметра α_{jk} получим

$$\frac{n/2}{|\alpha|}\frac{\partial|\alpha|}{\partial\alpha_{jk}} - \frac{1}{2}\mathrm{S}(x_j-\mu_j)(x_k-\mu_k) = 0. \quad (41.10)$$

Пусть A_{jk} — алгебраическое дополнение α_{jk} в $|\alpha|$. Заменяя μ соответствующим ему $\bar{x}$, получим

$$\hat{A}_{jk}/|\hat{\alpha}| = \frac{1}{n}\mathrm{S}(x_j-\hat{\mu}_j)(x_k-\hat{\mu}_k) = c_{jk} \quad (41.11)$$

Отсюда следует *), что

$$\hat{\gamma}_{jk} = c_{jk}. \tag{41.12}$$

В частности, выборочные дисперсии являются оценками МП теоретических дисперсий, а для коэффициентов корреляций получим

$$\hat{\rho}_{jk} = r_{jk} = \frac{\sum (x_j - \bar{x}_j)(x_k - \bar{x}_k)}{\{\sum (x_j - \bar{x}_j)^2 \sum (x_k - \bar{x}_k)^2\}^{1/2}}. \tag{41.13}$$

Все сказанное справедливо, когда оцениваются все параметры. Иных случаев мы не будем касаться, так как они серьезного практического интереса не представляют. Можно рекомендовать читателю примеры 18.14—15 (т. 2) и упражнение 18.14, где рассматривается двумерный случай.

41.5 Строя доверительные интервалы для параметров, мы, желая получить распределения типа Стьюдента или χ^2, сталкиваемся с теми же трудностями, что и в одномерном случае. Кроме того, мы сталкиваемся с новой проблемой, связанной с построением доверительных интервалов одновременно для всех компонент вектора. В качестве примера рассмотрим оценивание средних в случае, когда теоретическая матрица рассеяния известна.

В главе 15 мы видели, что линейным преобразованием случайные величины в (41.1) могут быть приведены к независимым нормальным случайным величинам с единичными дисперсиями. Отсюда следует, что величина

$$(\boldsymbol{x} - \boldsymbol{\mu})' \boldsymbol{\alpha} (\boldsymbol{x} - \boldsymbol{\mu})$$

имеет χ^2-распределение с p степенями свободы.

В пункте **41.6** мы бегло докажем, что распределение средних многомерной нормальной выборки совпадает с распределением исходных переменных с той только разницей, что матрица рассеяния здесь равна $\boldsymbol{\alpha}/n$. Отсюда и из (41.5) следует, что величина

$$n(\bar{\boldsymbol{x}} - \boldsymbol{\mu}')\boldsymbol{\gamma}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}) \tag{41.14}$$

имеет χ^2-распределение с p степенями свободы. Следовательно, и при заданном уровне вероятности P справедливо утверждение

$$\mathsf{P}\left(n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})' \boldsymbol{\gamma}^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}) \geqslant \chi_P^2\right) = P. \tag{41.15}$$

*) Вероятно, это не совсем очевидно. Мы использовали здесь просто доказываемую теорему (см. 8.9) о том, что если $\varphi_1, \ldots, \varphi_m$ — функции параметров $\theta_1, \ldots, \theta_m$, то оценки МП величин φ получаются, если в функциональных связях параметры θ заменить их оценками МП.

Отсюда также следует, что оценки МП частных и множественных корреляций даются соответствующими выборочными статистиками.

Так как, по предположению, $\boldsymbol{\gamma}$ известно, последнее неравенство и дает нам доверительную область. Последняя ограничена p-мерной поверхностью второго порядка. Практическое использование полученных результатов требует некоторой осторожности. Вопрос о смещении оценок мы обсудим в следующей главе, поскольку сейчас нас интересуют распределения.

Распределение Уишарта

41.6 Займемся исследованием совместного распределения выборочных средних и характеристик рассеяния для многомерного нормального случая. Предположим, что мы располагаем случайной выборкой, состоящей из n независимых векторов. Пусть x_{jl} — l-е наблюдение j-й координаты. Тогда

$$\boldsymbol{x} = (x_{jl}) = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1n} \\ x_{21} & x_{22} & \ldots & x_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ x_{p1} & x_{p2} & \ldots & x_{pn} \end{pmatrix}. \tag{41.16}$$

Распределение выборки дается формулой

$$dF = \frac{|\boldsymbol{\alpha}|^{n/2}}{(2\pi)^{np/2}} \exp\left\{-\frac{1}{2}\sum_{l=1}^{n}\sum_{j,k=1}^{p} \alpha_{jk}(x_{jl} - \mu_j)(x_{kl} - \mu_k)\right\} \prod_{l=1}^{n}\prod_{j=1}^{p} dx_{jl}. \tag{41.17}$$

Из примера 11.7 и **16.25** (т. 1) нам уже известно, что при $p = 1, 2$ это распределение можно «разложить» на два независимых распределения: одно для средних и другое для ковариаций. Прежде всего мы докажем, что то же самое верно и при любом p. Воспользуемся известным алгебраическим тождеством

$$\sum_l \sum_{j,k} \alpha_{jk}(x_{jl} - \mu_j)(x_{kl} - \mu_k) \equiv$$
$$\equiv \sum \sum \alpha_{jk}(x_{jl} - \bar{x}_j)(x_{kl} - \bar{x}_k) + n \sum \alpha_{jk}(\bar{x}_j - \mu_j)(\bar{x}_k - \mu_k). \tag{41.18}$$

Тогда экспоненциальный член в (41.17) разлагается на два сомножителя. Обратимся теперь к дифференциалам. Здесь удобно применить к переменным ортогональное преобразование Хельмерта того же типа, что использовалось в примере 11.3 (т. 1):

$$\left.\begin{aligned} y_1 &= (x_1 - x_2)/\sqrt{2}, \\ y_2 &= (x_1 + x_2 - 2x_3)/\sqrt{6}, \\ &\cdots\cdots\cdots\cdots \\ y_{n-1} &= \{x_1 + x_2 + \ldots + x_{n-1} - (n-1)x_n\}/\sqrt{n(n-1)}, \\ y_n &= \{x_1 + x_2 + \ldots + x_n\}/\sqrt{n} = \sqrt{n}\,\bar{x}. \end{aligned}\right\} \tag{41.19}$$

(Индексы означают номера выборок.) Обратные соотношения таковы:

$$\left.\begin{aligned} x_1 &= \frac{y_1}{\sqrt{2}} + \frac{y_2}{\sqrt{6}} + \ldots + \frac{y_{n-1}}{\sqrt{n(n-1)}} + \frac{y_n}{\sqrt{n}}, \\ x_2 &= -\frac{y_1}{\sqrt{2}} + \frac{y_2}{\sqrt{6}} + \ldots + \frac{y_{n-1}}{\sqrt{n(n-1)}} + \frac{y_n}{\sqrt{n}}, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n &= -\frac{n-1}{\sqrt{n(n-1)}} y_{n-1} + \frac{y_n}{\sqrt{n}}. \end{aligned}\right\} \quad (41.20)$$

Так как $|J| = 1$, дифференциальный элемент в (41.17) равен просто

$$\prod_{l=1}^{n} \prod_{j=1}^{p} dy_{jl}. \qquad (41.21)$$

Обратившись к (41.18) и вспоминая, что $y_n = \sqrt{n}\,\bar{x}$, мы видим, что второй сомножитель зависит лишь от y_{jn} $(j = 1, \ldots, p)$. В силу (41.20) первый сомножитель зависит только от $y_1, \ldots, y_{n-1}$.

Следовательно, мы можем выделить в первоначальном представлении сомножитель, соответствующий второму члену в (41.18), а также соответствующий дифференциальный элемент. Подобрав необходимый постоянный сомножитель, получаем, что для средних

$$dF = \frac{|\alpha|^{1/2}}{(2\pi)^{p/2}} \exp\left\{-\frac{n}{2}\sum_{j,k} \alpha_{jk}(\bar{x}_j - \mu_j)(\bar{x}_k - \mu_k)\right\} \prod_{j=1}^{p} d(\sqrt{n}\,\bar{x}_j). \quad (41.22)$$

В формулах присутствует величина n, во всем остальном совместное распределение для средних совпадает с распределением исходных переменных.

41.7 Таким образом, распределение выборочных дисперсий и ковариаций сосредоточено на произведении $(n-1)$-мерных пространств, ортогональных вектору выборочных средних. Так как ортогональное преобразование представляет собой просто поворот координатных осей, оно не меняет расстояния и углы. А поскольку лишь от них зависят дисперсии и ковариации (см. пример 11.7 и **16.24**), последние также остаются неизменными. Из (41.18) и (41.22) получаем, что «недифференциальная часть» распределения равна

$$f = \frac{|\alpha|^{(n-1)/2}}{(2\pi)^{(n-1)p/2}} \exp\left\{-\frac{n}{2}\sum_{jk} \alpha_{jk} c_{jk}\right\}. \qquad (41.23)$$

Наша основная задача — выразить дифференциальный элемент на языке c_{jk}. Пусть

$$\{u_{j1}, \ldots, u_{j(n-1)}\} \qquad (j = 1, \ldots, p) \qquad (41.24)$$

— результат некоторого ортогонального преобразования (вид которого мы уточним позднее) $(n-1)$-мерного вектора $\{y_{j1}, \ldots, y_{j(n-1)}\}$. Тогда ковариация c_{jk} дается формулой (см. также (41.20))

$$nc_{jk} = \sum_{l=1}^{n-1} u_{jl} u_{kl}. \qquad (41.25)$$

Заметим, что здесь и в дальнейшем мы будем понимать под n не степень свободы характеристик рассеяния, а размер выборки. Потребуем, чтобы $n - 1 \geqslant p$.

Следуя рассуждениям **16.24** (т. 1), рассмотрим p $(n-1)$-мерных гиперплоскостей, по одной для каждого u. Точки, соответствующие выборкам, обозначим $P_1, \ldots, P_p$. Будем по очереди варьировать P_1, затем P_2 при условии, что P_1 фиксировано, затем P_3 при условии, что фиксированы P_2 и P_1, и т. д. Для того чтобы изучить полную вариацию $P_1, \ldots, P_p$, объединим полученное.

Рассмотрим P_m при заданных $P_1, \ldots, P_{m-1}$. Пусть O — начало координат. Представим себе, что наши пространства (гиперплоскости) совмещены. При фиксированном OP_m и фиксированных углах $P_mOP_1, P_mOP_2, \ldots, P_mOP_{m-1}$ точка P_m расположена на сфере размерности $n - m$. Пусть t_m — длина перпендикуляра, опущенного из точки P_m на гиперплоскость, порожденную $O, P_1, \ldots, P_{m-1}$. Тогда «объем», в котором происходят допустимые изменения P_m, равен «поверхности» $(n-m)$-мерной гиперсферы радиуса t_m, которая равна *)

$$\frac{2\pi^{(n-m)/2} t_m^{n-m-1}}{\Gamma\{(n-m)/2\}}. \qquad (41.26)$$

Нам нужно умножить это на «дифференциальный элемент», соответствующий направлениям, перпендикулярным гиперсфере. В m-мерном пространстве, порожденном $O, P_1, \ldots, P_m$, рассмотрим преобразование, основанное на (41.25):

$$\xi_{mj} = nc_{mj} = \sum_{k=1}^{n-1} u_{mk} u_{jk}, \qquad j = 1, 2, \ldots, m. \qquad (41.27)$$

При этом ортогональное преобразование, соответствующее величинам u (см. (41.24)), будем считать как раз таким, чтобы величины $u_{m1}, \ldots, u_{mm}$ соответствовали координатам рассмат-

*) См. Kendall M. G., A course in the Geometry of n Dimensions, стр. 42.

риваемого m-мерного пространства. Тогда якобиан преобразования (41.27) равен

$$J=\frac{\partial(\xi_{m1},\ldots,\xi_{mm})}{\partial(u_{m1},\ldots,u_{mm})}=\begin{vmatrix} u_{11} & u_{12} & \ldots & u_{1m} \\ u_{21} & u_{22} & \ldots & u_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ 2u_{m1} & 2u_{m2} & \ldots & 2u_{mm} \end{vmatrix}. \qquad (41.28)$$

Последнее равно $2v_m$, где v_m — объем параллелепипеда (m-мерный параллелограмм), порожденного O, P_1, ..., P_m. Следовательно, дифференциальный элемент равен

$$\frac{1}{2v_m}\prod_{j=1}^{m} d\xi_{mj}. \qquad (41.29)$$

Умножая это на (41.26), получаем, что полная вариация P_m равна

$$\frac{\pi^{(n-m)/2}t_m^{n-m-1}}{\Gamma\{(n-m)/2\}\,v_m}\prod_{j=1}^{m} d\xi_{mj}. \qquad (41.30)$$

Но

$$t_m=v_m/v_{m-1}, \qquad (41.31)$$

и в силу (41.27)

$$|\xi_{kj}|_{(m)}=|u_{kj}|^2_{(m)}=v_m^2, \qquad (41.32)$$

где $|\xi_{kj}|_{(m)}$ — определитель матрицы (ξ_{kj}) при $k, j=1, \ldots, m$. То же относится и к $|u_{kj}|_{(m)}$. Таким образом, (41.30) равно

$$\frac{\pi^{(n-m)/2}}{\Gamma\{(n-m)/2\}}\cdot\frac{v_m^{n-m-2}}{v_{m-1}^{n-m-1}}\prod_{j=1}^{m} d\xi_{mj}. \qquad (41.33)$$

Перемножим теперь выражения типа (41.33) при $m=1, 2, \ldots, p$. Степени v сократятся, за исключением v_0 (которое равно единице) и v_p, и мы придем к выражению

$$\frac{\pi^{p(2n-p-1)/4}}{\prod_{j=1}^{p}\Gamma\{(n-j)/2\}}\,v_p^{n-p-2}\prod_{j=1}^{m}\prod_{k=m}^{p} d\xi_{jk}. \qquad (41.34)$$

Из (41.32) следует, что

$$v_p^2=|\xi_{jk}|=n^p|c_{jk}|. \qquad (41.35)$$

Объединяя (41.34—5) и (41.23), получаем распределение характеристик рассеяния:

$$dF=\frac{(n/2)^{p(n-1)/2}|\alpha|^{(n-1)/2}\cdot|c|^{(n-p-2)/2}}{\pi^{p(p-1)/4}\prod_{j=1}^{p}\Gamma\{(n-j)/2\}}\times$$
$$\times\exp\left\{-\frac{n}{2}\sum\alpha_{jk}c_{jk}\right\}\prod_{j\leqslant k}^{p} dc_{jk}. \qquad (41.36)$$

Это распределение Уишарта (1928). Равенство (11.41) в примере 11.17 соответствует $p=1$, (16.54) — слегка преобразованное выражение, соответствующее $p=2$.

Читателю, который найдет наше «геометрическое» доказательство сложным, можно рекомендовать обзор иных доказательств, содержащийся в статье Уишарта (1948). С самых разных точек зрения описанное распределение можно рассматривать как обобщение χ^2-распределения на p-мерный случай.

41.8 Отметим ряд дополнительных, но немаловажных свойств распределения:

(а) Показатель экспоненты в (41.36) содержит суммы по всем j, k, и, так как $\alpha_{kj}=\alpha_{jk}$, $c_{kj}=c_{jk}$, все члены суммы появляются парами, за исключением случая $j=k$. Поэтому существует ровно p членов типа $\alpha_{11}c_{11}$ и ровно $p(p-1)/2$ членов типа $\alpha_{12}c_{12}$. Например, при $p=2$ показатель экспоненты равен

$$-n(\alpha_{11}c_{11}+2\alpha_{12}c_{12}+\alpha_{22}c_{22})/2.$$

(б) Дифференциальный элемент содержит $p(p+1)/2$ сомножителей, а не p^2. Так, при $p=2$ этот элемент равен $dc_{11}dc_{12}dc_{22}$.

(в) Каждое из c_{jj} может меняться от 0 до ∞, для остальных c область изменения определить нелегко, так как мы связаны условием положительной определенности матрицы (c_{jk}). Поэтому довольно трудно провести интегрирование по каким-либо отдельным переменным для того, чтобы получить распределение остальных.

(г) Определяя выборочные дисперсии и ковариации, мы делили соответствующие суммы произведений на n. Можно производить деление и на $n-1$; в этом случае надо соответствующим образом откорректировать (41.36). Читатель может встретить этот вариант в литературе, поскольку общеупотребительного подхода здесь нет.

41.9 Мы можем теперь вычислить характеристическую функцию распределения Уишарта. Употребляя один знак интеграла для обозначения интегрирования по всей области изменения величин c, из (41.36) получаем

$$\int|c|^{(n-p-2)/2}\exp\left\{-n\sum\alpha_{jk}c_{jk}/2\right\}\prod dc_{jk}=k|\alpha|^{-(n-1)/2}, \qquad (41.37)$$

где k — некоторая постоянная. Если мы заменим α_{jj} на $\alpha_{jj} - 2\theta_{jj}/n$, α_{jk} на $\alpha_{jk} - \theta_{jk}/n$, $j \neq k$, а под обозначением θ_{jk} будем понимать обычную мнимую переменную it_{jk}, выписанный интеграл даст нам х. ф. для c. Производя соответствующую замену в правой части (41.37) и подбирая постоянную так, чтобы φ равнялась единице при $\theta = 0$, получаем

$$\varphi(\theta) = |\alpha|^{(n-1)/2} \cdot \begin{vmatrix} \alpha_{11} - 2\theta_{11}/n & \alpha_{12} - \theta_{12}/n & \dots & \alpha_{1p} - \theta_{1p}/n \\ \alpha_{21} - \theta_{21}/n & \alpha_{22} - 2\theta_{22}/n & \dots & \alpha_{2p} - \theta_{2p}/n \\ \dots & \dots & \dots & \dots \\ \alpha_{p1} - \theta_{p1}/n & \alpha_{p2} - \theta_{p2}/n & \dots & \alpha_{pp} - 2\theta_{pp}/n \end{vmatrix}^{-(n-1)/2}. \tag{41.38}$$

Этот прием замены, позволяющий избежать непосредственного интегрирования, весьма полезен и будет использоваться в дальнейшем.

Пример 41.1

В двумерном случае ($p = 2$), используя обычные обозначения (дисперсии единичны), имеем

$$\gamma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \quad |\gamma| = 1 - \rho^2,$$

$$\alpha = \begin{vmatrix} \frac{1}{1-\rho^2}, & -\frac{\rho}{1-\rho^2} \\ -\frac{\rho}{1-\rho^2}, & \frac{1}{1-\rho^2} \end{vmatrix}, \quad |\alpha| = (1-\rho^2)^{-1}.$$

Таким образом, (41.36) сводится к виду

$$dF = \frac{(n/2)^{n-1}}{(1-\rho^2)^{(n-1)/2}} \cdot \frac{s_1^{n-4} s_2^{n-4} (1-r^2)^{(n-4)/2}}{\pi^{1/2}\Gamma\{(n-1)/2\}\Gamma\{(n-2)/2\}} \times$$
$$\times \exp\left\{\frac{-n}{2(1-\rho^2)}(s_1^2 - 2\rho r s_1 s_2 + s_2^2)\right\} d(s_1^2)\, d(s_2^2)\, d(r s_1 s_2).$$

Используя формулу удвоения

$$\Gamma\{(n-1)/2\}\Gamma\{(n-2)/2\} = \pi^{1/2}\Gamma(n-2)/2^{n-3},$$

получаем

$$dF = \frac{n^{n-1}}{(1-\rho^2)^{(n-1)/2}} \frac{s_1^{n-4} s_2^{n-4}(1-r^2)^{(n-4)/2}}{4\pi\Gamma(n-2)} \times$$
$$\times \exp\left\{\frac{-n}{2(1-\rho^2)}(s_1^2 - 2\rho r s_1 s_2 + s_2^2)\right\} d(s_1^2)\, d(s_2^2)\, d(r s_1 s_2). \tag{41.39}$$

Последнему выражению легко придать форму, встречавшуюся в **16.26.**

Пример 41.2

Займемся вычислением моментов распределения ковариаций при $p = 2$. Из (41.38), положив $\theta_{11} = \theta_{22} = 0$, получаем

$$\varphi(\theta_{12}) \propto \begin{vmatrix} \frac{1}{1-\rho^2} & -\frac{\rho}{1-\rho^2} - \frac{\theta_{12}}{n} \\ -\frac{\rho}{1-\rho^2} - \frac{\theta_{12}}{n} & \frac{1}{1-\rho^2} \end{vmatrix}^{-(n-1)/2}. \tag{41.40}$$

Вычислив определитель и подобрав постоянную из условия $\varphi(0) = 1$, получаем

$$\varphi(\theta_{12}) = \left\{1 - \frac{2\rho\theta_{12}}{n} - \frac{(1-\rho^2)\theta_{12}^2}{n^2}\right\}^{-(n-1)/2}. \tag{41.41}$$

Беря логарифм и вычисляя коэффициенты при θ_{12}, находим

$$\varkappa_1 = \frac{n-1}{n}\rho, \tag{41.42}$$

$$\varkappa_2 = \frac{n-1}{n^2}(1+\rho^2), \tag{41.43}$$

$$\varkappa_3 = \frac{2(n-1)}{n^3}\rho(3+\rho^2), \tag{41.44}$$

$$\varkappa_4 = \frac{6(n-1)}{n^4}(1+6\rho^2+\rho^4). \tag{41.45}$$

При стандартной нормировке и n, стремящемся к бесконечности, распределение сходится к нормальному. При конечном n

$$\beta_1^2 = \frac{4}{n-1}\frac{\rho^2(3+\rho^2)^2}{(1+\rho^2)^3}, \tag{41.46}$$

$$\beta_2 = 3 + \frac{6}{n-1}\cdot\frac{1+6\rho^2+\rho^4}{(1+\rho^2)^2}\ ^{*)}. \tag{41.47}$$

Так что, даже если $\rho = 0$, распределение хотя и симметрично, но не является нормальным.

В примере 13.3 (т. 1) мы получили эти результаты иным путем.

Уишарт (1929) получил точные формулы, вплоть до четвертого порядка, для $p \leqslant 8$.

Аддитивное свойство распределения Уишарта

41.10 Следует отметить свойство распределения Уишарта, аналогичное одному свойству χ^2 в одномерном случае.

Предположим, что мы располагаем двумя независимыми выборками размера n_1 и n_2, соответствующими одному и тому

*) Под β_1 и $\beta_2 - 3$ автор подразумевает соответственно коэффициенты асимметрии и эксцесса, см. т. 1, стр. 125. (*Прим. ред.*)

же многомерному распределению. Если мы объединим их, характеристики рассеяния объединенной выборки будут, конечно, подчиняться распределению Уишарта для выборки размера n_1+n_2. Но мы можем рассмотреть также распределение характеристик рассеяния для каждой выборки. Образуем новую матрицу рассеяния, складывая соответствующие характеристики:

$$c_{jk}=c_{jk}^{(1)}+c_{jk}^{(2)}, \tag{41.48}$$

где верхние индексы соответствуют номеру выборки. Тогда распределение c_{jk} совпадает с распределением Уишарта, если в последнем n заменить на n_1+n_2-1.

Наиболее просто это, пожалуй, получить, исследуя характеристические функции. Если рассмотреть совместное распределение величин $c^{(1)}$ и $c^{(2)}$ и выписать соответствующий аналог (41.37), то легко увидеть, что х. ф. величин c сама представима в виде (41.38).

41.11 Выборочная теория характеристик рассеяния стала бы более полной и более красивой, если бы нам удалось, оттолкнувшись от распределения Уишарта, получить распределения каких-либо функций от дисперсий и ковариаций, интегрируя по соответствующим областям так, чтобы исключить ненужные нам переменные. К сожалению, в общем случае задача оказывается исключительно сложной. Однако мы можем, как и в примере 41.2, вычислить моменты и смешанные моменты характеристик рассеяния. Как правило, этого оказывается достаточным, и точные распределения не нужны. Тем не менее существует еще ряд ситуаций, где можно добиться определенного успеха.

Моменты определителя рассеяния (обобщенной дисперсии)

41.12 Рассмотрим распределение определителя рассеяния $|c|$. В силу (41.36) взятый по дисперсиям и ковариациям интеграл

$$\int \frac{(n/2)^{p(n-1)/2}\,|\alpha|^{(n-1)/2}\,|c|^{(n-p-2)/2}}{\pi^{p(p-1)/4}\prod\limits_{j=1}^{p}\Gamma\{(n-j)/2\}}\exp\left\{-n\sum\alpha_{jk}c_{jk}/2\right\}\prod_{j\leqslant k}^{p}dc_{jk}=1. \tag{41.49}$$

Пусть $n\alpha_{jk}=\beta_{jk}$. Тогда

$$\int \frac{|\beta|^{(n-1)/2}\,|c|^{(n-p-2)/2}}{\pi^{p(p-1)/4}}\exp\left\{-\sum\beta_{jk}c_{jk}/2\right\}\prod dc_{jk}=$$

$$=2^{p(n-1)/2}\prod_{j=1}^{p}\Gamma\{(n-j)/2\}. \tag{41.50}$$

Заменим n на $n+2t$ и разделим все на $|\beta|^t$. Имеем

$$\int \frac{|\beta|^{(n-1)/2}\,|c|^{t+(n-p-2)/2}}{\pi^{p(p-1)/4}}\exp\left\{-\sum\beta_{jk}c_{jk}/2\right\}\prod dc_{jk}=$$

$$=2^{pt+p(n-1)/2}\cdot|\beta|^{-t}\prod_{j=1}^{p}\Gamma\{t+(n-j)/2\}. \tag{41.51}$$

Если мы теперь заменим β_{jk} на $n\alpha_{jk}$ и разделим все на $2^{p(n-1)/2}\Pi\Gamma\{(n-j)/2\}$, мы получим слева математическое ожидание $|c|^t$. Итак

$$\mathsf{M}(|c|^t)=\frac{2^{pt}}{n^{pt}|\alpha|^t}\frac{\prod\limits_{j=1}^{p}\Gamma\{t+(n-j)/2\}}{\prod\limits_{j=1}^{p}\Gamma\{(n-j)/2\}}= \tag{41.52}$$

$$=\frac{2^{pt}}{n^{pt}}\prod_{j=1}^{p}\frac{\Gamma\{t+(n-j)/2\}}{\Gamma\{(n-j)/2\}}|\gamma|^t. \tag{41.53}$$

Отсюда можно получить моменты и семиинварианты какого угодно порядка. Отметим, что прием замены переменных снова помог нам избежать вычисления весьма громоздких интегралов.

41.13 Следует отметить еще одно следствие из (41.53). Пусть d_{jk} равно сумме произведений отклонений от среднего, так что $d_{jk}=nc_{jk}$. Тогда

$$\mathsf{M}\left(\frac{|d|^t}{|\gamma|^t}\right)=2^{pt}\prod_{j=1}^{p}\frac{\Gamma\{t+(n-j)/2\}}{\Gamma\{(n-j)/2\}}. \tag{41.54}$$

Моменты относительно нуля случайной величины χ^2 с ν степенями свободы равны

$$\mu_t'=2^t\frac{\Gamma(t+\nu/2)}{\Gamma(\nu/2)}.$$

Правая часть (41.54) равна произведению p таких сомножителей с $\nu=n-1,\ n-2,\ \ldots,\ n-p$. Вспомнив, что момент произведения независимых случайных величин равен произведению их моментов, получаем, что $|d|/|\gamma|$ может быть представлено как произведение p независимых сомножителей, распределенных как χ^2 с $n-1,\ldots,n-p$ степенями свободы.

Пример 41.3

При $p=1$ мы сталкиваемся с тем известным фактом, что нормированная (точнее, деленная на теоретическую дисперсию) сумма квадратов наших величин распределена как χ^2 с $n-1$ ст. св.

При $p=2$ для $|d|/|\gamma|$ из (41.54) находим, что

$$\mu_t' = \frac{\Gamma\{t+(n-1)/2\}}{\Gamma\{(n-1)/2\}} \cdot \frac{\Gamma\{t+(n-2)/2\}}{\Gamma\{(n-2)/2\}} 2^{2t}. \qquad (41.55)$$

Используя формулу удвоения для гамма-функций

$$\Gamma(x)\Gamma(x+1/2) = \pi^{1/2}\Gamma(2x)/2^{2x-1}, \qquad (41.56)$$

приводим (41.55) к виду

$$\mu_t' = \Gamma(n-2+2t)/\Gamma(n-2). \qquad (41.57)$$

Последнее является моментом порядка $2t$ величины, имеющей гамма-распределение с параметром $n-2$ (см. пример 4.4 т. 1). Поэтому момент порядка t величины $2|d|/|\gamma|$ равен моменту порядка $2t$ величины χ^2 с $2(n-2)$ ст. св. Это частный случай результата примера 11.9 т. 1.

Корреляционный определитель

41.14 Значительный интерес представляет изучение совместного распределения выборочных корреляций. В общем случае (ненулевых теоретических корреляций) распределение весьма сложно даже при $p=2$ (см. **16.26—7**). В «нулевом» случае, т. е. в случае, когда теоретические корреляции равны нулю, можно добиться некоторого успеха. Распределение Уишарта (41.36) тогда сводится к распределению

$$dF = \frac{(n/2)^{p(n-1)/2}\,|c|^{(n-p-2)/2}}{\pi^{p(p-1)/4}\Pi\Gamma\{(n-j)/2\}} \exp\left\{-n\sum_{j=1}^{p} c_{jj}/2\right\} \prod_{j\leqslant k} dc_{jk}. \qquad (41.58)$$

Перейдем к новым переменным с помощью равенств типа

$$c_{jk} = s_j s_k r_{jk}. \qquad (41.59)$$

Якобиан

$$J = 2^p \prod_{j=1}^{p} s_j^p. \qquad (41.60)$$

Он не зависит от r_{jk}, так же как и показатель степени в (41.58). Следовательно, члены, зависящие от s, могут быть факторизованы (выделены в качестве сомножителей), в результате чего мы получим распределение корреляций:

$$dF = \frac{|r|^{(n-p-2)/2}\,[\Gamma\{(n-1)/2\}]^p}{\pi^{p(p-1)/4}\Pi\Gamma\{(n-j)/2\}} \prod_{j<k} dr_{jk}, \qquad (41.61)$$

где $|r|$ — определитель выборочной матрицы корреляций, а постоянная подобрана так, чтобы интеграл от плотности был равен единице. Дальнейшему прогрессу вновь мешает сложность

области интегрирования. Однако мы можем найти моменты $|r|$, поступая в духе пункта **41.12**. Моменты

$$\mathbf{M}|r|^t = \frac{[\Gamma\{(n-1)/2\}]^p}{[\Gamma\{t+(n-1)/2\}]^p} \cdot \frac{\prod_{j=1}^{p}\Gamma\{t+(n-j)/2\}}{\prod_{j=1}^{p}\Gamma\{(n-j)/2\}}. \qquad (41.62)$$

Пример 41.4

Обозначив $L=\log|r|$, $\nu=(n-1)/2$, $m=(j-1)/2$, из (41.62) получаем, что

$$\mathbf{M}(e^{Lt}) = \prod_{j=1}^{p} \frac{\Gamma(\nu)}{\Gamma(\nu+t)} \cdot \frac{\Gamma(\nu-m+t)}{\Gamma(\nu-m)}. \qquad (41.63)$$

Правда, последнее было доказано только для целого t, но к общему случаю всех t можно перейти с помощью аналитического продолжения. Если мы теперь будем рассматривать t как мнимую величину, (41.63) даст нам характеристическую функцию L. Тогда производящая функция семиинвариантов равна

$$\psi = \sum_{j=1}^{p}\{\log\Gamma(\nu) + \log\Gamma(\nu-m+t) - \log\Gamma(\nu+t) - \log\Gamma(\nu-m)\} \qquad (41.64)$$

Воспользуемся разложением

$$\log\Gamma(x) = \\ = (x-1/2)\log x - x + \frac{1}{2}\log(2\pi) + \frac{1}{12x} - \frac{1}{360x^3} + \ldots \qquad (41.65)$$

При больших ν подстановка в (41.64) даст нам для члена в скобках представление

$$-\frac{mt}{\nu} + \frac{1}{\nu^2}\{-m(m+1)t/2 + mt^2/2\} + O(\nu^{-3})\}. \qquad (41.66)$$

Положим теперь $m=(j-1)/2$ и возьмем сумму по j от 1 до p. Получаем

$$\psi = -p(p-1)t/4\nu + O(t/\nu^2) + p(p-1)t^2/8\nu^2 + O(\nu^{-3}). \qquad (41.67)$$

Отсюда

$$\mathbf{M}L = -\frac{p(p-1)}{4\nu} + O(\nu^{-2}), \qquad (41.68)$$

$$\mathbf{D}L = \frac{p(p-1)}{4\nu^2} + O(\nu^{-3}). \qquad (41.69)$$

Далее

$$\left(\frac{d}{dx}\right)^k \log\Gamma(x) = \frac{(-1)^{k-2}(k-2)!}{x^{k-1}} + o(x^{-k+1}), \qquad k>1. \qquad (41.70)$$

Следовательно, из (41.64), положив $(j-1)/2=m$, с точностью до «больших степеней ν» получаем, что

$$\left.\frac{d^k\psi}{dt^k}\right|_{t=0}=\sum_{j=1}^{p}(-1)^{k-2}(k-2)!\left\{\frac{1}{(\nu-m+t)^{k-1}}-\frac{1}{(\nu+t)^{k-1}}\right\}_{t=0}=$$

$$=\sum_{j=1}^{p}\frac{(-1)^{k-2}(k-1)!}{\nu^k}m=(-2\nu)^{-k}2^{k-1}(k-1)!\,p(p-1)/2. \quad (41.71)$$

Сравнение с равенством (16.4) (т. 1) показывает, что при подходящем выборе начала координат величина $-2\nu\log|r|$ асимптотически распределена как χ^2 с $p(p-1)/2$ степенями свободы. Из (41.68) видно, что с точностью до порядка ν^{-1} среднее (иначе сказать, новое начало отсчета) равно нулю.

Итак, величина $-(n-1)\log|r|$ асимптотически распределена как χ^2 с $p(p-1)/2$ ст. св. Бартлетт (1951) получил более тонкий результат: он показал, что

$$-\{n-(2p+1)/6\}\log|r|$$

распределено как χ^2 с $p(p-1)/2$ ст. св. Дополнительный член получается, если учесть в среднем члены порядка n^{-1}. Для практики такое уточнение не имеет большого значения.

Пример 41.5

Интересно сравнить полученные в предыдущем примере результаты для определителя корреляций в «нулевом случае» с теми, что мы получили в **41.13** для определителя рассеяния.

Не теряя общности, предположим, что теоретические дисперсии равны единице, а теоретические коэффициенты корреляции — нулю. Тогда, из **41.13** получаем, что величина

$$n^p|c|=(ns_1^2)(ns_2^2)\ldots(ns_p^2)|r| \quad (41.72)$$

равна произведению p независимых χ^2-величин с $n-1$, $n-2$, ..., $n-p$ степенями свободы, в то время как величина $-(n-1)\log|r|$ асимптотически распределена как χ^2 с $p(p-1)/2$ степенями свободы.

Теперь, применяя (27.61) к выборочному, а не теоретическому распределению, получаем, что в наших обозначениях

$$1-R^{(2)}_{1(2\ldots p)}=|r|/R_{11}, \quad (41.73)$$

где $R_{1(2\ldots p)}$ — множественный коэффициент корреляции величины x_1 по отношению к x_2, ..., x_p, а R_{11} — алгебраическое дополнение r_{11} $(=1)$ в определителе корреляций. Многократно используя эту формулу, каждый раз переупорядочивая индексы,

получаем, что

$$\{1-R^2_{p(12\ldots(p-1))}\}\{1-R^2_{(p-1)(12\ldots(p-2))}\}\ldots\{1-R^2_{2(1)}\}=|r|, \quad (41.74)$$

где $R^2_{2(1)}$ — то же самое, что и корреляция нулевого порядка r^2_{21}. Кроме того, все x независимы, и, следовательно, все сомножители в левой части (41.74) независимы (см. **27.30**).

Распределение $U=1-R^2$, основанное на выборке n и q величин, равно (см. (27.74))

$$dF=\frac{(1-U)^{(q-3)/2}U^{(n-q-2)/2}}{B\{(q-1)/2,(n-q)/2\}}dU. \quad (41.75)$$

Таким образом, $|r|$ распределен в точности как произведение $p-1$ независимых бетта-величин с параметрами $((n-q)/2, (q-1)/2)$, $q=2,\ldots,p$.

Поступая так же, как и в предыдущем примере, мы можем найти моменты U, а следовательно, и характеристическую функцию $\log U$. На самом деле известно, что $-(n-1)\log U$ распределено приблизительно как χ^2 с $q-1$ степенями свободы.

Таким образом, $-(n-1)\log|r|$ приближенно равно сумме p независимых χ^2-величин с $(p-1)$, ..., 1 ст. св., а следовательно, равно χ^2-величине с $p(p-1)/2$ ст. св. Это согласуется с результатом, который мы получили в предыдущем примере.

Отношение $|r|/R_{11}$ равно также (см. (27.34)) $s^2_{1\cdot2\ldots p}/s^2_1$. В R_{11} соответствующее отношение равно $s^2_{2\cdot3\ldots p}/s^2_2$ и т. д. Итак, из (41.72—4) получаем

$$|n|^p|c|=\{ns^2_{1\cdot2\ldots p}\}\{ns^2_{2\cdot3\ldots p}\}\ldots\{ns^2_p\}. \quad (41.76)$$

Суммы квадратов типа $ns^2_{j(k)}$ представляют собой остатки, которые независимы и имеют соответственно $n-p$, $n-p+1$, ..., $n-1$ ст. св.

Таким образом, $n^p|c|$ равно произведению независимых χ^2-величин с теми же степенями свободы. Это подтверждает результаты **41.13**.

T^2-статистика Хотеллинга

41.15 Рассмотрим теперь полученное Хотеллингом обобщение отношения t Стьюдента *). Пусть, как и в **41.13** $d_{jk}=nc_{jk}$, а (D_{jk}) — матрица, обратная (d_{jk}). Положим

$$T^2=n(n-1)\sum_{j,k=1}^{p}D_{jk}\bar{x}_j\bar{x}_k. \quad (41.77)$$

*) В данном параграфе автор подразумевает (не оговаривая этого), что вектор теоретических средних исследуемой многомерной нормальной величины является «нулевым». (*Прим. ред.*)

При $p=1$ имеем

$$d_{11}=ns_1^2,\quad D_{11}=1/(ns_1^2),$$
$$T^2=\frac{n(n-1)\bar{x}^2}{ns_1^2}=\frac{(n-1)\bar{x}^2}{s_1^2}=t^2.$$

Мы видим, что в одномерном случае T сводится к статистике t Стьюдента. Пусть m_{jk} — суммы произведений отклонений от начала координат:

$$m_{jk}=d_{jk}+n\bar{x}_j\bar{x}_k. \tag{41.78}$$

Определитель $|(m_{jk})|$ может быть записан в виде

$$\begin{vmatrix} 1 & \bar{x}_1\sqrt{n} & \bar{x}_2\sqrt{n} & \dots & \bar{x}_p\sqrt{n} \\ 0 & d_{11}+n\bar{x}_1^2 & d_{12}+n\bar{x}_1\bar{x}_2 & \dots & d_{1p}+n\bar{x}_1\bar{x}_p \\ 0 & d_{21}+n\bar{x}_1\bar{x}_2 & d_{22}+n\bar{x}_2^2 & \dots & d_{2p}+n\bar{x}_2\bar{x}_p \\ \dots & \dots & \dots & \dots & \dots \\ 0 & d_{p1}+n\bar{x}_1\bar{x}_p & d_{p2}+n\bar{x}_p\bar{x}_2 & \dots & d_{pp}+n\bar{x}_p^2 \end{vmatrix}. \tag{41.79}$$

Вычитая из второй строки первую, помноженную на $\bar{x}_1\sqrt{n}$, из третьей строки первую, помноженную на $\bar{x}_2\sqrt{n}$, и так далее, находим

$$|(m_{jk})|=\begin{vmatrix} 1 & \bar{x}_1\sqrt{n} & \dots & \bar{x}_p\sqrt{n} \\ -\bar{x}_1\sqrt{n} & d_{11} & \dots & d_{1p} \\ -\bar{x}_2\sqrt{n} & d_{21} & \dots & d_{2p} \\ \dots & \dots & \dots & \dots \\ -\bar{x}_p\sqrt{n} & d_{p1} & \dots & d_{pp} \end{vmatrix}. \tag{41.80}$$

Раскладывая по первой строке и первому столбцу, получаем

$$|(m_{jk})|=|(d_{jk})|+\sum_{j,k}^{p} nD_{jk}\bar{x}_j\bar{x}_k\cdot|(d_{jk})|. \tag{41.81}$$

Тогда из (41.77) следует, что

$$\frac{|(d_{jk})|}{|(m_{jk})|}=\frac{1}{1+T^2/(n-1)}. \tag{41.82}$$

41.16 Обратимся к геометрической интерпретации полученного результата. При $p=1$ числитель и знаменатель (41.82) равны соответственно d_{11} и m_{11}, т. е. соответственно квадрату расстояния от выборочной точки P_1 (в n-мерном выборочном пространстве) до ее проекции на единичный вектор с равными направляющими углами, и квадрату расстояния от P_1 до начала координат O. Отношение равно квадрату синуса угла между OP_1 и «единичным вектором». Это и есть то геометрическое истолкование, к которому приводило нас распределение Стьюдента в примере 11.8 (т. 1).

В общем случае рассмотрим p совмещенных выборочных пространств, которые уже изучались при выводе распределения Уишарта в **41.7**.

Из соотношения, аналогичного (41.32), получаем, что $|(m_{jk})|$ равен квадрату объема многомерного параллелепипеда с одной вершиной в начале координат и со сторонами, параллельными $OP_1, \dots, OP_p$.

Далее, если H — гиперплоскость, перпендикулярная «единичному вектору», и O' — точка пересечения вектора с плоскостью, то, как легко видеть, проекции $P_1, \dots, P_p$ на H, скажем, $P_1', \dots, P_p'$ таковы, что d_{jk} как раз равны суммам квадратов или попарных произведений отклонений в H от O', как начала координат. Поэтому $|(d_{jk})|$ равен квадрату объема параллелепипеда в H, а отношение $|(d_{jk})|$ к $|(m_{jk})|$ равно квадрату косинуса угла между H и гиперплоскостью $OP_1 \dots P_p$. Обозначив этот угол θ, имеем

$$\frac{1}{1+T^2/(n-1)}=\cos^2\theta. \tag{41.83}$$

Далее, если выборочные точки P_j случайным образом распределены в n-мерном пространстве, гиперплоскость, которую они порождают, образует случайный угол с любым фиксированным вектором, в частности с тем единичным вектором, который рассматривался выше. Выборочное распределение θ представляет собой тогда распределение угла между фиксированным вектором и случайной гиперплоскостью. Но это то же самое, что и распределение между фиксированной гиперплоскостью и случайным вектором. Задачу о нахождении последнего распределения мы с несколько иной точки зрения решали в связи с множественным коэффициентом корреляции R. Мы видели (**27.26** т. 2), что R может рассматриваться как синус угла между «остаточной» величиной $x_{1\cdot 2\dots p}$ и пространством, содержащим остальные величины $x_2, \dots, x_p$. Если одно не зависит от другого, мы можем что-то одно зафиксировать. Таким образом, можно написать, что

$$\frac{1}{1+T^2/(n-1)}=1-R^2. \tag{41.84}$$

При этом мы должны помнить, что в распределении величины R^2, а именно в формуле

$$dF=\frac{1}{B\{(n-p)/2,\ (p-1)/2\}}(1-R^2)^{(n-p-2)/2}(R^2)^{(p-3)/2}dR^2 \tag{41.85}$$

p — полное число всех величин, а при составлении уравнения регрессии значения переменных центрируются их средними.

Применяя (41.85) к распределению Хотеллинга, мы должны увеличить p на единицу, так как к имеющимся величинам добавляется еще одна — «единичный вектор». Мы должны увеличить на единицу и n, так как варьирование не ограничивается отклонениями от среднего.

Осуществив эти изменения и проведя замену в (41.85) согласно (41.84), получаем распределение T^2:

$$dF = \frac{1}{B\{(n-p)/2,\ p/2\}} \cdot \frac{\{T^2/(n-1)\}^{(p-2)/2}}{\{1+T^2/(n-1)\}^{n/2}}\, d\left(\frac{T^2}{n-1}\right). \quad (41.86)$$

Иначе говоря,

$$\frac{(n-p)\,T^2}{p\,(n-1)} \text{ имеет распределение } F(p,\ n-p). \quad (41.87)$$

Чтобы изложенная выше теория оставалась строгой, мы, как и для распределения Уишарта в **41.7**, требуем, чтобы $n-1 \geqslant p$. Полученное распределение можно очевидным образом использовать для проверки того, что исследуемый вектор имеет среднее $\boldsymbol{\mu}_0$. Для этого надо измерять $\boldsymbol{x}$ относительно $\boldsymbol{\mu}_0$ как начала координат и воспользоваться распределением (41.87).

41.17 Этот же результат можно получить, используя прием замены из **41.12**. Заметим, что если в распределении Уишарта смешанные моменты относительно средних заменить теми же моментами относительно начала координат, скажем, c', распределение не меняется, за исключением того, что появляется дополнительная степень свободы. Тогда

$$dF = \frac{(n/2)^{pn/2}\,|\alpha|^{n/2}\,|c'|^{(n-p-1)/2}}{\pi^{p(p-1)/4}\Pi\Gamma\{(n+1-j)/2\}} \exp\left\{-\frac{n}{2}\sum \alpha_{jk}c'_{jk}/2\right\}\Pi\, dc'_{jk}. \quad (41.88)$$

Как и в (41.53), находим

$$\mathbf{M}(|c'|)^t = \frac{2^{pt}}{|\beta|^t}\prod_{j=1}^{p}\frac{\Gamma\{t+(n+1-j)/2\}}{\Gamma\{(n+1-j)/2\}}. \quad (41.89)$$

Теперь мы можем выписать также (41.88) в форме, подсказанной первоначальным видом распределения Уишарта:

$$dF = \frac{|\beta/2|^{(n-1)/2}\,|c|^{(n-p-2)/2}}{\pi^{p(p-1)/4}\Pi\Gamma\{(n-j)/2\}} \exp\left\{-\sum \beta_{jk}c_{jk}/2\right\}\Pi\, dc_{jk}\times$$
$$\times\frac{|\beta|^{1/2}}{(2\pi)^{p/2}}\exp\left\{-\sum \beta_{jk}\bar{x}_j\bar{x}_k/2\right\}\Pi\, dx. \quad (41.90)$$

Если мы умножим это на $|c'|^t$ и проинтегрируем, мы получим правую часть (41.89). Заменим n на $n+2u$ и разделим на

$|\beta/2|^u$. Разделив еще на соответствующую постоянную, получим

$$\mathbf{M}\{|c'|^t|c|^u\} =$$
$$= \frac{2^{p(t+u)}}{|\beta|^{t+u}}\prod_{j=1}^{p}\frac{\Gamma\{t+u+(n+1-j)/2\}\,\Gamma\{u+(n-j)/2\}}{\Gamma\{u+(n+1-j)/2\}\,\Gamma\{(n-j)/2\}}. \quad (41.91)$$

Напомним, что $|c|/|c'| = |d|/|m|$. Положим $t=-u$ в (41.91). Тогда

$$\mathbf{M}\{|d|/|m|\}^u = \prod_{j=1}^{p}\frac{\Gamma\{(n+1-j)/2\}\,\Gamma\{u+(n-j)/2\}}{\Gamma\{u+(n+1-j)/2\}\,\Gamma\{(n-j)/2\}} =$$
$$= \frac{\Gamma(n/2)\,\Gamma\{u+(n-p)/2\}}{\Gamma(u+n/2)\,\Gamma\{(n-p)/2\}} = \quad (41.92)$$
$$= \frac{B\{u+(n-p)/2,\ p/2\}}{B\{(n-p)/2,\ p/2\}}. \quad (41.93)$$

Это u-й момент распределения

$$dF = [B\{(n-p)/2,\ p/2\}]^{-1}\, x^{(n-p-2)/2}(1-x)^{(p-2)/2}dx, \quad (41.94)$$

которое однозначно определяется своими моментами. Следовательно, (41.94) — распределение отношения $|d|/|m|$ и переходя, согласно (41.82), к T, мы приходим к (41.86).

Существенной чертой статистики T^2 является то, что она представима в форме $\boldsymbol{z}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{z}$, где $\boldsymbol{z}$ — многомерный нормальный вектор с нулевыми математическими ожиданиями и матрицей рассеяния $\boldsymbol{V}$, а $\hat{\boldsymbol{V}}$ — оценка МП для $\boldsymbol{V}$, подправленная таким образом, чтобы было устранено смещение. Выше $\boldsymbol{z}$ равнялся $\bar{\boldsymbol{x}}$, а $\hat{\boldsymbol{V}} = \boldsymbol{d}/n(n-1)$. Аналогично мы можем определить статистику критерия для случая двух выборок (обобщая критерий t^2 Стьюдента для двух выборок), положив $\boldsymbol{z} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$, $\hat{\boldsymbol{V}} = \boldsymbol{P}(n_1^{-1}+n_2^{-1})$, где $\boldsymbol{P}$ — несмещенная оценка общей матрицы рассеяния двух совокупностей, вычисленная по объединенным выборкам. Условие, соответствующее условию $n-1 \geqslant p$ в **41.6**, здесь имеет вид $n_1+n_2-2 \geqslant p$. Противоположный случай см. у Дэмпстера (1958).

41.18 До сих пор мы обобщали на p-мерный случай результаты теории выборочных распределений для средних, характеристик рассеяния и отношений Стьюдента в нормальной схеме. Естественно выяснить, будет ли аналогично обобщаться критерий Фишера для отношения двух независимых дисперсий. Мы отложим обсуждение этого вопроса до следующей главы, где он естественно возникнет в задачах проверки гипотез. Отметим, однако, что точные распределения величин, соответствующих,

например, величине z Фишера, получить трудно, но моменты могут быть вычислены так же, как и в **41.12**. Представляет некоторый интерес также тот факт, что в обобщениях дисперсионного анализа в качестве статистики критерия появляется не отношение двух независимых определителей рассеяния, а отношения типа $|c_1|/|c_1+c_2|$, где c_1 и c_2 независимы.

В одномерном случае перейти от s^2/s_2^2 к $s_1^2/(s_1^2+s_2^2)$ просто, но это уже не представляется возможным при двух и более измерениях.

Результаты для больших выборок

41.19 Даже в нормальном случае, когда мы не имеем дело с семиинвариантами порядка большего двух, существует обескураживающее обилие параметров: p средних, p дисперсий, $(p-1)p/2$ ковариаций (или корреляций), т. е. всего $p(p+3)/2$ параметров. Даже при таком малом p, как 5, число параметров равно 20, а при $p=10$ их 65. Поэтому в общем случае задача о распределениях практически неприступна, хотя, ограничиваясь, например, случаем независимых величин, когда все корреляции равны нулю, мы можем избавиться от большого числа параметров.

В такой ситуации не следует пренебрегать результатами, касающимися больших выборок, хотя это часто и делают.

Учитывая лишь первый порядок по n, имеем

$$\mathbf{M}c_{jk}=\gamma_{jk}. \tag{41.95}$$

С точностью до того же порядка

$$\mathbf{M}(c_{jk}c_{lm})=n^{-2}\mathbf{M}\left\{\sum_{\alpha=1}^{n}x_{j\alpha}x_{k\alpha}\sum_{\beta=1}^{n}x_{l\beta}x_{m\beta}\right\}. \tag{41.96}$$

При $\alpha\neq\beta$ элементы в двух суммах справа независимы. Существует ровно $n(n-1)$ таких случаев, и для них математические ожидания равны $\gamma_{jk}\gamma_{lm}$. При $\alpha=\beta$ появляются n членов типа $M(x_{j\alpha}x_{k\alpha}x_{l\alpha}x_{m\alpha})$. В общем случае такое выражение содержит семиинварианты четвертого порядка, но при нормальном распределении, обращаясь к х. ф. для x

$$\varphi=\exp\left(-\sum_j\sum_k\gamma_{jk}t_jt_k/2\right),$$

находим, что

$$\mathbf{M}(x_{j\alpha}x_{k\alpha}x_{l\alpha}x_{m\alpha})=\gamma_{jk}\gamma_{lm}+\gamma_{jm}\gamma_{kl}+\gamma_{jl}\gamma_{km}. \tag{41.97}$$

Подстановка в (41.96) дает нам, что

$$\mathbf{M}(c_{jk}c_{lm})=n^{-1}(\gamma_{jm}\gamma_{kl}+\gamma_{jl}\gamma_{km})+\gamma_{jk}\gamma_{lm},$$

и, следовательно,

$$\operatorname{cov}(c_{jk},c_{lm})=n^{-1}(\gamma_{jm}\gamma_{kl}+\gamma_{jl}\gamma_{km}). \tag{41.98}$$

В частности, если $j=k=l=m=1$, мы получаем известный результат:

$$\mathbf{D}c_{11}=2\gamma_{11}^2/n=2\sigma_1^4/n. \tag{41.99}$$

Пример 41.6

В наших обозначениях безразлично, используем ли мы символы, соответствующие выборочным или теоретическим распределениям *),

$$r_{12}=\gamma_{12}/(\gamma_{11}\gamma_{22})^{1/2},$$

$$\frac{dr_{12}}{r_{12}}=\frac{d\gamma_{12}}{\gamma_{12}}-\frac{d\gamma_{11}}{2\gamma_{11}}-\frac{d\gamma_{22}}{2\gamma_{22}}. \tag{41.100}$$

Точно так же

$$\frac{dr_{13}}{r_{13}}=\frac{d\gamma_{13}}{\gamma_{13}}-\frac{d\gamma_{11}}{2\gamma_{11}}-\frac{d\gamma_{33}}{2\gamma_{33}}. \tag{41.101}$$

Перемножая (41.100) и (41.101) и беря математические ожидания, получаем

$$\frac{\operatorname{cov}(r_{12},r_{13})}{r_{12}r_{13}}=\frac{\operatorname{cov}(\gamma_{12},\gamma_{13})}{\gamma_{12}\gamma_{13}}-\frac{\operatorname{cov}(\gamma_{11},\gamma_{12})}{2\gamma_{11}\gamma_{12}}-\frac{\operatorname{cov}(\gamma_{11},\gamma_{13})}{2\gamma_{11}\gamma_{13}}-$$
$$-\frac{\operatorname{cov}(\gamma_{22},\gamma_{13})}{2\gamma_{22}\gamma_{13}}-\frac{\operatorname{cov}(\gamma_{33},\gamma_{12})}{2\gamma_{33}\gamma_{12}}+\frac{\mathbf{D}\gamma_{11}}{4\gamma_{11}^2}+$$
$$+\frac{\operatorname{cov}(\gamma_{11},\gamma_{33})}{4\gamma_{11}\gamma_{33}}+\frac{\operatorname{cov}(\gamma_{11},\gamma_{22})}{4\gamma_{11}\gamma_{22}}+\frac{\operatorname{cov}(\gamma_{22},\gamma_{33})}{4\gamma_{22}\gamma_{33}}. \tag{41.102}$$

Используя связи типа (41.98), приводим это к виду

$$n\operatorname{cov}(r_{12},r_{13})=r_{23}(1-r_{12}^2-r_{13}^2)+r_{12}r_{13}(r_{12}^2+r_{13}^2+r_{23}^2-1)/2. \tag{41.103}$$

Корреляцию r_{12} и r_{13} легко определить из тех соображений, что $n\mathbf{D}r=(1-r^2)^2$.

Собственные числа матрицы рассеяния

41.20 В последующих главах мы столкнемся с ситуациями, где нам понадобятся собственные числа случайных матриц. Если (c_{jk}) — матрица рассеяния, мы будем стремиться изучить поведение корней уравнения

$$|c_{jk}-\lambda\delta_{jk}|\equiv|c-\lambda I|=0, \tag{41.104}$$

*) Приведенные ниже выкладки имеют приближенный (асимптотически по n точный) смысл, так как автор весьма произвольно заменяет (там, где это ему удобно) теоретические коэффициенты корреляции (ϱ) выборочными (r) и наоборот. (*Прим. ред.*)

где δ_{jk} — символ Кронекера, равный нулю, если $j \neq k$, и единице, если $j = k$.

Мы воспользуемся тем фактом из теории матриц, что если $\boldsymbol{c}$ — положительно определенная матрица, все ее собственные числа действительны и неотрицательны. Лишь в исключительных случаях эти числа равны, и если q из них равны нулю, $\boldsymbol{c}$ вырождена и имеет ранг $p - q$.

41.21 По сути дела (41.104) является частным случаем уравнения

$$|c_{jk} - \lambda b_{jk}| = 0, \tag{41.105}$$

где b и c — независимые матрицы рассеяния, основанные на m и n наблюдениях соответственно.

Мы можем придать (41.105) другую, эквивалентную форму:

$$|u(c_{jk} + b_{jk}) - b_{jk}| = 0, \tag{41.106}$$

где

$$u = 1/(1 + \lambda), \quad \lambda = (1 - u)/u. \tag{41.107}$$

Сложность задачи о распределении связана с тем, что корни λ (или u) не являются алгебраическими функциями элементов матриц рассеяния. Легче получить выборочные распределения симметрических функций этих корней, чем распределения самих корней.

41.22 Предположим, что теоретическое распределение нормально с единичными дисперсиями и нулевыми коэффициентами корреляции. Из (41.58) находим, что плотность совместного распределения c_{jk} и b_{jk}

$$f \propto \frac{|c|^{(n-p-2)/2}\,|b|^{(m-p-2)/2}}{\prod_{j=1}^{p} \Gamma\{(n-j)/2\}\,\Gamma\{(m-j)/2\}} \exp\left\{-n \sum_{j=1}^{p} (c_{jj} + b_{jj})/2\right\}. \tag{41.108}$$

В главе 43 мы увидим, что каждому корню уравнения (41.106) соответствует собственный вектор $\boldsymbol{l}_j$ такой, что

$$\{\boldsymbol{b} - u_j(\boldsymbol{b} + \boldsymbol{c})\}\,\boldsymbol{l}_j = 0, \tag{41.109}$$

и мы можем выбрать $\boldsymbol{l}_j$ так, что

$$\boldsymbol{l}_j'(\boldsymbol{b} + \boldsymbol{c})\,\boldsymbol{l}_j = 1. \tag{41.110}$$

Если $\boldsymbol{l}$ — $(p \times p)$-матрица собственных векторов, то из (41.109) получаем

$$\boldsymbol{bl} = (\boldsymbol{b} + \boldsymbol{c})\,\boldsymbol{lu}, \tag{41.111}$$

где $\boldsymbol{u}$ — диагональная матрица с диагональными элементами $u_1, \ldots, u_p$. Предположим теперь, что u_j расположены в порядке убывания: $u_1 \geqslant \ldots \geqslant u_p$.

Из (41.109) получаем

$$\boldsymbol{l}_k'\boldsymbol{b}\boldsymbol{l}_j = u_j \boldsymbol{l}_k'(\boldsymbol{b} + \boldsymbol{c})\,\boldsymbol{l}_j \tag{41.112}$$

и, транспонируя ($\boldsymbol{b}$ и $\boldsymbol{c}$ симметричны),

$$\boldsymbol{l}_k'\boldsymbol{b} = \boldsymbol{l}_k' u_k (\boldsymbol{b} + \boldsymbol{c}).$$

Таким образом,

$$\boldsymbol{l}_k'\boldsymbol{b}\boldsymbol{l}_j = u_k \boldsymbol{l}_k'(\boldsymbol{b} + \boldsymbol{c})\,\boldsymbol{l}_j. \tag{41.113}$$

Из (41.112) и (41.113) следует, что если $u_j \neq u_k$, то

$$\boldsymbol{l}_k'(\boldsymbol{b} + \boldsymbol{c})\,\boldsymbol{l}_j = 0. \tag{41.114}$$

Умножая (41.111) слева на $\boldsymbol{l}'$ и используя (41.110), получаем

$$\boldsymbol{l}'\boldsymbol{b}\boldsymbol{l} = \boldsymbol{u}. \tag{41.115}$$

Из (41.110) и (41.114) следует, что

$$\boldsymbol{l}'(\boldsymbol{b} + \boldsymbol{c})\,\boldsymbol{l} = \boldsymbol{I}, \tag{41.116}$$

и, следовательно,

$$\boldsymbol{b} + \boldsymbol{c} = (\boldsymbol{l}')^{-1}\boldsymbol{l}^{-1}. \tag{41.117}$$

Точно так же из (41.115) получаем

$$\boldsymbol{b} = (\boldsymbol{l}')^{-1}\boldsymbol{u}\boldsymbol{l}^{-1}, \tag{41.118}$$

и, следовательно,

$$\boldsymbol{c} = (\boldsymbol{l}')^{-1}(\boldsymbol{I} - \boldsymbol{u})\,\boldsymbol{l}^{-1}. \tag{41.119}$$

41.23 Возвращаясь к (41.108), мы видим, что преобразования (41.118—19) приводят плотность вероятности к виду

$$f \propto \frac{\left\{\prod_{j=1}^{p}(1 - u_j)\right\}^{(n-p-2)/2} \left\{\prod_{j=1}^{p} u_j\right\}^{(m-p-2)/2}}{\prod \Gamma\{(n-j)/2\}\,\Gamma\{(m-j)/2\}} \times$$
$$\times \{\text{функция от } \boldsymbol{l}\}. \tag{41.120}$$

Рассмотрим якобиан преобразования.

Впоследствии выяснится, что распределение u не зависит от распределения l, и вклад последнего в общее распределение может быть «*факторизован*» (выделен в качестве сомножителя).

Из (41.118) и (41.119) получаем, что $\boldsymbol{b}$ и $\boldsymbol{c}$ в отдельности содержат $p(p+1)/2$ переменных, всего $p(p+1)$. С другой стороны, имеется p^2 переменных в $\boldsymbol{l}$ и p переменных u — всего снова $p(p+1)$. Якобиан для $\boldsymbol{b}$, $\boldsymbol{c}$ такой же, как и для $\boldsymbol{b}$, $\boldsymbol{b} + \boldsymbol{c}$.

Обозначив в (41.118) и (41.119) l^{-1} символом g, рассмотрим якобиан преобразования

$$b = g'ug, \tag{41.121}$$

$$b + c = g'g. \tag{41.122}$$

Мы докажем, что якобиан делится на величину $u_1 - u_2$. Хотя наши соображения носят общий характер, ради ясности приведем точные выкладки для $p = 2$.

В этом случае матрица b становится равной

$$\begin{pmatrix} g_{11}^2 u_1 + g_{21}^2 u_2 & g_{11}g_{12}u_1 + g_{21}g_{22}u_2 \\ g_{12}g_{11}u_1 + g_{22}g_{21}u_2 & g_{12}^2 u_1 + g_{22}^2 u_2 \end{pmatrix}, \tag{41.123}$$

а $b + c$ представляется так же, если u_j положить равными единице.

Для якобиана получаем

$$\frac{\partial\{b_{11}, b_{12}, b_{22}, (b+c)_{11}, (b+c)_{12}, (b+c)_{22}\}}{\partial\{u_1, u_2, g_{11}, g_{12}, g_{21}, g_{22}\}} =$$

$$= \begin{vmatrix} g_{11}^2 & g_{21}^2 & 2g_{11}u_1 & 0 & 2g_{21}u_2 & 0 \\ g_{11}g_{12} & g_{21}g_{22} & g_{12}u_1 & g_{11}u_1 & g_{22}u_2 & g_{21}u_2 \\ g_{12}^2 & g_{22}^2 & 0 & 2g_{12}u_1 & 0 & 2g_{22}u_2 \\ 0 & 0 & 2g_{11} & 0 & 2g_{21} & 0 \\ 0 & 0 & g_{12} & g_{11} & g_{22} & g_{21} \\ 0 & 0 & 0 & 2g_{12} & 0 & 2g_{22} \end{vmatrix}.$$

При $u_1 = u_2$ мы можем вычесть из трех верхних строк три нижних, умноженных на u, в верхних рядах уничтожатся все члены, кроме первых двух, и станет ясно, что определитель равен нулю.

Отсюда следует, что якобиан делится на каждое $u_j - u_k$, $j > k$. Произведение всех таких членов имеет степенной порядок $(p-1)p/2$. Якобиан не может содержать никакие другие члены, зависящие от u, поскольку он не может иметь более высокий степенной порядок.

Таким образом, только для u мы получаем из (41.120)

$$dF = K\frac{\{\Pi(1-u_j)\}^{(n-p-2)/2}\{\Pi u_j\}^{(m-p-2)/2}}{\Pi\Gamma\{(n-j)/2\}\,\Gamma\{(m-j)/2\}}\prod_{j<k}(u_j - u_k)\,\Pi du_j, \tag{41.124}$$

где K — некоторая константа. Определить ее непосредственным интегрированием трудно. Мы можем использовать следующий обходной прием. K определяют те элементы первоначальной плотности и якобиана, которые включают в себя p и $m+n$, а не только одно m. Запишем K как $K(p, m+n)$. Заметим, что $|b| = \Pi u_j$. Если мы увеличим в (41.124) m на $2t$ и проинтегрируем, мы получим t-й момент $|b|$ с точностью до множителя

$K(p, m+n+2t)$. Следуя **41.12**, можно найти этот момент, что приводит к выражению

$$\frac{K(p, m+n+2t)}{K(p, m+n)} = \prod_{j=1}^{p}\frac{\Gamma\{t + (m+n-1-j)/2\}}{\Gamma\{(m+n-1-j)/2\}}. \tag{41.125}$$

Отсюда следует, что

$$K(p, m+n) = K(p)\prod_{j=1}^{p}\Gamma\{(m+n-1-j)/2\}. \tag{41.126}$$

Функция $K(p)$ зависит только от p. Для того чтобы вычислить последнюю, заменим в (41.124) u_j на $2v_j/n$ и устремим n к бесконечности. Наше распределение принимает вид

$$dF = (\Pi v_j)^{(m-p-2)/2}\exp\left(-\sum v_j\right)\Pi(v_j - v_k)\,\Pi dv_j \times$$
$$\times\prod\frac{\Gamma\{(m+n-1-j)/2\}\,K(p)}{\Gamma\{(m-j)/2\}\,\Gamma\{(n-j)/2\}}. \tag{41.127}$$

Последнее может быть вычислено путем последовательной замены типа

$$v_1 = w_1, \qquad v_j = w_j + v_1, \qquad j > 1,$$

и выбора на каждой стадии m так, чтобы коэффициент при Πu_j равнялся нулю, что можно сделать, так как результат не зависит от m. Находим

$$\frac{K(p)}{2^{(p-1)p/2}}\prod\frac{\Gamma(p+1-j)}{\Gamma\{(p-j)/2\}} = 1. \tag{41.128}$$

Используя формулу удвоения (41.56) для гамма-функции, находим

$$K(p) = \frac{\pi^{p/2}}{\prod\Gamma\{(p+1-j)/2\}}.$$

Объединяя все полученное, окончательно приходим к распределению

$$dF = \pi^{p/2}\prod_{j=1}^{p}\frac{\Gamma\{(m+n-1-j)/2\}\,u_j^{(m-p-2)/2}(1-u_j)^{(n-p-2)/2}}{\Gamma\{(m-j)/2\}\,\Gamma\{(n-j)/2\}\,\Gamma\{(p+1-j)/2\}}\times$$
$$\times\prod_{j<k}(u_j - u_k)\,\Pi du_j. \tag{41.129}$$

Эта замечательная формула получена в 1939 г. Фишером, Сюем, С. Н. Роем и Гиршиком и независимо Мудом (см. Муд

(1951)). Распределение λ_j получаем с помощью простой замены $u = 1/(1+\lambda)$.

41.24 В случае (41.104), когда матрица $\boldsymbol{b}$ равна единичной матрице, получен несколько иной результат. Приведем распределение корней λ_j для указанного случая.

Читателю достаточно просмотреть предыдущие доказательства и внести там, где это необходимо, некоторые изменения, чтобы получить следующее:

$$dF = \frac{\pi^{p/2}}{2^{p(n-1)/2}} \prod_{j=1}^{p} \frac{\lambda_j^{(n-p-2)/2} \exp\left\{-\sum_{j=1}^{p} \lambda_j/2\right\}}{\Gamma\{(n-j)/2\}\,\Gamma\{(p+1-j)/2\}} \times$$

$$\times \prod_{j<k} (\lambda_j - \lambda_k)\, \Pi\, d\lambda_j, \qquad (41.130)$$

где в убывающем порядке расположены теперь величины λ_j.

Пример 41.7

За исключением ряда простых случаев, эти распределения весьма трудны в обращении. Рассмотрим случай $p = 2$. Из (41.130) получим

$$dF = \frac{\pi^{1/2} (\lambda_1\lambda_2)^{(n-4)/2} \exp\{-(\lambda_1+\lambda_2)/2\}}{2^{n-1}\Gamma\{(n-1)/2\}\,\Gamma\{(n-2)/2\}} (\lambda_1 - \lambda_2)\, d\lambda_1\, d\lambda_2. \quad (41.131)$$

С помощью формулы удвоения (41.56) для гамма-функций, это распределение сводится к виду

$$dF = \frac{(\lambda_1\lambda_2)^{(n-4)/2}}{4\Gamma(n-2)} \exp\{-(\lambda_1+\lambda_2)/2\}(\lambda_1 - \lambda_2)\, d\lambda_1\, d\lambda_2. \quad (41.132)$$

Взяв интеграл по λ_2 от 0 до λ_1, получим для λ_1 неполную гамма-функцию. С другой стороны, для функций

$$x = \lambda_1\lambda_2, \quad y = \lambda_1 + \lambda_2$$

$1/J = |\lambda_1 - \lambda_2|$ и распределение принимает вид

$$dF = \frac{x^{(n-4)/2} e^{-y/2}\, dx\, dy}{4\Gamma(n-2)}, \qquad 0 \leqslant x \leqslant y^2/4. \qquad (41.133)$$

Интегрируя по x, получаем

$$dF = \frac{(y/2)^{n-2} e^{-y/2} d(y/2)}{\Gamma(n-1)}, \qquad 0 \leqslant y \leqslant \infty. \qquad (41.134)$$

Действительно, определитель (41.104) в этом случае равен

$$\begin{vmatrix} s_1^2 - \lambda & s_1 s_2 r \\ s_1 s_2 r & s_2^2 - \lambda \end{vmatrix} = \lambda^2 - (s_1^2 + s_2^2)\lambda + s_1^2 s_2^2 (1 - r^2) = 0. \quad (41.135)$$

Сумма корней здесь равна сумме двух независимых случайных величин и имеет χ^2 распределение с $2n-2$ степенями свободы.

41.25 Свойства собственных чисел в случае больших выборок мы обсудим в главе 43.

Нецентральные распределения

41.26 Точно так же, как и в случае одномерных χ^2, t и F (отношение дисперсий), мы здесь сталкиваемся с необходимостью изучения нецентральных многомерных распределений, особенно при рассмотрении функций мощностей критериев, основанных на статистике T^2 или родственных ей.

Как и следовало ожидать, получающиеся в результате распределения оказываются очень громоздкими.

Отметим, в частности,

(а) нецентральное распределение Уишарта (см. Т. Андерсон (1946), последующие статьи и его книгу (1958));

(б) нецентральное T^2-распределение. Так как распределение T^2 имеет F-форму, фактически это нецентральное F-распределение (см. **42.22** ниже).

41.27 В заключение мы коснемся некоторых вопросов, которые не можем развить детально за недостатком места.

(1) Распределение (41.129) собственных чисел при $p = 1$ сводится к бетта-распределению. Поэтому Фостер и Рис (1957—8) назвали (41.129) «обобщенным бетта-распределением». Следуя методике, восходящей к С. Рою (1945) и Пиллаи (1956), они табулировали процентные точки наибольшего собственного числа при $p = 2, 3, 4$ и 5. Пиллаи (1960, 1967a), улучшив методику, составил таблицы (1964, 1965, 1967b) вплоть до $p = 7$. Сугияма (1967,a, b) получил точные формулы для распределения наибольшего собственного числа в случае (41.130) и для распределений наибольшего и наименьшего собственных чисел в случае (41.129).

(2) Уэйгл (1961), изучая распределения, пошел дальше и провел выборочные эксперименты на вычислительной машине. Задача отнюдь не простая, но при $p = 2, 3, 4$ для всех собственных чисел были получены вполне удовлетворительные результаты. Вычисления для больших p требуют лишь большего машинного времени.

(3) Индийская школа, начиная с работ Махаланобиса (1930) о сходстве рас, получила ряд интересных работ, основанных на

8

том, что сейчас называют D^2-статистикой. Смотрите, например, работы Бозе (1936), Бозе и С. Роя (1938) и многие другие, более поздние статьи С. Роя. D^2-статистика может быть определена формулой

$$D^2 = \sum_{j,k=1}^{p} a_{jk} (\bar{x}_{1j} - \bar{x}_{2j}) (\bar{x}_{1k} - \bar{x}_{2k}), \qquad (41.136)$$

где две выборки x_1 и x_2 взяты из двух p-мерных генеральных совокупностей, а (a_{jk}) — обратная матрица объединенной матрицы рассеяния. Соответствующий параметр

$$\Delta^2 = \sum \alpha_{jk} (\mu_{1j} - \mu_{2j}) (\mu_{1k} - \mu_{2k}) \qquad (41.137)$$

иногда называют обобщенным расстоянием Махаланобиса.

Фактически D^2 является простой функцией от статистики Хотеллинга T^2, определенной выше (см. (41.17)) для случая двух выборок, иначе говоря,

$$D^2 = (n_1^{-1} + n_2^{-1}) T^2.$$

(4) Если имеющие распределения Уишарта матрицы $\boldsymbol{c}_1$ и $\boldsymbol{c}_2$ основаны на p-мерных выборках объемами n_1 и n_2, то существует такая нижняя треугольная матрица $\boldsymbol{V}$, что $\boldsymbol{c}_1 + \boldsymbol{c}_2 = \boldsymbol{V}\boldsymbol{V}'$ (см. упражнение 41.16). Если $\boldsymbol{L}$ определяется из уравнения $\boldsymbol{c}_1 = \boldsymbol{V}\boldsymbol{L}\boldsymbol{V}'$, то $\boldsymbol{V}$ и $\boldsymbol{L}$ независимо распределены, и плотность вероятности $\boldsymbol{L}$ равна

$$f \propto |\boldsymbol{L}|^{(n_1-p-2)/2} |\boldsymbol{I} - \boldsymbol{L}|^{(n_2-p-2)/2}.$$

Этот результат впервые был получен Сюем (1939). Смотрите также работу Кширсагара (1961). Распределение $\boldsymbol{L}$ иногда называют многомерным бетта-распределением.

(5) Обзор работ по собственным числам можно найти у Джеймса (1964). Смотрите также работы Джеймса (1966) и Пиллаи (1966).

41.28 Чеймберс (1967) развил методику ассимптотических приближений для многомерных распределений, включая разложения Эджворта (см. 6.18 т. 1) и теорию возмущений. Эта методика обобщает методику 10.6.

УПРАЖНЕНИЯ

41.1 $\bar{\boldsymbol{x}}$ — среднее p-мерной выборки из нормальной совокупности с математическим ожиданием $\boldsymbol{\mu}$ и матрицей рассеяния $\boldsymbol{\gamma}$. Показать, что форма

$$n (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)' \boldsymbol{\gamma}^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)$$

имеет нецентральное χ^2-распределение с p степенями свободы и параметром нецентральности

$$n (\boldsymbol{\mu} - \boldsymbol{\mu}_0) \boldsymbol{\gamma}^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0),$$

где $\boldsymbol{\mu}_0$ — заданный вектор.

(Бозе, 1936.)

41.2 $\bar{\boldsymbol{x}}_1$ и $\bar{\boldsymbol{x}}_2$ — выборочные средние из двух p-мерных нормальных совокупностей со средними $\boldsymbol{\mu}_1$ и $\boldsymbol{\mu}_2$ и общей матрицей рассеяния $\boldsymbol{\gamma}$. Показать, что если $\boldsymbol{y} = \bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2$ и $\boldsymbol{\nu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, а n_1, n_2 — объемы выборок, то неравенство

$$(n_1 + n_2)^{-1} n_1 n_2 (\boldsymbol{y} - \boldsymbol{\nu})' \boldsymbol{\gamma}^{-1} (\boldsymbol{y} - \boldsymbol{\nu}) \geqslant \chi^2$$

дает доверительную область для $\boldsymbol{\nu}$.

41.3 Показать, что в 41.4 $\bar{\boldsymbol{x}}$ и выборочная матрица $\boldsymbol{a} (= \boldsymbol{c}^{-1})$ являются достаточными статистиками для $\boldsymbol{\mu}$ и $\boldsymbol{\gamma}^{-1}$.

41.4 Показать, что для четырехмерного нормального распределения корреляция между ковариациями c_{12} и c_{34} равна

$$(\rho_{13}\rho_{24} + \rho_{14}\rho_{23}) / \{(1 + \rho_{12}^2)(1 + \rho_{34}^2)\}^{1/2}.$$

(Уишарт, 1929.)

41.5 Среднее многомерной нормальной генеральной совокупности равно $\boldsymbol{\mu}$, все дисперсии равны σ^2, и все корреляции равны ρ. Пусть

$$(n-1) ps^2 = \sum_{i,j} (x_{ij} - \bar{x}_i)^2,$$

$$(n-1)(p-1) ps^2 r = \sum_{i \neq k} \sum_j (x_{ij} - \bar{x}_i)(x_{kj} - \bar{x}_k).$$

Показать, что величины

$$u = s^2 \{1 + (p-1) r\},$$
$$v = s^2 (1 - r)$$

имеют совместное распределение

$$dF \propto u^{(n-3)/2} v^{-1+(p-1)(n-1)/2} \exp \left\{ \frac{1}{2} (n-1) \left(\frac{u}{\alpha} + \frac{(p-1) v}{\theta} \right) \right\} du\, dv,$$

где

$$\alpha = \sigma^2 \{1 + (p-1) \rho\},$$
$$\theta = \sigma^2 (1 - \rho).$$

Вывести отсюда, что $u\theta/(v\alpha)$ имеет F-распределение с $n-1$, $(p-1)(n-1)$ ст. св., и получить доверительные интервалы для ρ.

(Гейссер, 1964.)

41.6 $\boldsymbol{c}$ имеет распределение Уишарта, соответствующее генеральной совокупности с матрицей рассеяния $\boldsymbol{\gamma}$. Показать, что $\boldsymbol{h}'\boldsymbol{c}\boldsymbol{h}$, где $\boldsymbol{h}$ — произвольная невырожденная $(p \times p)$-матрица, распределено также с соответствующим параметром $\boldsymbol{h}'\boldsymbol{\gamma}\boldsymbol{h}$.

41.7 *) Показать, что если $\boldsymbol{x}$ имеет распределение $N(\boldsymbol{0}, \boldsymbol{\gamma})$, иначе говоря, нормальное распределение с нулевым средним и матрицей рассеяния $\boldsymbol{\gamma}$, а $\boldsymbol{M}$ — ортогональная матрица, то $\boldsymbol{y} = \boldsymbol{M}\boldsymbol{x}$ имеет распределение $N(\boldsymbol{0}, \boldsymbol{M}\boldsymbol{\gamma}\boldsymbol{M}')$. Показать, в частности, что если $\boldsymbol{M}$ — матрица Хельмерта из 41.6, то $\boldsymbol{c}$ может быть представлена в виде

$$\sum_{l=1}^{n-1} y_l' y_l,$$

где y_l независимы и имеют распределение $N(\boldsymbol{0}, \boldsymbol{M}\boldsymbol{\gamma}\boldsymbol{M}')$.

*) Изложение этого упражнения несколько отличается от оригинала. (*Прим. ред.*)

Вывести отсюда аддитивное свойство матриц Уишарта, описанное в **41.10**.

41.8 Воспользовавшись примером 41.3, показать, что функция плотности величины $\{|c|/|\gamma|\}^{1/p}$ приближенно дается формулой

$$\alpha^{p(n-p)/2} y^{p(n-p)/2-1} e^{-\alpha y}/\Gamma\{p(n-p)/2\},$$

где

$$\alpha = np\{1-(p-1)(p-2)/2n\}^{1/p}.$$

(Хоул, 1937.)

41.9 Показать, что если c — выборочная матрица рассеяния, то при больших выборках $\{|c|/|\gamma|-1\}n^{1/2}$ имеет распределение с приблизительно нулевым векторным средним и дисперсией $2p$.

(Т. Андерсон, 1958.)

41.10 Пусть некая выборка принадлежит n-мерной нормальной совокупности, случайные переменные сгруппированы по k классам:

$$x_1, \ldots, x_{p_1};\quad x_{p_1+1}, \ldots, x_{p_2+p_1};\ \ldots;\ x_{p_{k-1}+\ldots+p_1+1}, \ldots, x_p.$$

Рассмотреть функцию

$$W = |r|/|r_{ij}^{(0)}|,$$

где $r_{ii}=1$, а $r_{ij}^{(0)}$ равно нулю, если переменные принадлежат различным классам и равны корреляциям r_{ij}, если переменные принадлежат одному и тому же классу.

Показать, что статистика критерия ОП независимости k классов равна

$$l = W^{n/2}$$

и

$$\mu_r'(W) = \prod_{t=1}^{k}\prod_{j=1}^{p_t}\left\{\frac{\Gamma\{(n-j)/2\}}{\Gamma\{r+(n-j)/2\}}\right\}\prod_{j=1}^{p}\frac{\Gamma\{r+(n-j)/2\}}{\Gamma\{(n-j)/2\}}$$

(Уилкс, 1935.)

41.11. Рассматривая частный случай последнего упражнения, показать, что если одна переменная x_1 не зависит от остальных $(x_2, \ldots, x_p)$, то

$$\mu_r'(W) = \frac{\Gamma\{(n-1)/2\}\,\Gamma\{r+(n-p)/2\}}{\Gamma\{r+(n-1)/2\}\,\Gamma\{(n-p)/2\}}.$$

Найти отсюда распределение множественного коэффициента корреляции при равенстве нулю теоретических коэффициентов.

(Уилкс, 1935.)

41.12. Показать алгебраически, что статистика Хотеллинга T^2 инвариантна относительно любых линейных преобразований p переменных.

41.13 Рассматривая пару нормальных величин с корреляцией ρ и определив v как

$$v = nc_{12}/\sigma_1\sigma_2(1-\rho^2),$$

показать, что функция плотности для v равна

$$f(v) = \frac{(1-\rho^2)^{(n-1)/2} e^{\rho v}}{\pi^{1/2} 2^{-1+n/2}\Gamma\{(n-1)/2\}}\left\{v^{-1+n/2} K_{-1+n/2}(v)\right\}$$

при $v>0$ и тому же выражению с заменой в фигурных скобках v на $-v$ при $v<0$ (здесь K — функция Бесселя второго рода с мнимым аргументом).

(Уишарт и Бартлетт, 1933.)

41.14 Используя (41.129), показать, что при $p=2$ распределение $y = (1-u_1)(1-u_2)$ дается формулой

$$dF = (1-\sqrt{y})^{n-p}(\sqrt{y})^{m-p-1}\, d\sqrt{y}/B(m-p, n-p+1).$$

41.15 Проверить (41.62).

41.16 (разложение Бартлетта). Пусть x_{jk}, $j=1, \ldots, p$; $k=1, \ldots, n$, — независимые $N(0,1)$-величины;

$$\begin{aligned} y_1 &= x_1,\\ y_2 &= x_2 - b_{21}' y_1,\\ &\cdots\cdots\cdots\\ y_p &= x_p - b_{p1}' y_1 - \ldots - b_{p,\,p-1}' y_{p-1}. \end{aligned}$$

Пусть все y_j ортогональны, т. е. $y_j' y_k = 0$, $j \neq k$. Тогда $b_{jk}' = y_k' x_j / y_k' y_k$. Пусть $b_{jk} = (y_k' y_k)^{1/2} b_{jk}'$. Показать, что

$$x'x = BB',$$

где B — треугольная матрица

$$\begin{pmatrix} (y_1'y_1)^{1/2} & 0 & 0 \ldots & 0\\ b_{21} & (y_2'y_2)^{1/2} & 0 \ldots & 0\\ \cdots & \cdots & \cdots & \cdots\\ b_{p1} & b_{p2} & \ldots & (y_p'y_p)^{1/2} \end{pmatrix}.$$

Показать, что b_{jk}, $k=1, \ldots, (p-1)$, — независимые $N(0,1)$-величины, и, следовательно, каждое $y_k' y_k$ является χ^2-величиной с $n-k+1$ степенями свободы, независимой от всех b и остальных $y_j' y_j$.

(Бартлетт, 1933; ср. Кширсагар, 1959.)

41.17 В предшествующем примере, взяв определители в равенстве $x'x = BB'$, показать, что $y_k' y_k$ равно отношению двух определителей «сумм произведений». Вывести отсюда, что k-й диагональный элемент матрицы, обратной матрице $x'x$, распределен так же, как величина, обратная χ^2 с $n-k+1$ степенями свободы.

(Вейсман, 1957; см. также Кширсагар, 1959.)

41.18 Используя предыдущий пример, доказать утверждение из **41.13** о том, что $|\alpha|/|\gamma|$ равно произведению p независимых χ^2-сомножителей со степенями свободы $n-1, n-2, \ldots, n-p$.

(Вейсман, 1957.)

41.19 Проверить результат (41.130).

41.20 Используя (41.53), показать, что $|c|$ — смещенная оценка $|\gamma|$.

Показать также, что смещение остается, если мы, для того чтобы получить ковариации, разделим «суммы произведений» не на n, а на $n-1$.

ГЛАВА 42

ПРОВЕРКА ГИПОТЕЗ В МНОГОМЕРНОМ АНАЛИЗЕ

42.1 Современный уровень знаний таков, что пока точному многомерному анализу, за редкими исключениями, поддаются лишь задачи, где рассматривается нормальный случай. В этом случае, как мы видели в предыдущей главе, оценки МП для средних и характеристик рассеяния совпадают с соответствующими выборочными статистиками. Общая же теория оценивания, опирающаяся не только на обычный метод МП, пока мало пригодна для практического использования. Методы, основанные на байесовских априорных вероятностях или соображения фидуциального характера обсуждались в литературе, но приводили порой к довольно аномальным результатам (см., например, Дэмпстер (1966)). Мы будем довольствоваться нормальными оценками МП, которые, во всяком случае, внушают определенное доверие.

Обсудим теперь вопрос о смещении. Рассмотрим, например, определитель рассеяния $|c|$ как оценку теоретического $|\gamma|$. Из равенства (41.53) при $t = 1$ находим

$$\mathbf{M}\,|c| = \prod_{j=1}^{p}(1 - j/n)\,|\gamma|. \tag{42.1}$$

С точностью до порядка n^{-1} мультипликативный сомножитель в правой части равен

$$1 - n^{-1}\sum_{j=1}^{p} j = 1 - p(p+1)/2n. \tag{42.2}$$

Смещение можно теперь оценить, а можно и просто убрать (или, по крайней мере, уменьшить до порядка n^{-2}) или, умножая $|c|$ на величину, обратную правой части (42.2), или используя прием, принадлежащий Кеную (**17.10**, т. 2). Если $|c|_n$ — определитель рассеяния, основанный на n наблюдениях, а $|c|_{n-1}$ — такой же определитель, основанный на $n-1$ наблюдениях, мы строим оценку *)

$$\mathrm{Est}\,|\gamma| = n\,|c|_n - (n-1)\,\mathrm{Av}\,|c|_{n-1}, \tag{42.3}$$

*) Символ Est X означает оценку X.

где среднее (символ Av) в правой части берется по всем n возможным определителям, получающимся после отбрасывания одного наблюдения. Отсюда, с точностью до порядка n^{-1}, получаем

$$\frac{\mathbf{M}\{\mathrm{Est}\,|\gamma|\}}{|\gamma|} = n\left\{1 - \frac{p(p+1)}{2n}\right\} - (n-1)\left\{1 - \frac{p(p+1)}{2(n-1)}\right\} = 1,$$

и оценка становится несмещенной с точностью до порядка n^{-1}. Наша идея совершенно прозрачна; трудности ее использования заключаются в большом объеме вычислений, в том числе и в вычислении определителей рассеяния, хотя для электронно-вычислительных машин последнее не непреодолимо.

42.2 В оставшейся части главы мы обсудим задачи проверки гипотез. Обзор, касающийся доверительных областей, толерантных областей и областей предсказания для многомерных нормальных распределений, дан Чу (1966).

Критерии однородности

42.3 Естественное обобщение обсуждавшегося в главе 35 дисперсионного анализа возникает, когда мы рассматриваем выборки из k различных p-мерных генеральных совокупностей и интересуемся, идентичны ли теоретические распределения. Обычно рассматривают три типа гипотез:

H: Генеральные совокупности идентичны, т. е. имеют одинаковые средние и ковариации.

H_1: Генеральные совокупности имеют одинаковые характеристики рассеяния, но отличаются средними.

H_2: Известно, что генеральные совокупности имеют одинаковые характеристики рассеяния. Гипотеза состоит в том, что одинаковы и средние.

Существуют, разумеется, и смешанные гипотезы. Например, некоторые характеристики рассеяния заданы — требуется определить, одинаковы ли остальные.

42.4 Для проверки простых гипотез лемма Неймана — Пирсона из **22.10** (т. 2) используется в многомерном анализе без изменений. Также и метод отношения правдоподобия из главы 24 может быть с той же уверенностью использован как статистический критерий для проверки сложных гипотез.

Следует отметить одно свойство процедуры максимизации. Если мы максимизируем, скажем, $f(\theta_1, \theta_2)$ по переменным θ_1 и θ_2, мы можем одновременно решать уравнения

$$\frac{\partial f}{\partial \theta_1} = 0, \qquad \frac{\partial f}{\partial \theta_2} = 0. \tag{42.4}$$

Это эквивалентно, однако, решению уравнения $\frac{\partial f}{\partial \theta_1}=0$ относительно θ_1, подстановке в f и затем решению уравнения $\frac{\partial f}{\partial \theta_2}=0$.

42.5 Рассмотрим теперь k многомерных нормальных генеральных совокупностей со средними μ_{jt} $(j=1, \ldots p,\ t=1, \ldots \ldots, k)$ и характеристиками рассеяния γ_{jlt} или, что то же, $\sigma_{jt}\sigma_{lt}\rho_{jlt}$.

Пусть объем выборки из t-й генеральной совокупности равен n_t. Если $\{\alpha_{jlt}\}$ есть обращение $\{\gamma_{jlt}\}$, общая функция правдоподобия всех выборок равна

$$\prod_{t=1}^{k}\frac{|\alpha_t|^{n_t/2}}{(2\pi)^{pn_t/2}}\exp\left\{-\frac{1}{2}\sum_{t=1}^{k}\underset{n_t}{S}\sum_{j,\,l=1}^{p}\alpha_{jlt}(x_{jt}-\mu_{jt})(x_{lt}-\mu_{lt})\right\}. \quad (42.5)$$

Если все μ и γ равны, соответствующая функция правдоподобия равна

$$|\alpha|^{n/2}/(2\pi)^{pn/2}\exp\left\{-\frac{1}{2}\underset{n}{S}\sum_{j,\,l=1}^{p}\alpha_{jl}(x_j-\mu_j)(x_l-\mu_l)\right\}, \quad (42.6)$$

где

$$n=\sum_{t=1}^{k}n_t. \quad (42.7)$$

Следуя обычной процедуре (глава 24), оценим параметры в (42.5) методом МП, подставим полученные оценки и найдем безусловный максимум L_1. Для того чтобы получить условный максимум L_0, точно так же поступим с (42.6). Отношение $l=L_0/L_1$ или некоторая монотонная функция от него и будет служить нам статистикой критерия.

Логарифм функции правдоподобия равен сумме слагаемых и, поскольку последние независимы, их можно максимизировать отдельно.

Как и следовало ожидать,

$$\hat{\mu}_{jt}=\bar{x}_{jt}, \quad (42.8)$$

$$\hat{\alpha}_{jlt}=a_{jlt}, \quad (42.9)$$

$$\hat{\gamma}_{jlt}=c_{jlt}. \quad (42.10)$$

Подстановка в экспоненциальный член в (42.5) дает постоянную, поскольку

$$\sum_{j,\,l}a_{jlt}c_{jlt}=1. \quad (42.11)$$

Отсюда с точностью до постоянной

$$L_1=\prod_{t=1}^{k}1/|c_{jlt}|^{n_t/2}. \quad (42.12)$$

Точно так же из (42.6) получаем

$$L_0=1/|c_{jl}|^{n/2}, \quad (42.13)$$

где c_{jl} — характеристика рассеяния объединенной выборки (полученной путем объединения всех k выборок в одну).

Тогда статистика критерия

$$l_H=L_0/L_1=\Pi|c_{jlt}|^{n_t/2}/|c_{jl}|^{n/2}=\prod_{t=1}^{k}\{|c_{jlt}|/|c_{jl}|\}^{n_t/2}. \quad (42.14)$$

Как и в одномерном случае, l_H меняется от 0 до 1. Чем ближе l_H к единице, тем более мы склоняемся к тому, чтобы принять гипотезу о равенстве всех средних и характеристик рассеяния.

42.6 Та же техника приводит нас к критериям для проверки H_1 и H_2. Мы приведем результаты без доказательства.

Пусть c_{jla} — средние характеристики рассеяния по всем генеральным совокупностям, а именно, пусть

$$c_{jla}=\frac{1}{n}\sum_{t=1}^{k}\sum_{u=1}^{n_t}(x_{ju}-\bar{x}_{jt})(x_{lu}-\bar{x}_{lt}). \quad (42.15)$$

Тогда для H_1

$$l_{H_1}=\prod_{t=1}^{k}\{|c_{jlt}|/|c_{jla}|\}^{n_t/2}. \quad (42.16)$$

а для H_2

$$l_{H_2}=\{|c_{jla}|/|c_{jl}|\}^{n/2}. \quad (42.17)$$

Заметим, что, так же как и в одномерном случае (см. упражнение 24.6),

$$l_H=l_{H_1}l_{H_2}. \quad (42.18)$$

Таким образом, наши статистики равны отношениям определителей рассеяния.

42.7 Чтобы использовать рассмотренные выше критерии, надо знать распределения их статистик. В некоторых случаях можно получить точные выражения для этих распределений, и во всех случаях, используя методику **41.12**, можно вычислить моменты. Однако для практических целей достаточно ограничиться приближениями Уилкса, использующими тот факт, что $(-2\log l)$ распределено как χ^2 с числом степеней свободы, равным числу связей, налагаемых гипотезами, т. е. числу оцениваемых пара-

метров в безусловной форме без числа оцениваемых параметров в условной. Доказательство сводится к простому обобщению сказанного в **24.7** (т. 2).

Мы оставили статистики в той форме, в какой они естественно возникли. Ясно, что и любая степень l хорошо служила бы нашим целям. В частности, возведенная в степень $2/n$ статистика для H_2 в (42.17) превращается в отношение определителей, а умноженный на $(-n)$ логарифм этого отношения распределен как χ^2.

42.8 Рассмотрим теперь моменты статистики ОП (42.14) для проверки H. Имеем

$$c_{jl}=\frac{1}{n}\sum_{n}(x_{jt}-\bar{x}_j)(x_{lt}-\bar{x}_l)=$$

$$=\frac{1}{n}\sum_{t=1}^{k}\sum_{n_t}(x_{jt}-\bar{x}_{jt})(x_{lt}-\bar{x}_{lt})+\frac{1}{n}\sum_{t=1}^{k}n_t(\bar{x}_{jt}-\bar{x}_j)(\bar{x}_{lt}-\bar{x}_l)=$$

$$=c_{jla}+c_{jlm}, \qquad (42.19)$$

где c_{jlm} — характеристики рассеяния выборочных средних относительно средних по объединенной выборке. Следуя приему, использованному в **41.17**, мы можем выписать функцию правдоподобия характеристик рассеяния в двух видах: первый будет содержать $|c|$, второй — $|c_{jla}|$ и $|c_{jlm}|$. Тогда

$$\mathbf{M}(l_H^r)=\prod_{t=1}^{k}\left(\frac{n}{n_t}\right)^{prn_t/2}\prod_{j=1}^{p}\frac{\Gamma\{[n_t(1+r)-j]/2\}}{\Gamma\{(n_t-j)/2\}}\cdot\frac{\Gamma\{(n-j)/2\}}{\Gamma\{[n(1+r)-j]/2\}}. \qquad (42.20)$$

Этот и следующие результаты принадлежат Уилксу (1932), в чьей статье можно ознакомиться со всеми подробностями.

Аналогично показывается, что

$$\mathbf{M}(l_{H_1}^r)=$$

$$=\prod_{t=1}^{k}\left[\left(\frac{n}{n_t}\right)^{prn_t/2}\prod_{j=1}^{p}\frac{\Gamma\{[n_t(1+r)-j]/2\}}{\Gamma\{(n_t-j)/2\}}\prod_{j=1}^{p}\frac{\Gamma\{(n-k+1-j)/2\}}{\Gamma\{[n(1+r)-k+1-j]/2\}}\right., \qquad (42.21)$$

$$\mathbf{M}(l_{H_2}^r)=\prod_{j=1}^{p}\frac{\Gamma\{[n(1+r)-k+1-j]/2\}\,\Gamma\{(n-j)/2\}}{\Gamma\{(n-k+1-j)/2\}\,\Gamma\{[n(1+r)-j]/2\}}. \qquad (42.22)$$

Заметим, что (так же как и в упражнении 24.6) моменты l_H равны произведению моментов остальных двух величин l. Это означает, что l_{H_1} и l_{H_2} независимы, если справедлива H. Последнее вполне естественно, поскольку средние и характеристики рассеяния независимы.

42.9 Укажем попутно на один из возможных источников заблуждений. В наших обозначениях n является объемом выборки, а не числом степеней свободы. Та форма функции плотности, что приведена, например, в (42.5), содержит символ n только в стоящих впереди постоянных. Если бы степенной показатель сводился к квадратичной форме, которая преобразуется в сумму $p-q$ квадратов, в соответствующих постоянных стояло бы не p, а $p-q$. А если бы сумма выборочных величин соответствовала не n, а $n-q$ величинам, мы должны были бы в постоянных заменить n на $n-q$. Показатель степени в (42.5) и (42.6) зависит от того, как определены характеристики рассеяния. В наших выкладках делитель всегда равен n. Некоторые авторы в определении характеристик рассеяния используют $\nu_t=n_t-1$ вместо нашего n_t и $\nu=n-k$ вместо n. Поступают так по той причине, что, как отмечено в примере 24.6 т. 2, критерий в этом случае становится менее смещенным.

42.10 При $p=1$ наши распределения, конечно, сводятся к распределениям, уже знакомым нам по одномерной теории. Читатель может убедиться в этом, проделав все необходимое в качестве упражнения. При $p=2$ находим из (42.22) для моментов l_{H_2}

$$\mu_r'=\frac{\Gamma\{(n(1+r)-k)/2\}\,\Gamma\{(n-1)/2\}\,\Gamma\}\,\{[n(1+r)-k-1]/2\}\,\Gamma\{(n-2)/2\}}{\Gamma\{(n-k)/2\}\,\Gamma\{[n(1+r)]/2-1\}\,\Gamma\{(n-k-1)/2\}\,\Gamma\{[n(1+r)/2]-2\}}, \qquad (42.23)$$

Формула удвоения для гамма-функции (41.56) позволяет свести последнее к виду

$$\mu_r'=\frac{\Gamma\{n(1+r)-k-1\}\,\Gamma(n-2)}{\Gamma(n-k-1)\,\Gamma\{n(1+r)-2\}}=$$

$$=B\{n(1+r)-k-1,\ k-1\}/B\{k-1,\ n-k-1\}. \qquad (42.24)$$

Моменты $(l_{H_2})^{1/n}$, иначе $\{|c_{jla}|/|c_{jl}|\}^{1/2}$ те же, что и у распределения

$$dF=(1-x)^{k-2}x^{n-k-2}\,dx/B(k-1,\ n-k-1). \qquad (42.25)$$

Если p четно, то, используя формулу удвоения для гамма-функции, можно моменты l-статистик выразить в виде произведений бетта-функций; при этом сами статистики оказываются произведениями некоторых независимых величин типа I. Это позволяет получить законченную форму для функций распределения, когда или p, или число связей, налагаемых гипотезой, четно (см. Шатцов (1966b), где приведены поправочные таблицы для асимптотических χ^2-аппроксимаций).

42.11 На практике наиболее полезными оказываются результаты, связанные с асимптотическими разложениями. Следуя Боксу (1949), мы разовьем общий метод, опирающийся на идею примера 41.4 предыдущей главы. По существу, этот метод пригоден для широкого класса статистик, зависящих от отношения правдоподобия. Пусть величина W такова, что

$$\mathsf{M}(W^t) = \text{const} \cdot \left[\frac{\prod_{j=1}^{k} y_j^{y_j}}{\prod_{j=1}^{m} x_j^{x_j}} \right]^t \frac{\sum_{j=1}^{m} \Gamma\{x_j(1+t)+\xi_j\}}{\prod_{j=1}^{k} \Gamma\{y_j(1+t)+\eta_j\}}, \quad (42.26)$$

где

$$\sum_{j=1}^{m} x_j = \sum_{j=1}^{k} y_j. \quad (42.27)$$

У нас x_j и y_j будут велики, порядка общего объема выборки n, и мы будем писать $O(n)$ как для $O(x)$, так и для $O(y)$.

Пусть

$$N = -2 \log W. \quad (42.28)$$

Найдем характеристическую функцию величины ρN, где ρ — лежащая между 0 и 1 постоянная масштаба, которую мы выберем позднее из соображений удобства. Пусть t — аргумент х. ф. Тогда х. ф. величины ρN равна

$$\varphi(t) = \mathsf{M}(\exp it\rho N) = \mathsf{M}(W^{-2\rho it}) =$$

$$= \text{const} \cdot \left[\frac{\prod_{j=1}^{k} \left(y_j^{y_j}\right)}{\prod_{j=1}^{m} \left(x_j^{x_j}\right)} \right]^{-2\rho it} \frac{\prod_{j=1}^{m} \Gamma\{x_j(1-2\rho it)+\xi_j\}}{\prod_{j=1}^{k} \Gamma\{y_j(1-2\rho it)+\eta_j\}}. \quad (42.29)$$

Положив теперь

$$(1-\rho)x_j = \beta_j, \qquad (1-\rho)y_j = \varepsilon_j, \quad (42.30)$$

получим производящую функцию семиинвариантов для ρN в виде

$$\psi(t) = g(t) - g(0),$$

где

$$g(t) = 2i\rho t \left[\sum_{j=1}^{m} x_j \log x_j - \sum_{j=1}^{k} y_j \log y_j \right] +$$

$$+ \sum_{j=1}^{m} \log \Gamma\{\rho x_j(1-2it)+\beta_j+\xi_j\} -$$

$$- \sum_{j=1}^{k} \log \Gamma\{\rho y_j(1-2it)+\varepsilon_j+\eta_j\}. \quad (42.31)$$

Используем теперь разложение гамма-функции (справедливое для комплексного аргумента)

$$\log \Gamma(x+h) = \log \sqrt{2\pi} + (x+h-1/2)\log x - x -$$

$$- \sum_{j=1}^{n} (-1)^j \frac{B_{j+1}(h)}{j(j+1)x^j} + R_{n+1}(x), \quad (42.32)$$

где величины B — полиномы Бернулли первого порядка (3.25, т. 1). Учитывая (42.31) и соответствующее разложение для $g(0)$, получаем

$$\psi(t) = -\frac{1}{2} f \log(1-2it) + \sum_{j=1}^{\infty} \omega_j \{(1-2it)^{-j} - 1\}, \quad (42.33)$$

где

$$f = -2\left\{ \sum_{j=1}^{m} \xi_j - \sum_{j=1}^{k} \eta_j - \frac{1}{2}(m-k) \right\}, \quad (42.34)$$

$$\omega_j = \frac{(-1)^{j+1}}{j(j+1)} \left\{ \sum_{i=1}^{m} \frac{B_{j+1}(\beta_i+\xi_i)}{(\rho x_i)^j} - \sum_{i=1}^{k} \frac{B_{j+1}(\varepsilon_i+\eta_i)}{(\rho y_i)^j} \right\}. \quad (42.35)$$

Мы должны помнить, что, как видно из (42.30), β и ε имеют порядок n, пока $1-\rho$ не мало. При $\rho = 1$ имеет $\omega_1 = O(n^{-1})$.

Таким образом,

$$\psi(t) = -\frac{1}{2} f \log(1-2it) + O(n^{-1}), \quad (42.36)$$

и, следовательно, с точностью до этого порядка $-2\log W$ распределена как χ^2 с f степенями свободы.

Делая в аппроксимации еще один шаг, находим, что, поскольку

$$B_2(x) = x^2 - x - 1/6,$$

$$\omega_1 = \frac{1}{2\rho} \left[\sum_{i=1}^{m} \left\{ \frac{(\beta_i+\xi_i)^2 - (\beta_i+\xi_i) + 1/6}{x_i} \right\} - \right.$$

$$\left. - \sum_{i=1}^{k} \left\{ \frac{(\varepsilon_i+\eta_i)^2 - (\varepsilon_i+\eta_i) + 1/6}{y_i} \right\} \right].$$

Последнее с помощью (42.30) сводится к виду

$$\omega_1 = \frac{1}{2\rho} \left[-(1-\rho)f + \sum_{i=1}^{m} \frac{\xi_i^2 - \xi_i + 1/6}{x_i} - \sum_{i=1}^{k} \frac{\eta_i^2 - \eta_i + 1/6}{y_i} \right]. \quad (42.37)$$

Выбирая теперь ρ так, чтобы $\omega_1 = 0$, получим

$$\psi(t) = -\left(\frac{1}{2}\right) f \log(1 - 2it) + O(n^{-2}). \qquad (42.38)$$

Итак, в общем случае существует постоянная ρ такая, что с точностью до порядка n^{-2} $-2\rho \log W$ распределено как χ^2 с f степенями свободы.

Бокс продвинулся в своих исследованиях значительно дальше, но, как мы уже отмечали, грубая аппроксимация (42.36) обычно достаточно хороша для практического использования. (См. также Лоли (1956b), чья работа кратко излагалась в **24.9** т. 2.)

Пример 42.1

Найдем χ^2 — аппроксимацию для распределения l_H, чьи моменты даны в (42.20). Сравнение с (42.26) показывает, что они имеют требуемую форму

$$k = p, \quad y_j = n/2, \quad \eta_j = -j/2; \quad j = 1, 2, \ldots, p;$$
$$m = pk, \quad x_j = n_t/2, \quad \xi_j = -j/2; \quad t = 1, 2, \ldots, k.$$

В качестве первого приближения полагаем, что $-2\log l$ распределено как χ^2 с числом степеней свободы, данным в (42.34), иначе говоря,

$$\begin{aligned} f = -2\{(-1/2)k(1/2)p(p+1) - \\ -(-1/2)(1/2)p(p+1) - (1/2)(pk-p)\} = \\ = (k-1)p(p+3)/2. \end{aligned} \qquad (42.39)$$

Фактически это число параметров в функции правдоподобия (42.5) без числа параметров в (42.6), т. е. число связей, налагаемых гипотезой.

Переходя ко второму приближению, из (42.37) находим, что

$$\begin{aligned} 0 = -(1-\rho)(1/2)(k-1)p(p+3) + \\ + p\left(\sum (1/n_j) - (1/n)\right)(2p^2 + 9p + 12)/12, \\ \rho = 1 - \left(\sum_{j=1}^{k} \frac{1}{n_j} - \frac{1}{n}\right) \frac{2p^2 + 9p + 12}{6(k-1)(p+3)}. \end{aligned} \qquad (42.40)$$

Анализируя выражения такого рода, нужно помнить, что в наших предположениях n_j и n — объемы выборок. Как отмечалось выше, иногда в литературе приводятся результаты для статистик, основанные на числах степеней свободы: $\nu_j = n_j - 1$ и

$\nu = n - k$. Последнее не влияет на (42.39), но меняет второй член в (42.40). В этом случае ξ_j равно $(1-j)/2$, $\eta_j = (k-j)/2$, а выражение, соответствующее (42.40), имеет вид

$$\rho = 1 - \left(\sum_{j=1}^{k} \frac{1}{\nu_j} - \frac{1}{\nu}\right) \frac{2p^2 + 3p - 1}{6(k-1)(p+3)} + \frac{1}{\nu}\left(\frac{p-k+2}{p+3}\right). \qquad (42.41)$$

Критерии независимости

42.12 За несколькими исключениями, построение всех рассматриваемых здесь критериев основано на использовавшихся ранее идеях: обращение к некоторой статистике правдоподобия, вычисление ее моментов, аппроксимация критерием χ^2 или каким-нибудь другим, более точным. Поэтому нет необходимости тратить время, отвлекаясь на детали, проверка которых может быть предоставлена читателю.

Прежде всего рассмотрим критерий независимости. Как обычно, нам дана выборка объема n из p-мерной нормальной совокупности и задано некоторое разбиение компонент (или координат) наблюденного вектора на q групп, содержащих $p_1, \ldots, p_q$ элементов. Требуется проверить гипотезу о независимости каждой группы от всех остальных. Мы будем особенно интересоваться случаями $q = 2$ и $q = p$.

Если теоретическая матрица рассеяния γ разбита на q^2 частей:

$$\begin{pmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1q} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2q} \\ \cdot & \cdot & \cdot & \cdot \\ \gamma_{q1} & \gamma_{q2} & \cdots & \gamma_{qq} \end{pmatrix}, \qquad (42.42)$$

то проверяемая гипотеза состоит в том, что

$$\gamma_{jk} = 0 \qquad j \neq k. \qquad (42.43)$$

Полагая альтернативой к (42.43) общий вид (42.42), находим статистику ОП

$$l_H = \frac{|c|^{n/2}}{\prod_{j=1}^{q} |c_{jj}|^{n/2}}, \qquad (42.44)$$

где, как обычно, (c_{jk}) — выборочная матрица рассеяния. Последнее перепишем в эквивалентной форме:

$$l_H^{2/n} = \frac{|c|}{\prod_{j=1}^{q} |c_{jj}|}. \qquad (42.45)$$

Если гипотеза справедлива, статистика l не зависит от ее знаменателя, а $|c_{jj}|$ не зависит от $|c_{kk}|$. Тогда из (41.53) получаем

$$\mathsf{M}(l^r) = \frac{\prod_{j=1}^{p} \Gamma\left\{\frac{1}{2}(n-j)+\frac{1}{2}rn\right\} \prod_{k=1}^{q} \prod_{j=1}^{p_k} \Gamma\left\{\frac{1}{2}(n-j)\right\}}{\prod_{j=1}^{p} \Gamma\left\{\frac{1}{2}(n-j)\right\} \prod_{k=1}^{q} \prod_{j=1}^{p_k} \Gamma\left\{\frac{1}{2}(n-j)+\frac{1}{2}rn\right\}}. \qquad (42.46)$$

Отсюда $-2\log l$ распределено приблизительно как χ^2 с

$$f = \frac{1}{2}\left\{(p(p+1) - \sum_{j=1}^{q} p_j(p_j+1)\right\} \qquad (42.47)$$

ст. св. При более точной аппроксимации имеем

$$\rho = 1 - \frac{2\left(p^3 - \sum p_j^3\right) + 9\left(p^2 - \sum p_j^2\right)}{6n\left(p^2 - \sum p_j^2\right)}. \qquad (42.48)$$

В случае $p_j = 1$ для всех j, т. е. в случае, когда мы проверяем независимость всех p компонент,

$$f = p(p-1)/2 \qquad (42.49)$$

$$\rho = 1 - \{(2p+11)/6n\}. \qquad (42.50)$$

Статистика в этом случае равна возведенному в степень $n/2$ отношению $|c|$ к произведению диагональных элементов, т. е. дисперсий, или, что эквивалентно возведенному в степень $n/2$ корреляционному определителю.

Обобщая работу Уилкса (1935), Консул (1967a) получил точное распределение величины (42.45) для $q = 2$, $p_1 \leqslant 6$ и всех p_2, а также для многих случаев при $q = 3$. См. также Т. Андерсон (1958). Дейли (1940) и Нарайн (1950) показали, что рассмотренный критерий ОП для проверки независимости является несмещенным.

Критерий сферичности

42.13 Теперь займемся проверкой гипотезы о том, что выборочная матрица рассеяния соответствует генеральной совокупности, теоретическая матрица рассеяния которой пропорциональна некоторой данной матрице $\boldsymbol{\gamma}$. Поскольку $\boldsymbol{\gamma}$ известна, линейным преобразованием мы можем свести ее к единичной матрице. Обозначив символом $\boldsymbol{c}$ соответствующим образом преобразованную выборочную матрицу, мы должны теперь проверить, действительно ли матрице $\boldsymbol{c}$ соответствует теоретическая матрица $\sigma^2\boldsymbol{I}$, где σ^2 неизвестно. По понятным причинам последнее называют проверкой сферичности (Мочли, 1940). Рассмотрим статистику критерия

$$l = \left[\frac{|c|}{\{(\operatorname{tr} c)/p\}^p}\right]^{n/2}. \qquad (42.51)$$

Моменты величины $l^{2/n}$ даются формулой

$$\mathsf{M}(l^{2r/n}) = p^{rp}\frac{\Gamma\{p(n-1)/2\}}{\Gamma\{pr+p(n-1)/2\}} \cdot \prod_{j=1}^{p} \frac{\Gamma\{r+(n-j)/2\}}{\Gamma\{(n-j)/2\}}, \qquad (42.52)$$

и, как обычно, $-n\log l^{2/n}$ распределено как χ^2 с

$$f = \frac{1}{2}p(p+1) - 1. \qquad (42.53)$$

Используя вторую аппроксимацию, имеем

$$\rho = 1 - \frac{2p^2+p+2}{6p(n-1)}. \qquad (42.54)$$

Консул (1967b) поолучил точное распределение (42.51) для $p = 2$, 3, 4 и 6. Глезер показал, что критерий сферичности не смещен.

Аналогично можно проверить, что $\boldsymbol{\gamma}$ равно некоторой заданной матрице $\boldsymbol{\gamma}_0$. Распределение статистики ОП для этого случая рассматривал Корин (1968).

42.14 Критерии однородности могут быть использованы для обобщения на p-мерный случай одномерных критериев ДА. Роль дисперсий теперь будут играть обобщенные дисперсии, т. е. определители рассеяния.

Пример 42.2

Следуя Пирсону и Уилксу (1963), рассмотрим один случай двумерного ДА ($p = 2$). В нашем распоряжении находятся пять выборок, в каждой по 12 членов ($k = 5$, $n_t = 12$ для всех t, $n = 60$), являющиеся результатом двух типов испытаний над алюминиевыми отливками.

По каждому из 60 образцов были измерены предел прочности при растяжении (в 1000 фунт. на 1 кв. дюйм), который мы обозначим x, и твердость (в единицах Роквелла), которую

обозначим y. Данные могут быть суммированы следующим образом:

Номер выборки t	x среднее	x (в. д.)$^{1/2}$	y среднее	y (в. д.)$^{1/2}$	Корреляция
1	33,399	2,565	68,49	10,19	0,683
2	28,216	4,318	68,02	14,49	0,876
3	30,313	2,188	66,57	10,17	0,714
4	33,150	3,964	76,12	11,18	0,715
5	34,269	2,715	69,92	9,88	0,805

Проверим прежде всего гипотезу H_1 о равенстве характеристик рассеяния *).

Получаем следующие результаты:

t	Суммы квадратов отклонений от среднего x	y	Сумма произведений отклонений от средних	Обобщенная дисперсия (определ. рассеяния)	$\log_{10}$ обобщенной дисперсии
1	78,948	1247,18	214,18	365,204	2,56254
2	223,695	2519,31	657,62	910,401	2,95923
3	57,448	1241,78	190,63	243,029	2,38566
4	187,618	1473,44	375,91	938,451	2,92741
5	88,456	1171,73	259,18	253,281	2,40360
Всего	636,165	7653,44	1697,52		13,28844

Для средних характеристик рассеяния (относительно соответствующих средних) по всем генеральным совокупностям имеем

$$c_{11a} = 636{,}165/60 = 10{,}6628,$$
$$c_{22a} = 127{,}5573,$$
$$c_{12a} = 28{,}2920,$$
$$|c_{jka}| = 552{,}018.$$

Согласно (42.16) статистика имеет вид $l^{2/n}$. Окончательно

$$\frac{2}{n}\log l = \frac{1}{5}\sum_{t=1}^{5}\{\log|c_{jlt}| - |c_{jla}|\} = \bar{1},91473,$$

*) В оригинале «that dispersions are homogeneous». (*Прим. перев.*)

где логарифм берется по основанию 10. Отсюда $l^{2/n} = 0{,}8217$ и $-n\log l^{2/n} = 11{,}78$. Число степеней свободы равно $3(k-1)$, т. е. 12. Наблюденное значение статистики критерия не противоречит предположению об однородности.

Предполагая теперь равенство характеристик рассеяния, перейдем к проверке гипотезы о равенстве средних (H_2). Для проверки этой гипотезы следует обратиться к (42.17). Вычислим дисперсию объединенной выборки. (В качестве среднего берем теперь среднее объединенной выборки.) Данные таковы:

Источник	Ст. св.	Сумма квадратов x	Сумма квадратов y	Сумма произведений xy
Межвыборочные	$k-1=4$	306,089	662,77	214,86
Внутривыборочные	$n-k=55$	636,165	7653,44	1697,52
Всего	$n-1=59$	942,254	8316,21	1912,38

Определитель рассеяния объединенной выборки равен тогда 1160,77, и значение статистики равно

$$-60\log(552{,}018/1160{,}77) = 44{,}59.$$

Число степеней свободы равно $2(k-1) = 8$. Этот результат отвергает H_2, так как иначе был бы слишком мал размер критерия.

Мы приходим к выводу об отсутствии однородности среди средних. Проверим теперь x и y отдельно.

	Оценки дисперсий x	y	Ст. св.
Межвыборочные	76,522	165,69	4
Внутривыборочные	11,566	139,15	55

Обычный F-критерий с однопроцентным уровнем значимости показывает, что неоднородность связана с различиями в пределах прочности на растяжение, а не с различиями в твердости.

Многомерная регрессия

42.15 Предположим теперь, что наши величины x линейно связаны с некоторыми фиксированными величинами z соотношением

$$\underset{p\times n}{\boldsymbol{x}} = \underset{p\times q}{\boldsymbol{\beta}}\ \underset{q\times n}{\boldsymbol{z}} + \underset{p\times n}{\boldsymbol{\varepsilon}}. \tag{42.55}$$

Здесь $\boldsymbol{\beta}$ — $(p\times q)$-матрица коэффициентов, а $\boldsymbol{\varepsilon}$ — $(p\times n)$-матрица ошибок. Если бы векторы-строки $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_p$ были независимы, мы бы имели дело с p независимыми регрессиями, по одной для каждого вектора $\boldsymbol{x}$. Мы такой независимости предполагать не будем. Приписывая переменной z_1 единичное значение, мы можем рассматривать ее как «фиктивную» переменную. Тогда величины $\beta_{j1}z_1$ можно рассматривать как средние и, следовательно, считать, что средние элементов ε равны нулю (см. пример 19.1). Наша цель — оценить $\boldsymbol{\beta}$ и матрицу рассеяния величин $\boldsymbol{\varepsilon}$, которую обозначим $\boldsymbol{\sigma}$. Пусть $\boldsymbol{\alpha} = \boldsymbol{\sigma}^{-1}$.

В других обозначениях (42.55) представляет обобщенную на p-мерный случай общую линейную модель, изложенную в главах 19, 24 и далее. Здесь, однако, мы предполагаем нормальность с самого начала.

42.16 Как и в одномерном случае, оценим $\boldsymbol{\beta}$, максимизируя функцию правдоподобия

$$L = \frac{|\boldsymbol{\alpha}|^{n/2}}{(2\pi)^{np/2}} \exp\left[-\frac{1}{2}\left\{\sum_{t=1}^{n} (\boldsymbol{x}_t - \boldsymbol{\beta}\boldsymbol{z}_t)' \boldsymbol{\alpha} (\boldsymbol{x}_t - \boldsymbol{\beta}\boldsymbol{z}_t)\right\}\right], \quad (42.56)$$

где индекс t относится к выборке.

Чтобы получить оценки МП для β_{jk}, выпишем уравнения

$$\frac{\partial \log L}{\partial \beta_{jk}} = -\sum_{t=1}^{n}\sum_{s=1}^{p}\left\{z_{kt}\alpha_{js}\left(x_{st} - \sum_{m=1}^{q}\hat{\beta}_{sm}z_{mt}\right)\right\} = 0. \quad (42.57)$$

Положим

$$u_{sk} = \sum_t z_{kt}x_{st} = \boldsymbol{x}_s\boldsymbol{z}'_k, \quad (42.58)$$

$$v_{sk} = \sum_t z_{kt}z_{st} = \boldsymbol{z}_s\boldsymbol{z}'_k. \quad (42.59)$$

Из (42.57) получаем

$$\sum_s \alpha_{js}\left(u_{sk} - \sum_{m=1}^{q}\hat{\beta}_{sm}v_{mk}\right) = 0,$$

и, поскольку $\boldsymbol{\alpha}$ не вырождена,

$$u_{sk} - \sum_{m=1}^{q}\hat{\beta}_{sm}v_{mk} = 0. \quad (42.60)$$

Последнее можно переписать в виде

$$\underset{p\times q}{\hat{\boldsymbol{\beta}}} = \underset{p\times q}{\boldsymbol{u}}\ \underset{q\times q}{\boldsymbol{v}^{-1}}. \quad (42.61)$$

Оценку для $\boldsymbol{\alpha}$ получим из уравнений

$$\frac{\partial \log L}{\partial \alpha_{jk}} = \frac{1}{2}n\frac{A_{jk}}{|\boldsymbol{\alpha}|} - \frac{1}{2}\sum_n\left(x_j - \sum\beta_{jl}z_l\right)\left(x_k - \sum\beta_{km}z_m\right) = 0, \quad (42.62)$$

где A_{jk} — алгебраическое дополнение α_{jk} в $|\boldsymbol{\alpha}|$. Помня, что $\boldsymbol{\alpha}$ — обращение $\boldsymbol{\sigma}$, находим

$$\hat{\sigma}_{jk} = \frac{1}{n}\sum\left(x_j - \sum_l \hat{\beta}_{jl}z_l\right)\left(x_k - \sum_m \hat{\beta}_{km}z_m\right). \quad (42.63)$$

Последнее можно переписать в виде

$$\hat{\boldsymbol{\sigma}} = \frac{1}{n}(\boldsymbol{x} - \hat{\boldsymbol{\beta}}\boldsymbol{z})(\boldsymbol{x} - \hat{\boldsymbol{\beta}}\boldsymbol{z})'. \quad (42.64)$$

42.17 Поскольку величины z — фиксированы, из (42.61) получаем

$$\mathbf{M}\hat{\boldsymbol{\beta}} = \{\mathbf{M}\boldsymbol{u}\}\boldsymbol{v}^{-1} = (\mathbf{M}\boldsymbol{x})\boldsymbol{z}'\boldsymbol{v}^{-1} = \boldsymbol{\beta}\boldsymbol{z}\boldsymbol{z}'\boldsymbol{v}^{-1} = \boldsymbol{\beta}. \quad (42.65)$$

Обозначая временно символом $\boldsymbol{V}$ обращение $\boldsymbol{v}$, из (42.61) получаем

$$\hat{\beta}_{jk} = \sum_m u_{jm}V_{mk}. \quad (42.66)$$

Учитывая (42.55), умножая на z_{lt} и суммируя по t, а затем умножая на V_{lk} и суммируя по l, получаем

$$\beta_{jk} = \sum_m u_{jm}V_{mk} - \sum_t\sum_l \varepsilon_{jt}z_{lt}V_{lk}. \quad (42.67)$$

Следовательно,

$$\hat{\beta}_{jk} - \beta_{jk} = \sum_t\sum_l \varepsilon_{jt}z_{lt}V_{lk}. \quad (42.68)$$

Вспоминая, что ε_{jt} и ε_{ku} независимы, пока t не равно u, и суммируя по соответствующим вспомогательным индексам, находим

$$\mathbf{M}(\hat{\beta}_{jk} - \beta_{jk})(\hat{\beta}_{lm} - \beta_{lm}) = \mathbf{M}\left(\sum_t\sum_\lambda \varepsilon_{jt}z_{\lambda t}V_{\lambda k}\right)\left(\sum_t\sum_\mu \varepsilon_{lt}z_{\mu t}V_{\mu m}\right) =$$

$$= \sigma_{jl}\sum_t\sum_\lambda\sum_\mu z_{\lambda t}z_{\mu t}V_{\lambda k}V_{\mu m} = \sigma_{jl}\sum_\lambda\sum_\mu v_{\lambda\mu}V_{\mu m}V_{\lambda k} =$$

$$= \sigma_{jl}\sum_\lambda v_{\lambda m}V_{\lambda k} = \sigma_{jl}V_{mk} = \sigma_{jl}V_{km}. \quad (42.69)$$

Имеется pq величин $\boldsymbol{\beta}$. Характеристики рассеяния оценок $\hat{\boldsymbol{\beta}}$ относительно средних $\boldsymbol{\beta}$ даны в (42.69). (42.69) можно записать в эквивалентной форме

$$\mathbf{M}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)'(\hat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l) = \sigma_{jl}\boldsymbol{v}^{-1}. \quad (42.70)$$

Будучи линейными функциями от x, величины $\hat{\beta}$ имеют совместное нормальное распределение.

Положив $p=1$ и транспонируя все матрицы в (42.55), мы возвращаемся к теории НК главы 19, в чем читатель может убедиться, обратившись к (42.61) и (42.70).

42.18 Имеем

$$S\left(x_j-\sum_l \beta_{jl}z_l\right)\left(x_k-\sum_m \beta_{km}z_m\right)=$$
$$=S\left(x_j-\sum_l \hat{\beta}_{jl}z_l\right)\left(x_k-\sum_m \hat{\beta}_{km}z_m\right)+S\sum_l(\hat{\beta}_{jl}-\beta_{jl})z_l(\hat{\beta}_{km}-\beta_{km})z_m,$$

и так как сумма перекрестных произведений равна нулю, и последнее выражение равно

$$S\left(x_j-\sum_l \hat{\beta}_{jl}z_l\right)\left(x_k-\sum_m \hat{\beta}_{km}z_m\right)+\sum_l\sum_m(\hat{\beta}_{jl}-\beta_{jl})(\hat{\beta}_{km}-\beta_{km})v_{lm}. \tag{42.71}$$

Последнее аналогично разложению в одномерном случае суммы квадратов ошибок на сумму квадратов остатков (отклонений от оцененной линии регрессии) и члена, описывающего отклонения оценок параметров от их истинных значений — ср. (19.42) т. 2. Исходя из тех же соображений, что привели нас к (42.69), можно показать, что математическое ожидание последнего члена правой части (42.71) равно σ_{jk}. Первый член правой части (42.71) представляет собой квадратичную форму от x, которые имеют многомерное нормальное распределение. Проверяя гипотезы, мы, однако, будем сравнивать не один член с другим, как в одномерном случае, а первый со всем выражением.

42.19 Для осуществления нашего плана нам потребуется теорема о том, что оценки характеристик рассеяния $(\hat{\sigma})$ из (42.64) имеют распределение Уишарта с заменой n на $n-q$. Из (42.64), (42.58), (42.59) и (42.60) получаем эквивалентную формулу

$$\hat{\sigma}=\{xx'-\hat{\beta}v\hat{\beta}'\}/n. \tag{42.72}$$

Мы будем действовать в том же направлении, что и в **27.22** т. 2, где было показано, что в нормальном случае частные корреляции имеют то же распределение, что и обычные корреляции, лишь с меньшим числом степеней свободы. Рассмотрим пространство n измерений. $(q\times n)$-матрица z задает в этом пространстве q векторов, скажем, $OP_1, \ldots, OP_q$, которые сами по себе лежат в q-мерном подпространстве. Далее

$$\varepsilon=x-\hat{\beta}z+(\hat{\beta}-\beta)z.$$

Слагаемые в правой части ортогональны в n-мерном пространстве. Действительно,

$$\sum_{t=1}^{n}\left(x_{jt}-\sum_l \hat{\beta}_{jl}z_{lt}\right)\sum_m(\hat{\beta}_{jm}-\beta_{jm})z_{mt}=$$
$$=\sum_m\left(u_{jm}-\sum \hat{\beta}_{jl}v_{lm}\right)(\hat{\beta}_{jm}-\beta_{jm}),$$

и первая скобка справа обращается в нуль в силу (46.58—60).

Таким образом, векторы $x-\hat{\beta}z$ ортогональны к q-мерному подпространству. Наши основные векторы ε представимы в виде суммы двух слагаемых: первое лежит в q-мерном подпространстве, второе — в ортогональном дополнении к этому подпространству.

В нашем n-мерном пространстве ортогональность влечет нулевую корреляцию, что в свою очередь влечет независимость для нормальных величин. Таким образом, оба слагаемых в правой части (42.71) независимы.

Отсюда следует, что в системе, представимой векторами $x-\hat{\beta}z$, характеристики рассеяния имеют распределение Уишарта, но не с параметром n, а с параметром $n-q$; при этом распределение сосредоточено в подпространстве, ортогональном q-мерному подпространству.

42.20 Теперь мы можем заняться проверкой гипотез, касающихся регрессии. Обычно требуется узнать, вносят ли существенный вклад в формирование x все β. Иначе говоря, если мы, используя обычную ковариационную технику, «исключим влияние» на x части величин β из какого-то подмножества, будут ли остатки существенно зависеть от оставшихся величин (βz)?

Рассмотрим q величин β и проверим гипотезу о том, что m из них $(m\leqslant q)$ равны нулю.

Предполагая, что это не так, оценим q величин β и σ и произведем подстановку в функцию правдоподобия (42.56). Теперь, если мы умножим (42.62) на $\hat{\alpha}_{jk}$ и просуммируем по j, k, мы получим, что показатель степени в функции правдоподобия становится равным постоянной. Итак, функция правдоподобия (42.56) с точностью до постоянной, сводится к $|\hat{\alpha}|^{n/2}$.

Предполагая справедливость гипотезы о равенстве нулю m величин β, мы точно так же получаем, что в этом случае функция правдоподобия сводится к некоторому числу, умноженному, скажем, на $|\hat{\alpha}_m|^{n/2}$, где α_m — обращение оценки матрицы рассеяния (42.64), или, что то же самое, (42.72), с тем исключением, что оценивается $q-m$ величин β.

Таким образом, отношение правдоподобия равно

$$l=\frac{|\hat{\sigma}|^{n/2}}{|\hat{\sigma}_m|^{n/2}}. \tag{42.73}$$

(42.73) распределено как отношение двух определителей Уишарта, основанных на выборках объемов $n-q$ и $n-(q-m)$. Кроме того, те же соображения, что использовались в **42.19**, указывают на то, что векторы, соответствующие $q-m$ ненулевым величинам β, ортогональны в q-мерном пространстве векторам, соответствующим m величинам β, которые предполагаются нулевыми. Следовательно, функции, составляющие $\hat{\sigma}_m$, представляют независимое подмножество среди тех, что входят в $\hat{\sigma}$.

Таким образом, статистика l может быть использована так же, как другие l-статистики, обсуждавшиеся раньше в этой главе. Проиллюстрируем метод следующим примером.

Пример 42.3 (Данные М. Барнард (1935) и Бартлетта (1947)).

Мисс Барнард располагала четырьмя группами черепов египтян; 91 относились к додинастическому периоду, 162 — к периоду между шестой и двенадцатой династиями, 70 — между двенадцатой и тринадцатой династиями и 75 — к династии Птоломеев. Всего 398. Над каждым из черепов производилось четыре измерения (в миллиметрах):

x_1 — максимальная ширина,
x_2 — длина основания альвеол,
x_3 — высота носа,
x_4 — высота вдоль темени.

Средние для групп, полученные Барнард, таковы:

	Группа I $n_1=91$	Группа II $n_2=162$	Группа III $n_3=70$	Группа IV $n_4=75$
x_1	133,582418	134,265432	134,371429	135,306664
x_2	98,307692	96,462963	95,857143	95,040000
x_3	50,835165	51,148148	50,100000	52,093333
x_4	133,000000	134,882716	133,642857	131,466667

(42.74)

Для сумм квадратов и произведений внутри групп (394 ст. св.) имеем:

	x_1	x_2	x_3	x_4
x_1	9661,997440	445,573301	1130,623900	2148,584219
x_2		9073,115207	1239,221990	2255,812722
x_3			3938,320351	1271,051662
x_4				8741,508829

(42.75)

Те же суммы для всех наблюдений вместе (397 ст. св.) даны в следующей таблице:

	x_1	x_2	x_3	x_4
x_1	9785,178098	214,197666	1217,929248	2019,820216
x_2		9559,460890	1131,716372	2381,126040
x_3			4088,731856	1133,473898
x_4				9382,242720

(42.76)

Наконец, суммы квадратов между классами (3 ст. св.) таковы:

	x_1	x_2	x_3	x_4
x_1	123,180658	−231,375635	87,305348	−128,763994
x_2		486,345863	−107,505618	125,313318
x_3			100,411505	−137,580764
x_4				640,733891

(42.77)

Прежде всего мы можем проверить данные на однородность, в частности, проверить, есть ли существенные различия между средними разных групп. Соответствующий критерий дается статистикой l_2 (42.17), т. е. отношением определителей (42.75) и (42.76). Имеем

$$l_2^{2/n} = 2426{,}898/2954{,}474 = 0{,}8214.$$

Отсюда выражение $-n\log l_2^{2/n}$, где $n=397$ — число степеней свободы, равно 77,3. Число степеней свободы для соответствующего χ^2-критерия равно $3\cdot 4=12$. Таким образом, мы заключаем, что данные не однородны даже по средним.

Теперь возникает сразу несколько вопросов, на которые хотелось бы иметь ответ. Например, (42.76) ясно показывает, что внутри групп есть значительная корреляция между величинами. Спрашивается, определяются ли различия между средними влиянием всех четырех величин или, например, x_3 и x_4 вносят свой вклад в эти различия лишь в силу коррелированности с x_1 и x_2? Для того чтобы ответить на это, вычислим регрессии x_3 и x_4 по x_1 и x_2, исключим их из общей схемы и подвергнем проверке остаточные матрицы. Таким образом, x_3 и x_4 у нас будут давать матрицу зависимых величин (x из **42.16**), а x_1 и x_2 играть роль величин z.

В нашем конкретном примере матрица рассеяния x_1, x_2 (см. (42.75)) равна

$$\frac{1}{394}\begin{bmatrix} 9661{,}977\,040 & 445{,}573\,301 \\ & 9073{,}115\,207 \end{bmatrix}. \qquad (42.78)$$

Матрица, обратная ей, равна

$$394\cdot 10^{-4}\begin{bmatrix} 1{,}037\,332 & -0{,}050\,942 \\ & 1{,}104\,659 \end{bmatrix}. \qquad (42.79)$$

Из (42.72) получаем, что вариация, обусловленная регрессией x по z, дается формулой

$$\hat{\beta} v \hat{\beta}' = (uv^{-1})(v)(v^{-1})' u' = uv^{-1}u'.$$

В нашем случае x относится к x_3 и x_4, а z — к x_1 и x_2, так что для последнего выражения получаем

$$10^{-4}\begin{bmatrix} 1130{,}623\,900 & 1239{,}221\,990 \\ 2148{,}584\,210 & 2255{,}812\,722 \end{bmatrix}\begin{bmatrix} 1{,}037\,332 & -0{,}050\,942 \\ -0{,}050\,942 & 1{,}104\,659 \end{bmatrix}\times$$
$$\times\begin{bmatrix} 1130{,}623\,900 & 2148{,}584\,210 \\ 1239{,}221\,900 & 2255{,}812\,722 \end{bmatrix}=\begin{bmatrix} 287{,}967\,620 & 534{,}238\,796 \\ 534{,}238\,796 & 991{,}621\,041 \end{bmatrix}. \qquad (42.80)$$

Вычитая это из матрицы, порожденной x_3 и x_4, получаем в качестве остатка

$$\begin{bmatrix} 3650{,}353\,731 & 736{,}815\,836 \\ & 7749{,}887\,788 \end{bmatrix}. \qquad (42.81)$$

с $394-2=392$ ст. св.

Точно так же обрабатывая (42.76), для полных сумм произведений получаем остаток

$$\begin{bmatrix} 3809{,}335\,190 & 611{,}698\,381 \\ & 8393{,}755\,848 \end{bmatrix}. \qquad (42.82)$$

Возникает вопрос: существенно ли отличаются матрицы (42.81) и (42.82)? Последнюю можно рассматривать как остаток в регрессии x_3 и x_4 по x_1 и x_2 плюс вектор средних, у первой же считать средние исключенными в каждом классе. Отношение определителей (42.73) равно

$$l^{2/n} = 0{,}277\,469/0{,}316\,003 = 0{,}8781.$$

При $n=392$ величина $-n\log l^{2/n}=51{,}39$, а соответствующее число степеней свободы равно $3\times 2=6$. Однородность, таким образом, отвергнута, и мы приходим к выводу, что влияние величин x_3 и x_4 существенно в том смысле, что различие между средними определяется не только величинами x_1 и x_2.

Далее мисс Барнард интересовалась, может ли каждая из этих величин иметь линейную регрессию по времени. Введем в рассмотрение время. Интервалы между четырьмя сериями предполагались взятыми в пропорции 2, 1, 2. Тогда для удобства можно величине времени приписать значения —5, —1, 1, 5.

Сделав это, получаем

$$\begin{aligned} S(t-\bar{t})^2 &= 4307{,}663\,32, \\ Sx_1(t-\bar{t}) &= 781{,}762\,86, \\ Sx_2(t-\bar{t}) &= -1407{,}260\,75, \\ Sx_3(t-\bar{t}) &= -410{,}101\,94, \\ Sx_4(t-\bar{t}) &= -733{,}427\,58. \end{aligned}$$

Рассмотрим теперь регрессию каждого из x по величине иного типа — по времени. Суммы квадратов и произведений, обусловленные регрессией (1 ст. св.), равны

	x_1	x_2	x_3	x_4
x_1	119,930358	—234,810812	68,428625	—122,377258
x_2		459,734449	—133,975163	149,601596
x_3			39,042852	—69,824358
x_4				124,874099

(42.83)

Здесь, например, член в первой строке и втором столбце равен

$$\frac{Sx_1(t-\bar{t})\,Sx_2(t-\bar{t})}{S(t-\bar{t})^2}=\frac{(781{,}762\,86)(-1407{,}260\,75)}{4307{,}663\,32}.$$

Исключим регрессию по времени из основной матрицы, вычитая (42.83) из (42.76). Получаем остаток:

	x_1	x_2	x_3	x_4
x_1	9665,247740	449,088478	1149,501013	2142,197474
x_2		9099,726441	1265,691535	2231,524444
x_3			4049,689004	1203,298256
x_4				9257,368621

(42.84)

Число ст. св. здесь равно 396.

Проверим теперь остаток на однородность, сравнивая рассеяние внутри серий, которое дается (42.79), с (42.84). Отношение определителей равно $l^{2/n}=0{,}9031$; $-n\log l^{2/n}=40$. Число ст. св. равно $2\times 4=8$. Гипотеза об однородности отвергается.

Мы приходим к выводу, что если регрессия по времени линейна, существует различие между сериями, обусловленное невременными эффектами.

Шатцов (1966b) и Консул (1966) получили в законченной форме точное распределение для l отдельно для случаев четного p или m и для случая $p\leqslant 4$.

42.21 Статистики правдоподобия, которые мы рассматривали до сих пор, представляют собой отношения определителей рассеяния того или иного вида и алгебраически относительно сходны между собой. При проверке гипотез могут оказаться полезными также статистики, представляющие собой одно или несколько собственных чисел матрицы рассеяния. Например, равенство двух матриц рассеяния $\boldsymbol{A}$ и $\boldsymbol{B}$ зависит от близости к единице корней уравнения $|\boldsymbol{A}-\lambda\boldsymbol{B}|$. Поэтому критерий, основанный на отношениях типа $|\boldsymbol{A}|/|\boldsymbol{B}|$, каким-то образом связан с собственными числами. Однако, как отмечалось в **41.27**, соответствующие распределения еще плохо затабулированы, и хотя для больших выборок мы и можем получить некоторые приближения, последние не так хороши, как в случае критериев правдоподобия, где методами **42.11** можно добиться любой степени точности. В действительности некоторые из наших статистик отношения правдоподобия являются симметрическими функциями собственных чисел. Так, произведение всех корней уравнения $|\boldsymbol{A}-\lambda\boldsymbol{B}|$ равно $|\boldsymbol{A}|/|\boldsymbol{B}|$, в чем легко убедиться, заметив, что $\boldsymbol{A}=\boldsymbol{A}\boldsymbol{B}^{-1}\boldsymbol{B}$. Задачи, где мы будем больше интересоваться отдельными корнями, появятся в следующей главе.

Пример 42.4 (Фостер и Рис, 1957. Данные Эштона, Липтона и Хили, 1957)

Рассматриваются данные двух типов измерений ($p=2$) над тремя группами особей мужского пола (человек, шимпанзе и орангутан). Измерения касались длины (x_1) и ширины (x_2) второго верхнего малого коренного зуба; результаты были прологарифмированы для того, чтобы стабилизировать рассеяние.

Суммы произведений таковы:

	Ст. св.	Суммы произведений		
		x_1^2	x_1x_2	x_2^2
Межгрупповые	2	0,544941	0,525765	0,509075
Внутригрупповые	154	0,137786	0,069342	0,092792
Всего	156	0,682727	0,595107	0,601867

(42.85)

Чтобы проверить однородность, можно рассмотреть корни уравнения

$$|\boldsymbol{A}-\lambda\boldsymbol{B}|,$$

где $\boldsymbol{A}$ — «межгрупповая» матрица, а $\boldsymbol{B}$ — общая матрица. В результате приходим к квадратному уравнению с корнями

$$\lambda_1=0{,}856\,543,$$
$$\lambda_2=0{,}020\,238.$$

В качестве статистики возьмем наибольший корень. Чем он больше, тем с большей уверенностью можно ожидать однородность. Из таблиц Фостера — Риса (см. **41.27** (1)) следует, что 99-процентный уровень равен приблизительно 0,08. Наблюденное значение гораздо больше.

Что касается средних для групп, то они оказались такими:

	Размер группы	$\bar{x}_1$	$\bar{x}_2$
Человек	59	1,846	1,981
Шимпанзе	55	1,865	2,008
Орангутан	43	1,986	2,119

Если же мы не знаем, что матрица $\boldsymbol{B}$ «больше» $\boldsymbol{A}$, чтобы проверить гипотезу о равенстве $\boldsymbol{A}$ и $\boldsymbol{B}$, мы должны, кроме того, проверить, насколько мал меньший корень. В общем случае мы здесь сталкиваемся с теоретическими трудностями, так как совместное распределение большего и меньшего корней неизвестно. Однако можно ожидать, что при больших p эти величины асимптотически независимы.

Мощность критериев

42.22 Мы уже отмечали обескураживающее обилие параметров, появляющихся в многомерной ситуации. Статистики наших критериев их не содержали, но если мы захотим точно задать альтернативы, чтобы установить мощность критерия, положение несколько осложнится.

Для основанного на статистике T^2 критерия проверки значений среднего (см. **41.16**) мощность может быть установлена по существующим таблицам. Действительно, если теоретический вектор средних равен $\boldsymbol{\mu}$, а T^2 базируется на другом векторе $\boldsymbol{\mu}_0$, т. е.

$$T^2=n(\bar{\boldsymbol{x}}-\boldsymbol{\mu}_0)'\boldsymbol{c}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu}_0), \tag{42.86}$$

распределение статистики T^2 имеет нецентральное F-распределение с p, $n-p$ степенями свободы и параметром нецентральности $n(\boldsymbol{\mu}-\boldsymbol{\mu}_0)'\boldsymbol{\gamma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)$. Следовательно, можно воспользоваться нецентральным F-распределением или одной из его аппроксимаций (см. **24.32**) (при условии, если точно определено $\boldsymbol{\gamma}$). Можно также показать (Симаик, 1941), что T^2 равномерно наиболее мощен в классе критериев, чья мощность зависит только от параметра нецентральности (одномерный случай см. в **24.36**). Аналогичные замечания справедливы и для случая двух выборок (см. **41.17**).

42.23 Для более детального изучения распределений и критериев в многомерном случае можно обратиться к книгам Т. Андерсона (1958) и Лемана (1959).

Обобщение на многомерный случай задач, связанных с критерием Беренса — Фишера различия двух средних в случае неравенства дисперсий, порождает целый ряд спорных вопросов и парадоксов (см. Молдон (1955)). Мы в **21.15** т. 2 отмечали, что задача на самом деле может быть решена методом, восходящим к Шеффе, которому удалось избежать присущих этим задачам трудностей. Этот метод был обобщен Беннетом (1951) на многомерный случай. Смотри также Т. Андерсон (1958, пункт 5.6) и (1964) и упражнение 42.12. О функциях мощностей можно прочесть у Себера (1964b), Дэрроча и Силвн (1963) и Хогга (1961). Дас Гупта, Т. Андерсон, Мудхолкар (1964), а также Т. Андерсон и Дас Гупта (1964a, b) установили монотонность функций мощностей ряда критериев для проверки некоторых многомерных гипотез.

Арнольд (1964) изучал влияние на распределение T^2 перестановок, Холловэй и Данн (1967) рассматривали устойчивость этого распределения по отношению к различиям между характеристиками рассеяния в случае двух выборок. Ито и Шулль (1964) обсуждали вопрос об устойчивости T_0^2-критерия, а также принадлежащие Лоли (1939) и Хотеллингу (1951) обобщения на случай нескольких выборок двухвыборочного T^2-критерия для проверки равенства векторных средних. Соответствующее распределение изучал Константин (1966). Пиллаи и Джаячандран (1967) вычислили мощности T_0^2-критерия ОП (42.17) и других критериев. Эти критерии асимптотически эквивалентны; в случае малых выборок они отличаются, но несущественно. Шатнов (1966b) сравнивает критерии, использующие различные статистики и приходит к выводу, что различие между (42.17) и T_0^2 несущественно. Пиллаи и Джаячандран (1967) пришли к сходному выводу для критериев независимости, где (42.14) является критерием ОП.

Постен и Баргман (1964) предложили метод для вычисления мощности критерия ОП, который одинаково пригоден для всех гипотез, налагающих одну или две связи.

42.24 Тамура (1966) построил асимптотическую теорию для свободных от распределений критериев равенства векторных параметров расположения в классе непрерывных многомерных распределений, отличающихся лишь сдвигом. Бхаттачария (1967), изучая их эффективность, пришел к выводу, что методы, использующие систему нормальных меток, предпочтительнее.

П. Сен и Пурн (1967) получили одновыборочный многомерный критерий расположения, использующий ранги.

УПРАЖНЕНИЯ

42.1 Показать, что при $p = 1$ использование статистики l_{H_2} из **42.6** эквивалентно использованию F-критерия.

42.2 Рассматривая процесс максимизации по различным множествам параметров, показать, что (42.18) всегда верно.

42.3 Следуя примеру 42.1, показать, что для статистики l_{H_1} величина $-2p \log l$ приблизительно распределена как χ^2 с $(k-1)p(p+1)/2$ ст. св. и

$$\rho = 1 - \left(\sum \frac{1}{\nu_j} - \frac{1}{\nu}\right) \frac{2p^2 + 3p - 1}{6(p+1)(k-1)}.$$

42.4 Показать, что $l^{2/n}$ из (42.45) может быть представлено в виде произведения независимых величин y_{jk}:

$$\prod_{j=2}^{q} \left\{ \prod_{k=1}^{p_j} y_{jk}^2 \right\},$$

где y_{jk} — бетта-величины с параметрами $(n - \bar{p}_j - k)/2$, $\bar{p}_j/2$;

$$\bar{p}_j = \sum_{\alpha=1}^{j-1} p_\alpha.$$

42.5 Получить равенство (42.41).

42.6 Получить равенства (42.52)—(42.54).

42.7 Выборка объема n взята из одной p-мерной нормальной совокупности. Рассмотреть гипотезы:

H: Средние для всех координат равны, все характеристики рассеяния равны.

H_1: Независимо от значений средних все характеристики рассеяния равны (т. е. все дисперсии равны и все ковариации равны).

H_2: При заданном равенстве характеристик рассеяния средние равны.

Показать, что статистики отношения правдоподобия даются формулами

$$l^{2/n} = \frac{|c_{jk}|}{(1 - r_0^2)^{p-1} (s_0^2)^p \{1 + (p-1) r_0\}},$$

$$l_1^{2/n} = \frac{|c_{jk}|}{(1 - r)^{p-1} (s^2)^p \{1 + (p-1) r\}},$$

$$(l_1 l_2)^{2/n(p-1)} = l,$$

где s^2 есть средняя дисперсия $\sum_{j=1}^{p} s_j^2/p$,

$$r^2 = \sum_{j \neq k} c_{jk} / \{p(p-1) s_0^2\},$$

а s_0^2, r_0^2 — дисперсия и корреляция, вычисленные «по всем координатам», точнее говоря,

$$s_0^2 = \frac{1}{p} \sum_{j=1}^{p} \{s_j^2 + (\bar{x}_j - \bar{x})^2\}, \qquad r_0 s_0^2 = r s^2 - \frac{1}{p(p-1)} \sum (\bar{x}_j - \bar{x})^2.$$

Показать, что $-2 \log l$, $-2 \log l_1$, $-2 \log l_2$ распределены приблизительно как χ^2 с $\frac{1}{2} p(p+3) - 3$, $\frac{1}{2} p(p+1) - 2$, $p - 1$ ст. св. соответственно.

(Уилкс, 1946.)

42.8 Используя распределение (42.25), подтвердить вывод примера 42.4.

42.9 Проверить результаты 42.12.

42.10. Получить T^2 как критерий отношения правдоподобия в форме

$$l = (1 + T^2/(n-1))^{n/2},$$

а также получить его распределение для больших выборок в нулевом случае.

42.11 Показать далее, что в общем случае, при p-мерной выборке из $N(\boldsymbol{\mu}, \boldsymbol{\gamma})$ и

$$T^2 = n(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)' c^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)$$

распределение $T^2(n-p)/(n-1)p$ имеет нецентральную F-форму с p, $n-p$ ст. св. и параметром нецентральности $\tau^2 = n(\mu-\mu_0)'\gamma^{-1}(\mu-\mu_0)$.

42.12 $x_j^{(1)}$ $(j=1, 2, \ldots, n_1)$, $x_j^{(2)}$ $(j=1, 2, \ldots, n_2)$ — выборки из двух p-мерных совокупностей $N(\mu_1, \gamma_1)$, $N(\mu_2, \gamma_2)$. Пусть

$$y_j = x_j^{(1)} - \sqrt{\frac{n_1}{n_2}}\, x_j^{(2)} + \frac{1}{\sqrt{n_1 n_2}} \sum_{j=1}^{n_1} x_j^{(2)} - \frac{1}{n_2} \sum_{k=1}^{n_2} x_k^{(2)},$$

$$j=1, \ldots, n_1, \quad k=1, \ldots, n_2,$$

так что $\bar{y} = \bar{x}^{(1)} - \bar{x}^{(2)}$.
Показать, что матрица ковариации y

$$(u_{\alpha\beta}) = \left(\delta_{\alpha\beta}\left(\gamma_1 + \frac{n_1}{n_2}\gamma_2\right)\right).$$

Определив W равенством

$$(n_1-1)W = \sum_{j=1}^{n_1} (y_j - \bar{y})(y_j - \bar{y})',$$

показать, что $T^2 = n_1 \bar{y}' W^{-1} \bar{y}$ распределено как T^2 с (n_1-1) ст. св.

(Беннет, 1951.)

42.13 Показать, что критерий, основанный на T^2, инвариантен относительно любого невырожденного линейного преобразования величин с некоторой матрицей M. Рассматривая преобразование, сводящее матрицу рассеяния c к единичной, показать, что единственной инвариантной функцией, зависящей только от $\bar{x}$ и c, является $\bar{x}'c\bar{x}$.

42.14 Обращаясь к упражнениям 42.11, 42.13, показать, что распределение T^2 может быть записано в виде постоянной, умноженной на

$$\exp\left\{-\frac{\tau^2}{2}\right\} \sum_{j=0}^{\infty} \frac{\left(\frac{\tau^2}{2}\right)\{T^2/(n-1)\}^{j+p/2}\,\Gamma\left\{\frac{n}{2}+j\right\}}{j!\,\Gamma\left\{\frac{n}{2}+p\right\}\{1+T^2/(n-1)\}^{j+n/2}}.$$

Заметив, что наиболее мощный критерий, использующий T^2 для проверки того, что $\tau=0$, представляет отношение соответствующей плотности к ее значению при $\tau=0$, показать, что такой критерий равномерно наиболее мощен.

42.15 В случае многомерной регрессии рассмотрите разбиение $\boldsymbol{\beta}$ на $\boldsymbol{\beta}_1\boldsymbol{\beta}_2$ с q_1 и q_2 столбцами соответственно. Проверяя гипотезу H: $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^*$, показать, что возведенное в степень $2/n$ отношение правдоподобия может быть записано в виде

$$l^{2/n} = \frac{|n\hat{\sigma}|}{|n\hat{\sigma} + (\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*)V_{11.2}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*)|},$$

где $\boldsymbol{V}_{11.2} = \boldsymbol{V}_{11} - \boldsymbol{V}_{12}\boldsymbol{V}_{22}^{-1}\boldsymbol{V}_{21}$, и все V есть разбиения матрицы $\boldsymbol{V}$ (42.59) на q_1, q_2 строк и столбцов, т. е. в виде

$$\begin{pmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{pmatrix}.$$

(Т. Андерсон, 1958.)

42.16 Показать, что в предыдущем примере в случае справедливости гипотезы H моменты l даются формулой

$$\mathbf{M}(l^t) = \prod_{j=1}^{p} \frac{\Gamma\left\{\frac{1}{2}(n-q_1+1-j)+t\right\}\Gamma\left\{\frac{1}{2}(n-q_2+1-j)\right\}}{\Gamma\left\{\frac{1}{2}(n-q_1+1+j)\right\}\Gamma\left\{\frac{1}{2}(n-q_2+1-j)+t\right\}}.$$

(Т. Андерсон, 1958.)

ГЛАВА 43

КАНОНИЧЕСКИЕ ВЕЛИЧИНЫ *)

43.1 Кроме проблем, которые ставит математика распределений, многомерный анализ сталкивается с одним серьезным препятствием и в сфере практического использования. Мы имеем в виду трудности, связанные с описанием сложной внутренней зависимости между случайными величинами и интерпретацией результатов анализа. Эти трудности заставляют нас пытаться, с одной стороны, уменьшить число рассматриваемых величин, а с другой — свести их к независимым. Методы настоящей главы и служат какой-то одной или обеим указанным целям.

Анализ главных компонент

43.2 Как обычно, рассмотрим n-мерные векторы-строки $\boldsymbol{x}_j$, $j = 1, 2, \ldots, p$, компонентами которых являются наблюдения над j-й компонентой исходной p-мерной случайной величиной x_{jk}, $k = 1, 2, \ldots, n$, так что в результате имеем $(p \times n)$-матрицу $\boldsymbol{x}$. Нам часто будет удобно полагать каждую компоненту x_j центрированной средним по всем ее n значениям. В этом случае наблюденная матрица рассеяния дается формулой

$$\boldsymbol{c} = \frac{1}{n}\boldsymbol{x}\boldsymbol{x}'. \tag{43.1}$$

Напомним, что если ранг $\boldsymbol{c}$ равен $m \leqslant p$, то на $\boldsymbol{x}$ наложено $p - m$ линейных связей. Последнее означает, что существует по крайней мере одно линейное преобразование, приводящее к новым величинам, число которых равно m. Можно сказать, что истинная размерность равна m и меньше p. Этот результат следует из того просто доказываемого факта, что ранг произведения матрицы на транспонированную к ней совпадает с рангом самой матрицы.

*) В оригинале — Canonical variables. Перевод предполагает, что прилагательное «случайные» опускается. Вариант «канонические переменные», по-видимому, хуже, так как в русской литературе переменными обычно называют аргументы функций. (*Прим. перев.*)

Пример 43.1

Рассмотрим $(p \times p)$ матрицу

$$\begin{vmatrix} 1 & \rho & \rho & \ldots & \rho \\ \rho & 1 & \rho & \ldots & \rho \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \rho & \rho & \rho & \ldots & 1 \end{vmatrix}. \tag{43.2}$$

Сложим столбцы и вынесем общий множитель $1 + (p-1)\rho$. Первый столбец, состоящий из единиц, умножим на ρ и вычтем из остальных. Тогда мы немедленно получим, что определитель матрицы равен

$$(1-\rho)^{p-1}\{1 + (p-1)\rho\}. \tag{43.3}$$

За исключением специальных случаев $\rho = 1$ и $\rho = -1/(p-1)$, (43.3) не равно нулю и, следовательно, ранг (43.2) равен p. Итак, мы не можем свести одинаково коррелированные величины к меньшему чем p числу величин.

Отметим без доказательства (которое можно найти у Ледермана, 1937), что число независимых связей, налагаемых на *симметрическую* матрицу, при которых она имеет ранг m, равно $(1/2)(p-m)(p-m-1)$.

43.3 Геометрически матрицу $\boldsymbol{x}$ можно понимать двояко. Можно определить p-мерное евклидово пространство размерность которого совпадает с размерностью случайного вектора и в котором точка задается выборочной последовательностью x_{jk} $(j = 1, \ldots, p)$ так, что вся наша выборка представляет собой «гроздь» из n точек. Можно также определить n-мерное пространство, размерность которого совпадает с числом наблюдений, в котором вся выборка представлена p векторами (лежащими в некотором p-мерном пространстве, вложенном в n-мерное). В любом из указанных пространств вырожденность $\boldsymbol{x}$ до ранга $m < p$ означает, что n выборочных точек лежат в некотором m-мерном подпространстве.

43.4 Рассмотрим переход к новым величинам

$$\boldsymbol{\xi} = \boldsymbol{a}\boldsymbol{x}, \tag{43.4}$$

где $\boldsymbol{a}$ — некоторая матрица коэффициентов.

Ограничим наше внимание лишь такого рода линейными преобразованиями, так как нелинейные ситуации значительно более трудны для рассмотрения. Если возникает подозрение, что мы столкнулись именно с такой ситуацией, следует сначала попытаться линеаризировать данные, например, логарифмируя их. Более того, потребуем, чтобы матрица $\boldsymbol{a}$ была ортогональной, в дальнейшем обозначая ее $\boldsymbol{l}' = (l_{mk})$. Итак,

$$\boldsymbol{l}\boldsymbol{l}' = \boldsymbol{l}'\boldsymbol{l} = \boldsymbol{I}. \tag{43.5}$$

Для характеристик рассеяния ξ, скажем $V(\xi)$, имеем

$$V(\xi) = l'cl. \tag{43.6}$$

Отсюда, конечно, следует, что

$$|V(\xi)| = |c|. \tag{43.7}$$

Имеется p^2 коэффициентов l. Равенство (43.5) налагает на них $p(p+1)/2$ связей: $p(p-1)/2$ — на недиагональные произведения и p — на диагональные. Следовательно, преобразование обладает $p(p-1)/2$ степенями свободы. Геометрически это означает вращение в нашем p-мерном пространстве.

Имеется по крайней мере одно такое преобразование, что все величины ξ не коррелированы. Последнее условие налагает $p(p-1)/2$ связей. Если такие ξ имеют дисперсии $\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2$, представленные диагональной матрицей Σ, то

$$l'cl = \Sigma, \tag{43.8}$$

и, следовательно, в силу (43.5)

$$c = l\Sigma l', \tag{43.9}$$

что эквивалентно

$$cl = l\Sigma. \tag{43.10}$$

Рассмотрев первый столбец равенства $cl = l\Sigma$, имеем *)

$$\sum_{k=1} c_{mk} l_{1k} = \sigma_1^2 l_{1m}, \qquad m = 1, 2, \ldots, p, \tag{43.11}$$

или

$$(c - \sigma_1^2 I)\, l_1 = 0, \tag{43.12}$$

где l_1 — первый столбец l. Следовательно,

$$|c - \sigma_1^2 I| = 0. \tag{43.13}$$

Аналогичные равенства справедливы и для остальных σ_r^2. Следовательно, p значений σ^2 являются собственными числами c. Величины ξ будут соответствовать проекциям на собственные векторы матрицы c. Мы будем называть такие величины *главными компонентами*.

43.5 В общем случае существует p различных собственных чисел λ матрицы c. Ниже мы увидим, что их удобно располагать в порядке убывания, т. е. считать, что $\sigma_1^2 \geqslant \sigma_2^2 \geqslant \ldots \geqslant \sigma_p^2$. Ранг матрицы c равен $p-q$ тогда и только тогда, когда последние q чисел равны нулю. Таким образом, по собственным числам можно построить критерий для ранга матрицы. Можно пойти

*) В оригинале в формуле (43.11) и в некоторых текстовых пояснениях к ней были неточности. (*Прим. ред.*)

дальше и утверждать, что если σ_p^2 мало, все изменения осуществляются в $(p-1)$-мерном пространстве и т. д. Из хода предыдущих рассуждений ясно, что все оси в нашем p-мерном пространстве первого типа ортогональны. Но мы перешли к некоррелированным величинам. Следовательно, и в нашем p-мерном пространстве второго типа (**43.3**) соответствующие векторы ортогональны. Очевидно, что примененное преобразование единственно, так как, за исключением вырожденного случая, у c имеется лишь одно множество собственных чисел. Следовательно, только наше преобразование одновременно порождает ортогональность в обоих p-мерных пространствах.

43.6 Рассмотрим дисперсию ξ_j:

$$\mathrm{D}\xi_j = l_j' c l_j, \tag{43.14}$$

где l_j — j-й столбец в l. Предположим, что мы стремимся ее максимизировать с учетом условия

$$l_j' l_j = 1.$$

Введем множитель Лагранжа λ и будем искать безусловный максимум выражения

$$l_j' c l_j - \lambda l_j' l_j. \tag{43.15}$$

После дифференцирования по l_{jk}, $k = 1, 2, \ldots, p$, приходим к системе уравнений

$$(c - \lambda I)\, l_j = 0. \tag{43.16}$$

Сравнение с (43.13) показывает, что величины λ снова оказываются собственными числами.

Из (43.14) и (43.16) следует, что максимальное значение $\mathrm{D}\xi_j$ фактически является соответствующим собственным числом. Таким образом, наша новая величина ξ_1 имеет максимальную дисперсию среди всех линейных функций от x, ξ_2 будет обладать максимальной дисперсией среди всех линейных функций, ортогональных к ξ_1 (некоррелированных с ξ_1) и т. д.

43.7 Поучительно рассмотреть ту же задачу с геометрической точки зрения. В нашем p-мерном пространстве первого типа рассмотрим n выборочных точек, координаты которых центрированны своими средними и нормированны так, что дисперсии равны единице. Итак

$$\sum_{\alpha=1}^{n} x_{j\alpha} = 0, \tag{43.17}$$

$$\frac{1}{n} \sum_{\alpha=1}^{n} x_{j\alpha}^2 = 1. \tag{43.18}$$

Рассмотрим некоторую прямую с текущими координатами X_j и направляющими косинусами u_i, так что для некоторых чисел m_i

$$\frac{X_1 - m_1}{u_1} = \frac{X_2 - m_2}{u_2} = \ldots = \frac{X_p - m_p}{u_p}. \qquad (43.19)$$

Пусть S — сумма квадратов расстояний от наших n точек до прямой. Тогда

$$S = \sum_{\alpha=1}^{n} \left[\sum_{j=1}^{p} (x_{j\alpha} - m_j)^2 - \left\{ \sum_{j=1}^{p} u_j (x_{j\alpha} - m_j) \right\}^2 \right]. \qquad (43.20)$$

Пусть числа m_j и u_j таковы, что S минимально. Равенство нулю частных производных (43.20) по каждому из m_j дает

$$-\sum_{\alpha} (x_{j\alpha} - m_j) + \sum_{\alpha} u_j \sum_{j=1}^{p} u_j (x_{j\alpha} - m_j) = 0. \qquad (43.21)$$

В силу (43.17) отсюда вытекает, что

$$\frac{m_j}{u_j} = \text{const}, \qquad j = 1, 2, \ldots, p. \qquad (43.22)$$

Следовательно, начало координат находится на прямой (43.19), и мы можем положить все m равными нулю. Впрочем, то, что прямая проходит через «центр тяжести» точек, можно было ожидать. Отсюда, используя (43.18), имеем

$$S = pn - \sum_{\alpha=1}^{n} \left(\sum_{j=1}^{p} u_j x_{j\alpha} \right)^2. \qquad (43.23)$$

Величины u подчинены условию $\sum u_j^2 = 1$. Будем искать безусловный минимум выражения

$$p - \frac{1}{n} \sum_{\alpha=1}^{n} \left(\sum_{j=1}^{p} u_j x_{j\alpha} \right)^2 + \lambda \sum_{j=1}^{p} u_j^2. \qquad (43.24)$$

Дифференцирование по u_k ведет к тому, что

$$\frac{1}{n} \sum_{j=1}^{p} \sum_{\alpha=1}^{n} x_{j\alpha} x_{k\alpha} u_j - \lambda u_k = 0, \qquad (43.25)$$

или

$$\sum_{j=1}^{p} r_{jk} u_j - \lambda u_k = 0. \qquad (43.26)$$

Исключая u, снова имеем

$$|\boldsymbol{r} - \lambda \boldsymbol{I}| = 0. \qquad (43.27)$$

Итак, рассматриваемое λ является собственным числом корреляционной матрицы. Если бы мы не ввели нормировку, после которой дисперсии становятся единичными, мы получили бы собственные числа не корреляционной матрицы, а матрицы рассеяния.

Более того, из (43.23) и (43.25) находим, что

$$S = (p - \lambda) n. \qquad (43.28)$$

Отсюда следует, что собственное число, при котором S минимально, является наибольшим собственным числом. Наша прямая соответствует ξ_1 и такова, что сумма квадратов расстояний от выборочных точек до нее минимальна.

Мы можем теперь спроектировать все наши точки на гиперплоскость, перпендикулярную прямой (43.19) и повторить процесс, строя такую прямую в этой гиперплоскости, что сумма квадратов расстояний от проекций точек до этой прямой минимальна. Наша прямая в $(p-1)$-мерном пространстве будет задаваться вторым по величине корнем уравнения (43.27). Последнее очевидно не сразу. Однако мы видели в **43.5**, что второй собственный вектор ортогонален первому и, следовательно, лежит в нашем $(p-1)$-мерном пространстве, а также то, что он получен путем максимизации дисперсии, что эквивалентно минимизации суммы квадратов расстояний до прямой.

43.8 Краткого упоминания заслуживают следующие факты:

(1). Все собственные числа матрицы рассеяния действительны и неотрицательны. Это вытекает из неотрицательной определенности матрицы. Формальное доказательство можно найти в большинстве учебников по теории матриц. Обратите внимание, однако, на предупреждение в **43.36**.

(2) Вообще говоря, собственные числа не равны. Но в частных случаях некоторые из них, и даже все, могут оказаться равными. В случае равенства не существует основанного на величине дисперсии критерия, выделяющего одну из компонент (среди группы с равными собственными числами) как имеющую какой-то приоритет. В любом ортогональном к указанной группе множестве компонент это можно сделать.

(3) Из (43.13) следует, что сумма собственных чисел равна сумме диагональных членов матрицы рассеяния, т. е. следу. Подобно этому произведение собственных чисел равно определителю матрицы рассеяния.

(4) Если $\boldsymbol{A}$ и $\boldsymbol{B}$ — две невырожденные матрицы рассеяния, собственные корни уравнения $|\boldsymbol{A} - \lambda \boldsymbol{B}| = 0$ такие же, как и у $|\boldsymbol{B}^{-1}\boldsymbol{A} - \lambda \boldsymbol{I}| = 0$. В частности, при $\boldsymbol{A} = \boldsymbol{I}$ мы видим, что

собственные числа обратной матрицы равны величинам, обратным собственным числам исходной.

(5) Величины l, соответствующие какому-либо собственному числу, определяются однозначно, за исключением того, что все они могут одновременно изменить знак.

43.9 Вопрос нормировки требует большего внимания. Принято, особенно в работах по психологии, нормировать матрицу рассеяния путем деления на соответствующие (выборочные) стандартные отклонения и, следовательно, сводить ее к корреляционной. В этом случае сумма дисперсий величин ξ_j равна размерности p. Такая процедура приводит к тому, что все p компонент ξ_j имеют одинаковый порядок наблюдаемых значений и их случайного рассеяния.

Однако собственные числа и векторы не инвариантны относительно изменения масштаба. Так, в геометрической картине из **43.7** перпендикуляры, опущенные из точек на прямую, после указанного изменения перестают быть таковыми. Таким образом, в общем случае мы получаем различные результаты в зависимости от того, менялся вначале в системе масштаб или нет. Этот факт иллюстрируется ниже в примере 43.4. Окончательное решение о желательности нормировки следует принимать, исходя из нестатистических соображений. С точки же зрения статистика, нормировка, особенно при выборочных обследованиях, приносит ряд неудобств, так как осложняет анализ распределений.

43.10 Численное решение уравнения (43.13) с помощью настольных компьютеров весьма утомительно. Описание процесса приближений к решению можно найти у Кендалла (1961 b). Появление электронно-вычислительных машин полностью изменило ситуацию. Большинство машин могут самостоятельно оперировать весьма большими матрицами и вычислять соответствующие собственные числа и векторы. Поэтому проблемы вычислений мы здесь касаться не будем.

Пример 43.2 (Лоли и Максвелл, 1963)

123 индивидуума были подвергнуты пяти психологическим тестам. Корреляции между результатами различных тестов таковы:

Тест 1	2	3	4	5
1	0,438	−0,137	0,205	−0,178
	1	0,031	0,180	−0,304
		1	0,161	0,372
			1	−0,013
				1

(43.29)

Анализ главных компонент дает нам следующие собственные числа и векторы:

Собственные числа	Коэффициенты при x_1	x_2	x_3	x_4	x_5
1,75714	0,55550	0,56470	−0,27000	0,23572	−0,49403
1,33070	−0,18568	−0,24745	−0,66199	−0,55654	−0,39538
0,78086	0,21597	0,43969	0,32041	−0,7870	0,19478
0,70916	0,64078	−0,30765	−0,39839	−0,1481	0,57950
0,42214	0,44688	−0,57611	0,47696	−0,12266	−0,47523

(43.30)

Матрица l дается пятью столбцами справа. Например,

$$\xi_1 = 0{,}55550x_1 + 0{,}56470x_2 - 0{,}27000x_3 + \\ + 0{,}23572x_4 - 0{,}49403x_5, \qquad (43.31)$$

или, читая сверху вниз,

$$x_1 = 0{,}55550\xi_1 - 0{,}18568\xi_2 + 0{,}21597\xi_3 + \\ + 0{,}64078\xi_4 + 0{,}44688\xi_5. \qquad (43.32)$$

Конечно, вряд ли исходные данные отражены в этих результатах с точностью до пятого знака. Тем не менее нам будет удобно оставить эти результаты в целях проверки. В работах по психологии принято рассматриваемые коэффициенты представлять в модифицированной форме. Вместо величин ξ рассмотрим

$$\zeta_j = \xi_j/\sqrt{\lambda_j}, \qquad (43.33)$$

так что все величины ζ имеют единичную дисперсию. Матрица коэффициентов при x для величин ζ тогда такова:

x	ζ 1	2	3	4	5
1	0,73635	−0,21419	0,19084	0,53961	0,29035
2	0,74855	−0,28545	0,38855	−0,25908	−0,37432
3	−0,35790	−0,76364	0,28313	−0,33349	0,30989
4	0,31247	−0,64200	−0,69548	−0,12247	−0,07969
5	−0,65488	−0,45609	0,48801	0,48801	−0,30877

(43.34)

Так, например,

$$x_1 = 0{,}73635\zeta_1 - 0{,}21419\zeta_2 + 0{,}19081\zeta_3 + \\ + 0{,}53961\zeta_4 + 0{,}29035\zeta_5. \qquad (43.35)$$

В тех случаях, когда величины ζ рассматривают как факторы, определяющие структуру ситуации, а коэффициенты — как веса,

с которыми эти факторы участвуют в различных величинах, эти коэффициенты называют коэффициентами (или факторами) нагрузки.

Вопросами проверки и оценивания мы займемся впоследствии. Рассматривая данные так, как они расположены, мы видим, что дисперсии величин ξ (которые по определению не коррелированы) составляют по порядку номеров 35, 27, 16, 14 и 8 процентов от общей суммы. Эти числа получаются делением собственных чисел на p (в нашем примере на 5). Довольствуясь приближенным анализом, мы можем опустить последнюю случайную величину, так как в этом случае дисперсии оставшихся четырех составляют 92 процента. Можно опустить даже последние две величины, так как дисперсии оставшихся трех составляют 78 процентов. Но для того, чтобы вычислить оставшиеся три величины ζ, мы должны измерять все пять величин x.

43.11 Мы уже отмечали, что собственные числа определяются матрицей рассеяния единственным образом. Из (43.16), независимо от того, фигурируют там выборочные или теоретические величины, ясно, что $\boldsymbol{l}_j$ также определяются единственным образом, если не учитывать возможность смены знака. Последний можно всегда определять так, чтобы, например, l_{j1} было положительным. Таким образом, установлено взаимно однозначное соответствие между собственными числами и собственными векторами с одной стороны, и матрицей рассеяния и векторами средних — с другой. Так как в нормальном случае выборочные значения последних являются оценками МП соответствующих теоретических значений, выборочные значения собственных чисел или векторов являются оценками МП соответствующих теоретических величин.

Вопрос о смещении рассматривал Дэмпстер (1966). Точные выражения довольно сложны.

Проверка статистических гипотез о собственных числах

43.12 В силу ряда причин трудно построить точную теорию, позволяющую исследовать собственные числа. Распределения здесь сложны, нормировка, как уже отмечалось, еще более усложняет дело. Кроме того, могут оказаться интересными те специальные случаи, когда собственные векторы не определены в том смысле, что некоторые из собственных чисел оказываются равными.

Выясним, какого типа гипотезы мы хотим проверять. Прежде всего будем интересоваться тем, является ли обсуждавшееся выше преобразование само по себе ценным, точнее говоря, отличаются ли собственные числа друг от друга. Если нет, то с точки зрения представлений исходные значения x не хуже, чем величины ξ. Иначе, мы хотим выяснить, являются ли значения x независимыми.

Один критерий для проверки этой гипотезы уже обсуждался в **42.12**. Статистикой там служил корреляционный определитель, логарифм которого, умноженный на $-n$, распределен приблизительно как χ^2 с $\frac{1}{2}p(p-1)$ ст. св. Более же точно, как χ^2 распределена случайная величина

$$-n\left(1-\frac{2p+11}{6n}\right)\log|r|.$$

Пример 43.3

Для данных из примера 43.2 величина корреляционного определителя (будучи произведением собственных чисел) равна 0,54659, n равно 123. Таким образом, $-123\log 0{,}54659 = 74{,}3$ приближенно является значением χ^2 с 10 ст. св. Результат крайне неблагоприятствует принятию гипотезы.

Пользуясь более точным приближением, получаем, что

$$-\left(n-\frac{2p+11}{6}\right)\log|r| = 72{,}2$$

является значением χ^2 также с 10 ст. св. Вывод остается тем же.

Это один из критериев независимости. Если же мы хотим сразу проверить и независимость и равенство дисперсий, нам следует использовать критерий сферичности из **42.13**, примененный к матрице рассеяния $\boldsymbol{c}$. Статистика здесь равна

$$-n\{\log|c| - p\log(\operatorname{tr}\boldsymbol{c}/p)\}. \tag{43.36}$$

Число ст. св. равно $\frac{1}{2}p(p+1)-1$. Чтобы получить более тонкий критерий, надо, используя (42.54), заменить n на

$$n-\frac{2p^2+p+2}{6p(n-1)}. \tag{43.37}$$

43.13 Критерии такого типа показывают, дает ли нам преимущество переход к каноническим величинам ξ.

В некотором смысле для проверки отличия собственных чисел от нуля не требуется никакого критерия. Ни одно большее нуля значение не может быть получено из генеральной совокупности, у которой соответствующее теоретическое значение равно нулю: в частности, если бы некоторые из теоретических собственных чисел равнялись нулю, то область изменений исследуемой случайной величины лежала бы в некотором подпространстве, и ни одна выборочная точка не могла бы лежать вне этого подпространства. Это не обязательно так, если значения наблюдаемой переменной подвержены ошибкам

наблюдения и измерения. Однако рассмотрение этого случая нам следует отложить до того момента, когда мы будем изучать факторный анализ.

43.14 Естественно, однако, задать следующий вопрос. Предположим, что некоторые собственные числа велики и составляют большую долю дисперсий. Значительно ли отличаются друг от друга оставшиеся числа или они характеризуют подпространство, где соответствующие величины сферичны или по крайней мере не коррелированы? Другими словами, являются ли оставшиеся λ и связанные с ними собственные векторы *различимыми*?

Бартлетт (1954 и более ранние работы), исходя из отчасти эвристических соображений, предположил, что такого рода гипотезу можно проверять приближенным χ^2-критерием. Предположим, что мы выделили первые k собственных чисел и хотим проверить, равны ли оставшиеся $p-k$ некоторому неизвестному числу. Допустим, что выборочные ошибки настолько малы по сравнению с различиями между $\lambda_1, \ldots, \lambda_k$, что мы можем установить почти точное соответствие между выборочными и теоретическими значениями λ_j при $j=1, \ldots, k$. Поскольку определитель рассеяния равен произведению собственных чисел, представляется разумным сравнить этот определитель с тем, что получается, если последние $p-k$ чисел положить равными. В том частном случае, когда рассматривается корреляционный определитель, последний равен

$$\lambda_1\lambda_2 \ldots \lambda_k \left\{\frac{p-\lambda_1-\lambda_2-\ldots-\lambda_{k-1}-\lambda_k}{p-k}\right\}^{p-k}. \qquad (43.38)$$

Предлагаемая статистика является отношением двух определителей и равна

$$(\lambda_{k+1}\lambda_{k+2} \ldots \lambda_p)^{-1} \left\{\frac{\lambda_{k+1}+\lambda_{k+2}+\ldots+\lambda_p}{p-k}\right\}^{p-k}, \qquad (43.39)$$

где все λ — выборочные значения. Последнее можно рассматривать как возведенное в степень $p-k$ отношение арифметического и геометрического среднего чисел $\lambda_{k+1}, \ldots, \lambda_p$.

Предположение состоит в том, что логарифм (43.39), умноженный на некоторый множитель, включающий n, может служить статистикой критерия, подчиненной χ^2 распределению с

$$\frac{1}{2}(p-k-1)(p-k+2) \text{ ст. св.} \qquad (43.40)$$

Лоли (1956а) показал, что если в качестве множителя взять

$$n-k-1-\frac{1}{6}\left\{\frac{2(p-k)^2+p-k+2}{p-k}\right\}+\lambda^2\sum_{j=1}^{k}\frac{1}{(\lambda_j-\lambda)^2}, \qquad (43.41)$$

а λ оценивать, как среднее от $\lambda_{k+1}, \ldots, \lambda_p$, статистика имеет моменты, совпадающие с моментами χ^2 распределения с точностью до $O(n^{-3})$. Если $\lambda_1, \ldots, \lambda_k$ велики по сравнению с λ, последний член в (43.41) можно опустить.

43.15 Строго говоря, эти результаты применимы к матрицам рассеяния с единицами по диагонали. Применение их к корреляционным матрицам ухудшает точность, поскольку мы производим нормировку с помощью выборочных дисперсий. Оказывается, что в этом случае статистика не подчиняется χ^2-распределению. Однако за неимением лучшего можно примириться и с грубым критерием, используя результаты **43.14**, как если бы они применялись к корреляционной матрице.

В связи с этим рассмотрим снова данные примеров 43.2 и 43.3. Предположим, что мы решили, что два наибольших собственных числа достаточно велики, чтобы быть уверенными, что соответствующие теоретические значения отличны между собой и отличны от остальных.

Произведение оставшихся трех чисел равно 0,23376, а их среднее 0,63739. Множитель (43.41) без двух последних членов равен 120 и значение статистики равно

$$120\{3\log 0{,}63739-\log 0{,}23376\}=12{,}3.$$

Число ст. св. в силу (43.40) равно 5. Наблюденное значение превышает 5-процентный уровень, но не превышает 1-процентный. Возникает подозрение, что последние три числа в действительности не равны.

Статистическое исследование собственных чисел в случае больших выборок

43.16 Мы продвинемся дальше, если рассмотрим асимптотическую теорию, а именно стандартные ошибки и ковариации в случае, когда все теоретические собственные числа различны. Результаты были впервые получены Гиршиком (1939).

Точность наших приближений будет такова, что нам будет безразлично, фигурируют ли в формулах выборочные или теоретические значения. Будем использовать выборочные значения. Тогда

$$\sum_{\alpha=1}^{p} l_{j\alpha}l_{k\alpha}=\delta_{jk}, \qquad (43.42)$$

$$\sum_{\alpha=1}^{p} c_{j\alpha}l_{k\alpha}=\lambda_k l_{kj}, \qquad (43.43)$$

и, используя, как и раньше, символ Кронекера, получаем из (43.43), что

$$\sum_{k=1}^{p}\sum_{\alpha=1}^{p} c_{j\alpha} l_{k\alpha} l_{km} = \lambda_k \delta_{jm}. \tag{43.44}$$

Переходя к дифференциалам в (43.43) и рассматривая случай $k = j$, получаем

$$\sum_{\alpha} c_{j\alpha}\, dl_{j\alpha} + \sum_{\alpha} dc_{j\alpha} l_{j\alpha} = \lambda_j\, dl_{jj} + d\lambda_j l_{jj}. \tag{43.45}$$

Не ограничивая общности, можно теперь предположить, что оси совпадают с «ξ-осями», так что $c_{jj} = \lambda_j$ и $c_{jk} = 0$ при $j \neq k$. Тогда первые члены в (43.45) слева и справа взаимно уничтожаются и

$$dc_{jj} = d\lambda_j. \tag{43.46}$$

Таким образом, $\operatorname{cov}(\lambda_j, \lambda_k) = \operatorname{cov}(c_{jj}, c_{kk})$, откуда, используя (41.98), получаем, что в нормальном случае

$$\operatorname{cov}(\lambda_j, \lambda_k) = \frac{2\lambda_j^2}{n}\delta_{jk}. \tag{43.47}$$

Следовательно, при $j \neq k$ λ_j и λ_k не коррелированы и

$$\mathbf{D}\lambda_j = \frac{2\lambda_j^2}{n}. \tag{43.48}$$

При нашей точности аппроксимации это означает, что

$$\mathbf{D}(\log \lambda_j) = 2/n. \tag{43.49}$$

Это удобная форма, так как дисперсия не зависит от параметров λ.

43.17 Здесь снова результаты для корреляционной матрицы по сравнению с матрицей рассеяния могут оказаться значительно более сложными. Приведем результаты Гиршика (1939):

$$\operatorname{cov}(\lambda_j, \lambda_k) = \frac{2}{n}\left\{\sum_{\alpha,\beta} l_{\alpha j}^2 l_{\beta k}^2 r_{\alpha\beta}^2 - (\lambda_j + \lambda_k)\sum_{\alpha} l_{\alpha j}^2 l_{\alpha k}^2\right\}. \tag{43.50}$$

$$\mathbf{D}\lambda_j = \frac{2}{n}\left(\lambda_j^2 + \sum_{\alpha,\beta} l_{\alpha j}^2 l_{\beta j}^2 r_{\alpha\beta}^2 - 2\lambda_j \sum_{\alpha} l_{\alpha j}^4\right), \tag{43.51}$$

где $r_{\alpha\beta}$ означают корреляции.

43.18 Тем же методом можно получить дисперсии и ковариации координат собственных векторов. Мы приведем эти результаты, поскольку они интересны, но следует помнить, что на практике нам вряд ли может понадобиться исследование отдельных направляющих косинусов.

$$\operatorname{cov}(l_{jk}, l_{mt}) = \frac{-\lambda_k \lambda_t l_{jt} l_{mk}}{n(\lambda_k - \lambda_t)^2}, \qquad k \neq t, \tag{43.52}$$

$$\mathbf{D}l_{jk} = \frac{l_{jk}^2}{4n} + \frac{\lambda_k^2}{n}\left(\frac{l_{j1}^2}{(\lambda_k - \lambda_1)^2} + \ldots + \frac{l_{jk}^2}{4\lambda_k^2} + \ldots + \frac{l_{jp}^2}{(\lambda_k - \lambda_p)^2}\right). \tag{43.53}$$

Величина $\operatorname{cov}(l_{jk}, l_{mk})$ задается (43.53) с заменой каждого l_{js}^2 на $l_{js} l_{ms}$.

По поводу некоторых более поздних результатов см. Т. Андерсон (1963). Им доказана асимптотическая нормальность распределений собственных чисел и векторов и исследован случай, когда некоторые из собственных чисел равны между собой *).

43.19 Как инструмент статистики анализ главных компонент, по-видимому, наиболее пригоден при нахождении «эффективной размерности», или при исследовании того, насколько «доминантна» та или иная линейная комбинация величин. Существует случай, когда в использовании такого анализа пошли дальше, возможно, даже переходя границы допустимого. Предположим, что наибольшее собственное число доминирует, составляя скажем 70 или 80 процентов от сумм дисперсий. Можно действовать довольно жестко и отбросить остальные компоненты, так сказать, пустив все изменения в направлении первого собственного вектора. Но может случиться так, что такой образ действий лишь субъективно желателен. Например, если значения x означают индексы **) деловой активности, банковские вклады, стоимости перевозок, характеристики импорта и т. д., мы можем просто желать, чтобы показатель общей деловой активности выражался одним числом, определяемым значением первой главной компоненты. Значение ξ_1 становится тогда просто взвешенным индексом **) и x выступают как составляющие взвеси. Является ли индекс искусственной характеристикой или действительно отражает некоторую «реальную» силу деловой активности, есть вопрос интерпретации. Решается он в свете знаний об экономической структуре изучаемой системы.

Кендалл (1961 в) показал, как ранговыми методами можно достигнуть хорошего приближения к упорядочению, связанному с выделением первой главной компоненты.

Еще меньше известно об использовании в каноническом анализе методов, свободных от распределений.

*) См. также П. В. Архаров, О предельных теоремах для характеристических корней выборочных ковариационных матриц при больших размерностях. Сборник «Статистические методы классификации», вып. 2, МГУ, 1972. (*Прим. ред.*).

**) Здесь индекс — экономический термин. (*Прим. перев.*).

Бил, Кендалл и Д. Манн (1967) дают численный метод уменьшения размерности от p до некоторого заданного q, состоящий в максимизации наименьшего из множественных коэффициентов корреляций между каждой из «отбрасываемых» величин и оставшимися величинами. Авторы сравнивают их методику с анализом главных компонент.

Пример 43.4 (Крэддок, 1965 + дополнительная информация, любезно переданная им в частной переписке)

Мэнли (1953 и позже) построил замечательно длинный ряд месячных температур Центральной Англии с 1680 по 1963 г. Температура измерялась в градусах по Фаренгейту с точностью до десятой градуса. Для удобства анализа будем считать, что год начинается в ноябре и кончается в следующем октябре.

Каждый год представлялся 12-мерным вектором, по одному значению на каждый месяц. Таким образом, проблемы масштаба не возникают. Случайные величины центрированы не отдельными месячными средними, а средним всего ряда. Это сохраняет картину изменения температуры в течение года, так что годовые вариации основных показателей нас не должны удивлять.

Таким образом, имелось 283 набора среднемесячных температур, охватывающих период с ноября 1680 по октябрь 1963 г.

Ранее в этой главе мы предполагали, что значения каждой компоненты x центрированы средним этой компоненты. Если мы центрируем какой-либо другой величиной, наши деленные на n суммы произведений являются уже не ковариациями, а моментами второго порядка. Анализ остается верным, но мы должны ожидать, что одна из компонент, возможно первая, соответствует направлению от нового среднего до выборочного среднего.

Матрица смешанных моментов дана в таблице 43.1.

Первые десять собственных чисел этой матрицы таковы:

Номер собственного числа	Процент от общей суммы
1	92,38
2	2,05
3	1,12
4	0,98
5	0,67
6	0,58
7	0,49
8	0,45
9	0,41
10	0,36
Сумма первых десяти	99,47

Таблица 43.1

Матрица смешанных моментов для температур Центральной Англии, ноябрь 1680 — октябрь 1963.
Данные представляют отклонения от общего среднего — 48,45° F

Месяц	Ноябрь	Декабрь	Январь	Февраль	Март	Апрель	Май	Июнь	Июль	Август	Сентябрь	Октябрь
Ноябрь	39,365	54,697	64,048	54,976	41,495	13,035	−20,322	−53,065	−70,060	−66,125	−42,323	−3,296
Декабрь	54,697	93,814	104,526	89,371	66,342	21,303	−33,432	−84,421	−110,803	−105,328	−67,910	−5,906
Январь	64,048	104,526	133,779	108,865	79,234	25,020	−40,215	−101,108	−131,932	−125,252	−80,368	−7,037
Февраль	54,976	89,371	108,865	102,280	71,219	22,593	−34,478	−86,008	−114,819	−108,928	−70,748	−5,630
Март	41,495	66,342	79,234	71,219	57,236	17,522	−25,164	−65,221	−85,348	−80,824	−51,806	−3,615
Апрель	13,096	21,303	25,020	22,593	17,522	10,216	−7,560	−19,716	−25,729	−24,331	−15,728	−0,756
Май	−20,322	−33,432	−40,215	−34,478	−25,164	−7,560	18,137	35,331	45,067	42,970	27,787	3,340
Июнь	−53,045	−84,421	−101,108	−88,008	−65,221	−19,716	35,331	90,261	114,083	107,854	69,624	7,044
Июль	−70,060	−110,803	−131,932	−114,819	−85,348	−25,729	45,067	114,083	151,505	141,129	90,662	9,108
Август	−66,125	−105,328	−125,252	−108,928	−80,824	−24,331	42,970	107,854	141,129	136,237	86,696	9,030
Сентябрь	−42,323	−67,910	−80,368	−70,748	−51,806	−15,728	27,787	69,624	90,662	86,696	59,000	6,237
Октябрь	−3,296	−5,905	−7,037	−5,630	−3,615	−0,756	3,240	7,044	9,108	9,030	6,237	5,507
Общее среднее	48,45	48,45	48,45	48,45	48,45	48,45	48,45	48,45	48,45	48,45	48,45	48,45
Дисперсии	6,27	9,69	11,57	10,11	7,57	3,20	4,26	9,50	12,31	11,67	7,68	2,35

Таблица 43.2

Матрица ковариаций для температур в Центральной Англии, ноябрь 1680 — октябрь 1963.
Данные представляют отклонения от месячных средних

Месяц	Ноябрь	Декабрь	Январь	Февраль	Март	Апрель	Май	Июнь	Июль	Август	Сентябрь	Октябрь
Ноябрь	6,051	1,722	0,887	−0,129	0,595	0,334	0,876	0,616	−0,021	0,308	0,424	0,782
Декабрь	1,722	9,574	4,090	1,745	1,305	1,106	0,276	0,909	0,570	0,311	0,066	0,580
Январь	0,887	4,090	14,032	4,392	1,693	−0,940	−0,026	0,629	0,855	0,698	0,677	0,695
Февраль	−0,129	1,745	4,392	11,132	3,568	1,594	0,585	0,752	1,031	0,957	−0,039	1,116
Март	0,595	1,305	1,693	3,568	7,025	1,929	0,860	0,658	0,637	0,734	0,675	1,392
Апрель	0,334	1,106	0,940	1,584	1,929	5,374	0,521	0,742	0,973	0,997	0,569	0,798
Май	0,876	0,276	−0,026	0,585	0,860	0,521	4,648	1,186	0,501	0,699	0,587	0,645
Июнь	0,616	0,909	0,629	0,752	0,658	0,742	1,186	3,827	1,269	0,847	0,768	0,475
Июль	−0,021	0,570	0,855	1,031	0,637	0,973	0,501	1,269	4,259	1,464	0,792	0,534
Август	0,308	0,371	0,698	0,957	0,734	0,997	0,699	0,847	1,464	3,763	1,452	0,898
Сентябрь	0,424	0,066	0,677	−0,039	0,675	0,569	0,587	0,768	0,792	1,452	4,148	1,004
Октябрь	0,782	0,580	0,695	1,116	1,392	0,798	0,645	0,475	0,534	0,898	1,004	5,008
Месячные средние	42,68	39,27	37,51	38,91	41,37	46,25	52,13	57,75	60,59	59,96	55,86	49,16
Дисперсии	2,46	3,09	3,75	3,34	2,65	2,32	2,16	1,96	2,06	1,94	2,04	2,24

Величина дисперсии, относящаяся к первому собственному вектору, необычно высока, но на это, конечно, есть причина — основное изменение происходит за счет смены сезонов. Коэффициенты при первых четырех собственных векторах даны в таблице 43.3 (ниже). Если выписать коэффициенты при первом собственном векторе рядом с месячными средними, приведенными в таблице 43.2, можно увидеть, что они почти точно следуют тенденции сезонных изменений.

В психологических или экономических работах мы вряд ли стали бы тратить усилия на рассмотрение остальных собственных чисел. Однако настоящий пример интересен тем, что о физической системе, которую характеризуют наши данные, мы знаем достаточно много, чтобы сделать попытку дальнейшей интерпретации. Крэддок, в статье которого читатель может найти интересующие его подробности, отождествляет вторую компоненту с климатическими изменениями средней годовой температуры, третью и четвертую — с изменением зимней температуры.

В таблице 43.2 приведены ковариации, причем при их вычислении величины центрировались месячными средними.

Первые четыре собственных числа этой матрицы таковы:

Номер собственного вектора	Процент от общей суммы
1	27,50
2	15,20
3	10,84
4	10,78
Сумма первых четырех	64,32

Теперь после удаления сезонных компонент картина остаточных изменений значительно менее ясна. Из таблицы 43.3, где даны коэффициенты при собственных векторах, видно, что первый вектор (чьи коэффициенты все положительны) характеризует изменения типа вековых (за весьма большие периоды), второй и третий характеризуют колебательное движение в течение года и, аналогично с анализом таблицы 43.1, по-видимому, связаны с изменениями типа изменений зимней температуры.

Интересно посмотреть, как изменится наш анализ, если, используя нормировку перейти от ковариаций таблицы 43.2 к корреляциям.

Таблица 43.3

Коэффициенты при собственных векторах, представляющих изменения температуры в Центральной Англии, ноябрь 1680—октябрь 1963

собственный вектор	вклад (в %)	Ноябрь	Декабрь	Январь	Февраль	Март	Апрель	Май	Июнь	Июль	Август	Сентябрь	Октябрь
Матрица ковариаций—отклонения от общего среднего													
1	92,38	−0,201	−0,323	−0,387	−0,337	−0,248	−0,077	0,128	0,322	0,421	0,399	0,258	0,024
2	2,05	0,037	0,234	0,473	0,397	0,239	0,203	0,177	0,297	0,361	0,343	0,244	0,193
3	1,12	−0,139	−0,399	−0,515	0,560	0,430	0,180	0,063	−0,057	−0,044	−0,014	−0,061	0,110
4	0,97	0,498	+0,498	−0,485	−0,268	0,238	0,213	0,205	0,099	−0,046	0,003	0,054	0,194
Матрица ковариаций—отклонения от месячных средних													
1	27,50	0,108	0,377	0,640	0,501	0,293	0,179	0,072	0,110	0,121	0,113	0,074	0,132
2	15,20	0,038	0,341	0,673	−0,374	−0,415	−0,223	−0,157	−0,047	−0,017	−0,086	−0,041	−0,178
3	10,84	−0,399	−0,423	0,449	0,226	−0,219	−0,271	−0,246	−0,222	−0,279	−0,165	−0,159	−0,218
4	10,78	−0,377	0,426	−0,261	0,489	−0,003	−0,037	−0,235	−0,109	−0,114	−0,211	−0,384	−0,310
Матрица корреляций—отклонения от месячных средних													
1	22,54	0,169	0,244	0,262	0,310	0,343	0,317	0,237	0,316	0,315	0,350	0,271	0,279
2	11,51	−0,037	0,252	0,286	0,498	0,380	0,136	−0,204	−0,260	−0,260	−0,331	−0,395	−0,020
3	10,29	0,482	0,471	0,450	−0,157	−0,341	−0,338	0,044	0,140	−0,138	−0,153	−0,027	−0,152
4	8,93	0,347	0,015	−0,467	−0,033	0,221	0,090	0,560	0,170	−305	−0,288	−0,275	0,091

Соответствующие результаты таковы:

Номер собственного числа	Процент от общей суммы
1	22,54
2	11,51
3	10,29
4	8,93
Сумма первых четырех	53,27

Хотя различия не столь велики, они все же ощутимы. Следовательно, даже в том случае, когда все p компонент вектора измеряются в одних и тех же единицах измерения, нормировка меняет дело. Мы должны были бы ожидать еще больших изменений, если бы компоненты измерялись в единицах разного масштаба.

Канонические корреляции

43.20 Преобразование совокупности величин x к канонической совокупности величин ξ по сути является преобразованием квадратичной формы к сумме квадратов. Обратимся теперь к общей теории связей между двумя совокупностями случайных величин: $x_1, \ldots, x_p$ и $x_{p+1}, \ldots, x_{p+q}$, причем $p \leqslant q$. Следуя Хотеллингу (1936) покажем, что всегда найдется линейное преобразование к таким случайным величинам $\xi_1, \ldots, \xi_p$; $\xi_{p+1}, \ldots, \xi_{p+q}$ *), что:

(а) все ξ имеют единичную дисперсию и нулевое среднее;

(б) каждое ξ в «p-группе» не коррелировано с любым ξ из этой же группы;

(в) каждое ξ в «q-группе» не коррелировано с любым ξ из этой же группы;

(г) корреляции между любым ξ из p-группы и любым ξ из q-группы равны нулю, за исключением p корреляций $\rho_1, \ldots, \rho_p$, которые можно считать корреляциями между ξ_1 и ξ_{p+1}, ξ_2 и $\xi_{p+2}, \ldots, \xi_p$ и ξ_{2p}.

В этом случае говорят, что величины ξ представлены в канонической форме, а все ρ называют каноническими корреляциями. Мы уже обсуждали канонические корреляции в связи с анализом категоризованных данных в **33.44—9** т. 2.

*) Имеется в виду, что совокупность $x_1, \ldots, x_p$ преобразуется в совокупность $\xi_1, \ldots, \xi_p$, а $x_{p+1}, \ldots, x_{p+q}$ — в $\xi_{p+1}, \ldots, \xi_{p+q}$. (*Прим. перев.*)

В случае единственной совокупности мы могли перейти к такой совокупности величин ξ, что их значения, рассмотренные по очереди, говорили нам о полной вариации столь много, сколь это возможно. Теперь этого сделать уже нельзя. Оптимизация здесь на самом деле ограничена тем, что число ненулевых корреляций не может быть меньше некоего минимально возможного числа.

Предположим, что наши величины x имеют нулевые средние и характеристики рассеяния γ_{jk}. Характеристики рассеяния в p-группе обозначим греческими индексами: $\gamma_{\alpha\beta}$, а в q-группе — латинскими индексами: γ_{jk}. Для ковариации «p-величины» и «q-величины» введем обозначение с одним греческим и одним латинским индексом: $\gamma_{\alpha j}$.

Для упрощения обозначений будем опускать индексы, относящиеся к выборке. На самом деле мы можем пойти дальше и опускать другие индексы, идентифицирующие ξ и η (см. ниже) и соответствующие коэффициенты в преобразовании. Рассмотрим теперь какую-либо пару величин (по одной на каждую группу), определенную соотношениями

$$\xi = \sum_{\alpha} l_\alpha x_\alpha, \qquad \alpha = 1, \ldots, p, \tag{43.54}$$

$$\eta = \sum_{a} m_a x_a, \qquad a = p+1, \ldots, p+q. \tag{43.55}$$

Потребуем, чтобы у них были единичные дисперсии, т. е. чтобы

$$\sum_{\alpha, \beta} l_\alpha l_\beta \gamma_{\alpha\beta} = 1, \tag{43.56}$$

$$\sum_{a, b} m_a m_b \gamma_{ab} = 1. \tag{43.57}$$

Будем теперь искать условие на коэффициенты l и m, при котором корреляция между ξ и η

$$R = \sum_{\alpha, a} l_\alpha m_a \gamma_{\alpha a}, \tag{43.58}$$

как функция от коэффициентов l и m принимала стационарное значение, иначе говоря, точка (l, m) была бы стационарной точкой для R. Взяв два, пока неопределенных множителя $\lambda/2$ и $\mu/2$, обратимся к безусловной стационарности выражения

$$\sum l_\alpha m_a \gamma_{\alpha a} - \frac{\lambda}{2} \sum l_\alpha l_\beta \gamma_{\alpha\beta} - \frac{\mu}{2} \sum m_a m_b \gamma_{ab}. \tag{43.59}$$

Дифференцирование приводит к уравнениям

$$\begin{aligned} \sum_{\alpha} l_\alpha \gamma_{\alpha a} - \mu \sum_{b} m_b \gamma_{ab} &= 0, \\ \sum_{a} m_a \gamma_{\alpha a} - \lambda \sum_{\beta} l_\beta \gamma_{\alpha\beta} &= 0. \end{aligned} \tag{43.60}$$

Умножая первое уравнение на m_a, а второе на l_α и суммируя в первом случае по a, а во втором по α, получаем в силу (43.56)—(43.58), что

$$R = \lambda = \mu. \tag{43.61}$$

Тогда уравнения (43.60) разрешимы относительно l и m, если соответствующий определитель равен нулю. Заменив μ на λ, имеем (порядок определителя равен $(p+q)^2$):

$$\begin{vmatrix} -\lambda\gamma_{\alpha\beta} & \gamma_{ab} \\ \gamma_{\alpha\beta} & -\lambda\gamma_{ab} \end{vmatrix} = 0, \qquad \begin{aligned} &\alpha, \beta = 1, \ldots, p, \\ &a, b = 1, \ldots, q. \end{aligned} \tag{43.62}$$

Умножая первые p строк на $-\lambda$ и деля последние q столбцов на $-\lambda$, получаем

$$(-\lambda)^{q-p} \begin{vmatrix} \lambda^2\gamma_{\alpha\beta} & \gamma_{ab} \\ \gamma_{\alpha\beta} & \gamma_{ab} \end{vmatrix} = 0. \tag{43.63}$$

Если мы умножим левую часть (43.63) на другой определитель порядка $(p+q)^2$, равенство сохранится. Мы будем умножать на

$$\begin{vmatrix} I_p & -\gamma_{ab}\gamma_{ab}^{-1} \\ 0 & \gamma_{ab}^{-1} \end{vmatrix}.$$

Поскольку определитель произведения равен произведению определителей, (43.63) переходит в уравнение

$$(-\lambda)^{q-p} \begin{vmatrix} \lambda^2\gamma_{\alpha\beta} - \gamma_{ab}\gamma_{ab}^{-1}\gamma_{\alpha\beta} & 0 \\ \gamma_{ab}^{-1}\gamma_{\alpha\beta} & I_q \end{vmatrix} = 0$$

или

$$(-\lambda)^{q-p} \left| \lambda^2\gamma_{\alpha\beta} - \gamma_{ab}\gamma_{ab}^{-1}\gamma_{\alpha\beta} \right| = 0. \tag{43.64}$$

Последнее уравнение имеет порядок $p+q$. $q-p$ корней равны нулю. Оставшиеся появляются парами, после того как приравнять $(p \times p)$-определитель из (43.64) нулю. Мы можем записать ненулевые корни в виде $\pm\rho_1, \ldots, \pm\rho_p$. Выберем в качестве корней те, что неотрицательны, и докажем, что они и являются теми каноническими корреляциями, что мы определили выше.

43.21 То, что ξ удовлетворяют условию (а), было получено в начале **43.20.** Подставляя в (43.60) простой корень уравнения (43.64), получаем коэффициенты l и m с точностью до знака. Если корень t-кратный, коэффициенты определяются с «точностью до $(t-1)$-й» неслучайной «постоянной»; этот результат без доказательства мы заимствуем из теории алгебраических форм.

Нам осталось доказать, что ξ в каждой группе не коррелированы и что, за исключением канонических корреляций, ξ из

одной группы не коррелировано с η из другой. Допустим, что мы выбрали некоторый корень ρ_i, определили соответствующие постоянные l_i и m_i, а следовательно, и пару соответствующих величин ξ_i и η_i. Тогда из (43.60) получаем

$$\sum l_{i\alpha}\gamma_{\alpha\alpha} = \rho_i \sum m_{ib}\gamma_{ab}, \tag{43.65}$$

$$\sum m_{i\alpha}\gamma_{\alpha\alpha} = \rho_i \sum l_{i\beta}\gamma_{\alpha\beta}. \tag{43.66}$$

Такое же равенство получим для второй пары ξ_j и η_j. Между этими четырьмя величинами существует шесть корреляций; две из них ρ_i и ρ_j. Достаточно показать, что оставшиеся четыре — нулевые. Они таковы:

$$\mathsf{M}(\xi_i\xi_j) = \sum l_{i\alpha}l_{j\beta}\gamma_{\alpha\beta}, \quad \mathsf{M}(\eta_i\eta_j) = \sum m_{i\alpha}m_{jb}\gamma_{ab}, \tag{43.67}$$

$$\mathsf{M}(\xi_i\eta_j) = \sum l_{i\alpha}m_{jb}\gamma_{ab}, \quad \mathsf{M}(\xi_j\eta_i) = \sum l_{j\alpha}m_{ib}\gamma_{ab}. \tag{43.68}$$

Умножим (43.65) на $m_{j\alpha}$ и просуммируем. В силу (43.68) имеем

$$\mathsf{M}(\xi_i\eta_j) = \rho_i\mathsf{M}(\eta_i\eta_j). \tag{43.69}$$

Точно так же, умножая (43.66) на $l_{j\alpha}$, находим, что

$$\mathsf{M}(\xi_j\eta_i) = \rho_i\mathsf{M}(\xi_i\xi_j). \tag{43.70}$$

Переставляя i и j, из (43.69) и (43.70) находим, что

$$\rho_i\mathsf{M}(\eta_i\eta_j) = \rho_j\mathsf{M}(\xi_i\xi_j). \tag{43.71}$$

Переставляя i и j в (43.71), получаем

$$\rho_j\mathsf{M}(\eta_i\eta_j) = \rho_i\mathsf{M}(\xi_i\xi_j). \tag{43.72}$$

Отсюда, если $\rho_i^2 \neq \rho_j^2$,

$$\mathsf{M}(\eta_i\eta_j) = \mathsf{M}(\xi_i\xi_j) = 0. \tag{43.73}$$

Аналогично показывается равенство нулю и других ковариаций.

Нам осталось, просматривая доказательство, убедиться, что если ρ имеет кратность t, все остается в силе. Это следует из того, что мы можем в этом случае выбрать все наши l и m подчиняющимися условиям ортогональности, которые обеспечивают равенство

$$\mathsf{M}(\xi_i\xi_j) + \mathsf{M}(\eta_i\eta_j) = 0.$$

Тогда из (43.72) будет следовать равенство нулю всех математических ожиданий, если только не равны нулю одновременно ρ_i и ρ_j. Но и в этом случае в силу (43.69) и (43.70) два математических ожидания оказываются нулевыми. Тогда мы можем выбрать наши «неслучайные» постоянные так, чтобы и остальные математические ожидания были нулевые.

43.22 Если переменные приведены к канонической форме, матрица рассеяния становится равной матрице

$$\left(\begin{array}{c:c:c} I_p & \begin{matrix}\rho_1 & & 0\\ & \rho_2 & \\ & & \ddots \\ 0 & & \rho_p\end{matrix} & 0 \\ \hdashline \begin{matrix}\rho_1 & & 0\\ & \rho_2 & \\ & & \ddots \\ 0 & & \rho_p\end{matrix} & I_p & 0 \\ \hdashline 0 & 0 & I_{q-p} \end{array}\right) \tag{43.74}$$

с определителем, равным

$$(1-\rho_1^2)(1-\rho_2^2)\ldots(1-\rho_p^2). \tag{43.75}$$

Пример 43.5 (Хотеллинг (1936), данные принадлежат Т. Келли)

140 школьникам седьмой ступени были предложены четыре теста: а) на скорость чтения; б) на способность к чтению; в) на скорость вычислений; г) на способность к вычислениям. Требуется найти канонические величины для двух тестов по чтению и двух тестов по арифметике.

Корреляции между случайными величинами таковы:

	x_1	x_2	x_3	x_4
x_1	1,0000	0,6328	0,2412	0,0586
x_2	0,6328	1,0000	—0,0553	0,0655
x_3	0,2412	—0,0553	1,0000	0,4248
x_4	0,0586	0,0655	0,4248	1,0000

Определитель (43.62) является здесь симметричным определителем:

$$\begin{vmatrix} -\lambda & (-0{,}6328)\lambda & 0{,}2412 & 0{,}0586 \\ & -\lambda & -0{,}0553 & 0{,}0655 \\ & & -\lambda & (-0{,}4248)\lambda \\ & & & -\lambda \end{vmatrix}. \tag{43.76}$$

Отсюда $(0{,}491\,370)\lambda^4 - (0{,}078\,803\,4)\lambda^2 + 0{,}000\,362\,490 = 0$

и

$$\lambda^2 = 0{,}155\,635 \quad \text{или} \quad 0{,}004\,740;$$

т. е.

$$\lambda = 0{,}3945 \quad \text{или} \quad 0{,}0688.$$

Чтобы найти сами преобразованные величины, используем (43.60). Например, если корень $\lambda = 0{,}3945$, имеем

$$l_1 + (0{,}6328)\,l_2 + (0{,}6114)\,m_1 - (0{,}1485)\,m_2 = 0, \tag{43.77}$$

$$(0{,}6328)\,l_1 + l_2 + (0{,}1402)\,m_1 - (0{,}1660)\,m_2 = 0, \tag{43.78}$$

$$-(0{,}6114)\,l_1 + (0{,}1402)\,l_2 + m_1 + (0{,}4248)\,m_2 = 0, \tag{43.79}$$

$$-(0{,}1485)\,l_1 - (0{,}1660)\,l_2 + (0{,}4248)\,m_1 + m_2 = 0. \tag{43.80}$$

Последнее уравнение линейно зависит от трех других и, следовательно, ничего не дает. Из оставшихся трех получаем отношения для l и m:

$$l_1 : l_2 : m_1 : m_2 = -2{,}7772 : 2{,}2655 : -2{,}4404 : 1.$$

Итак, преобразованные величины таковы:

$$k_1\xi_1 = -2{,}7772x_1 + 2{,}2655x_2, \tag{43.81}$$

$$k_2\eta_1 = -2{,}4404x_3 + x_4, \tag{43.82}$$

где k_1 и k_2 при желании можно выбрать из условия единичности дисперсий. Такие же уравнения, порожденные корнем 0,0688, приведут нас к другой паре канонических величин. Те, что мы получили раньше, имели максимальную корреляцию, другая пара имеет минимальную, так что представляет меньший интерес.

43.23 Стандартные ошибки можно получать в духе **43.16**. Начнем с того, что

$$\sum l_\alpha l_\beta c_{\alpha\beta} = 1, \tag{43.83}$$

$$\sum m_a m_b c_{ab} = 1, \tag{43.84}$$

$$\sum l_\alpha m_a c_{\alpha a} = r. \tag{43.85}$$

Беря полные дифференциалы, получаем

$$2\sum c_{\alpha\beta} l_\alpha\, dl_\beta + \sum l_\alpha l_\beta\, dc_{\alpha\beta} = 0, \tag{43.86}$$

$$2\sum c_{ab} m_a\, dm_b + \sum m_a m_b\, dc_{ab} = 0, \tag{43.87}$$

$$dr = \sum l_\alpha m_a\, dc_{\alpha a} + \sum l_\alpha c_{\alpha a}\, dm_a + \sum m_a c_{\alpha a}\, dl_\alpha. \tag{43.88}$$

Не ограничивая общности, можно предположить, что величины представлены в канонической форме. Все l, m, за исключением l_1 и m_1, равны нулю, так что имеем

$$2\,dl_1 + dc_{11} = 0, \tag{43.89}$$

$$2\,dm_1 + dc_{p+1,\,p+1} = 0, \tag{43.90}$$

$$dr_1 = r_1(dm_1 + dl_1) + dc_{1,\,p+1}. \tag{43.91}$$

Осуществляя подстановку из первых двух уравнений в третье, получаем

$$dr_1 = dc_{1,\,p+1} - \frac{1}{2} r_1 (dc_{11} + dc_{p+1,\,p+1}). \tag{43.92}$$

Сходные уравнения используются и для других простых корней. Например,

$$dr_2 = dc_{2,\,p+2} - \frac{1}{2} r_2 (dc_{22} + dc_{p+2,\,p+2}). \tag{43.93}$$

Перемножив (43.92) и (43.93), взяв математические ожидания и используя (41.98), находим, что

$$\operatorname{cov}(r_1, r_2) = 0. \tag{43.94}$$

Точно так же

$$\mathrm{D}r_1 = \frac{1}{n}(1 - \rho_1^2)^2. \tag{43.95}$$

Аналогичные формулы верны и для других корреляций. Следует отметить, что эта формула аналогична формуле для смешанных моментов в случае больших выборок.

Хотеллинг (1936), которому принадлежат эти выкладки, показал, что если $p = 2$, $q > 2$, а нулевой корень имеет кратность t, то nr^2 распределено как χ^2 с $t - 1$ ст. св. Если каноническая корреляция равна нулю и $p = q$, (43.95) справедливо с той оговоркой, что мы допускаем, что выборочные значения возле нулевого корня могут быть положительны или отрицательны, или что распределение r совпадает с распределением модуля нормальной величины. Лоли (1959) получил выражения для третьего и четвертого семиинвариантов r. Там же рассматривалось включающее $\operatorname{arctg} r$ преобразование, стабилизирующее рассеяние, но результаты не столь удовлетворительны, как результаты для смешанных моментов в **16.33** (т. 1).

43.24 Из (43.64) следует, что все ρ^2 являются корнями уравнения

$$|\rho^2 I - \gamma_{\alpha\beta}^{-1}\gamma_{ab}\gamma_{ab}^{-1}\gamma_{\alpha\beta}| = 0, \tag{43.96}$$

или

$$|\rho^2 I - \gamma_{11}^{-1}\gamma_{12}\gamma_{22}^{-1}\gamma_{21}| = 0, \tag{43.97}$$

где γ_{11} — матрица, соответствующая p-величинам $x_1, \ldots, x_p$;

γ_{22} — такая же матрица для $x_{p+1}, \ldots, x_{p+q}$; γ_{12} — ковариационная матрица между p-величинами и q-величинами, а γ_{21} определяется аналогично γ_{12}. Таким образом, ρ^2 — собственные корни произведения матриц в (43.97).

43.25 Результаты анализа канонических корреляций труднее интерпретировать, чем даже результаты анализа компонент. Эти результаты лучше всего рассматривать как вспомогательные, помогающие осознать структуру изучаемой многомерной системы. В любом случае они говорят, какой может быть максимальная корреляция линейных функций от двух групп переменных. В литературе по этому вопросу содержится немного примеров полезного практического применения; см. например, работу Барнетта и Льюиса (1963), в которой с этих позиций проводятся некоторые педагогические исследования. По этой причине оставшиеся теоретические факты мы отметим бегло и без доказательства.

(а) Для простоты изложения мы полагали $q \geqslant p$. Если $q < p$, мы должны просто группы величин поменять ролями.

(б) Заменив матрицы в (43.97) их МП-оценками, мы получим МП-оценки канонических корреляций.

(в) Рассматривая матрицы, входящие в (43.64), мы видим, что одна из них дает нам рассеяние в p-группе, а другую (произведение трех) можно интерпретировать как вклад от регрессии p-группы по q-группе. Таким образом, теория регрессии (**42.15—20**) здесь используется. Распределение собственных корней ρ^2 в (43.97) такое же, как и у λ в **41.22—3** при условии, что p-группа и q-группа независимы. К сожалению, это наименее интересный случай.

(г) Бартлетту (1947) принадлежит критерий, аналогичный критерию из **43.14** и основанный на представлении корреляционного определителя как произведения p сомножителей $1 - \rho_j^2$. Если предположить, что k канонических корреляций не равны нулю, то статистика критерия для проверки того, что остальные равны нулю, такова:

$$-\left\{n - 1 - k - \frac{1}{2}(p + q + 1) + \sum_{j=1}^{k} r_j^{-2}\right\} \log \prod_{j=k+1}^{p} (1 - r_j^2). \quad (43.98)$$

Последнее распределено приближенно как χ^2 с $(p - q)(q - k)$ ст. св. Лоли (1959) исследовал этот критерий, получив весьма удовлетворительные результаты.

(д) В одном случае, а именно, когда все канонические корреляции, за исключением одной, равны нулю, в практическом использовании можно ожидать некоторого прогресса. (См. Бартлетт (1947)).

(е) Дэмпстер (1966) методом Кенуя решал задачу устранения смещения у оценок канонических корреляций (ср. **17.10** т. 2).

Единое изложение канонического анализа и связанных с ним проблем многомерности дано у Е. Унлльямса (1967).

Факторный анализ

43.26 До сих пор в этой главе мы обсуждали методы, приспособленные к изучению системы с точки зрения структуры, которой она обладает. Теперь мы возьмемся за изучение проблемы, так сказать, с другого конца.

Более конкретно, предположим, как обычно, что у нас есть $(p \times n)$-матрица n наблюдений над $(p \times 1)$-вектором $\boldsymbol{x}$; предположим, что наблюдаемые x в действительности являются линейными функциями от некоторых величин ζ, которые подлежат изучению и называются факторами. Всего таких величин $m < p$. Таким образом,

$$x_j = \sum_{k=1}^{m} l_{jk}\zeta_k + \varepsilon_j. \quad (43.99)$$

Коэффициенты l не являются теперь характеристиками вращения до новых осей координат. Как и при анализе компонент, мы будем интерпретировать их как коэффициенты нагрузки. (Термин идет от ранних работ по психологии и соответствует более близкому статистикам термину «вес».) Предполагается, что все ζ независимы и нормально распределены с нулевыми средними и единичными дисперсиями. Так как в общем случае комплекс x имеет размерность p, представление их в виде линейных функций от величин ζ требует введения члена ε, указывающего на ошибку. В нашей модели мы предполагаем, что ε_j не зависит от ε_k и всех ζ. Задача состоит в том, чтобы оценить постоянные l и дисперсии σ_j^2 величин ε_j.

Это не модель регрессии. Наши ζ случайны, так что мы не можем обращаться с ними, как с фиксированными величинами, наподобие независимых переменных в уравнении регрессии. Здесь связь носит структурный характер в смысле главы 29 т. 2.

43.27 Первое, что следует отметить в нашей модели, это ее неопределенность. Мы представляем p-мерный комплекс в виде функций от $m + p$ случайных величин. Соотношения (43.99) содержат pm постоянных l, mn величин ζ и pn величин ε. (43.99), как система алгебраических уравнений, имеет много решений. Мы уже ввели условие, состоящее в том, что величины ζ имеют распределение $N(0, 1)$. Предположим, что все ε также нормальны и имеют нулевые средние. Спрашивается, становится ли задача оценивания pm постоянных l и p постоянных

σ^2 определенной после введения условия независимости всех величин ζ и ε? Так как все ζ и ε нормальны, x также нормальны. Имеем

$$\operatorname{cov}(x_j, x_k) = \mathsf{M}\{\sum l_{jt}\zeta_t + \varepsilon_j\}\{\sum l_{kt}\zeta_t + \varepsilon_k\} = \sum_{t=1}^{m} l_{jt}l_{kt}, \quad j \neq k, \tag{43.100}$$

$$\mathsf{D}x_j = \sum_{t=1}^{m} l_{jt}^2 + \sigma_j^2. \tag{43.101}$$

Объединим эти соотношения:

$$\gamma = ll' + \Sigma, \tag{43.102}$$

где l — $(p \times m)$-матрица коэффициентов l_{jt}; а Σ — диагональная матрица с σ_j^2 по диагонали $(j = 1, \ldots, p)$.

Число характеристик рассеяния в левой части (43.102) равно $p(p+1)/2$, число постоянных в правой равно $p(m+1)$. Следовательно, если $m + 1 > (p+1)/2$, соотношений (43.102) недостаточно, чтобы определить постоянные. На самом деле мы будем нормировать постоянные l, полагая что

$$\sum_{t=1}^{p} l_{tj}l_{tk}/\sigma_t^2 = 0, \qquad j \neq k, \tag{43.103}$$

или, эквивалентно, что

$$l'\Sigma^{-1}l = J, \tag{43.104}$$

где J — некоторая диагональная $(m \times m)$-матрица. Это накладывает еще $m(m-1)/2$ условий на оцениваемые постоянные. Система уравнений остается неопределенной, если

$$\frac{1}{2}p(p+1) < p(m+1) - \frac{1}{2}m(m-1)$$

или

$$(p-m)^2 < p + m. \tag{43.105}$$

Поэтому мы будем предполагать, что выполнено противное.

Пример 43.6

Неравенство, обратное к (43.105), эквивалентно

$$\left\{\left(p + \frac{1}{2}\right) - m\right\}^2 \geqslant \frac{1}{4}(8p+1). \tag{43.106}$$

Например, если $p = 5$, наша модель является неопределенной при $m > 2$, и мы не должны приписывать модели 5-мерного комплекса более чем два фактора. Если $p = 10$, наибольшее пригодное значение m равно 5.

43.28 Укажем на причины введения условий ортогональности (43.103). Рассмотрим невырожденное ортогональное преобразование величин ζ к новым переменным η:

$$\zeta = H\eta.$$

Величины η будут также независимыми и стандартно-нормальными и вместо (43.102) мы должны написать

$$\gamma = lH(lH') + \Sigma = lHH'l' + \Sigma = ll' + \Sigma.$$

Короче говоря, наши ζ определены с точностью до ортогонального преобразования. Введение условия (43.104) представляет весьма удобный способ устранить эту неопределенность, однако следует отметить, что выбранный нами путь, конечно, не единствен.

43.29 Если, как мы впредь и будем предполагать, $(p-m)^2 > p + m$, в системе (43.102)—(43.103) содержится больше уравнений, чем неизвестных. Не имея возможности привести чисто алгебраическое решение, мы привлечем некоторую компромиссную вспомогательную процедуру. Следуя Лоли (Лоли и Максвелл, 1963), используем метод максимального правдоподобия.

Одна из особенностей нашей ситуации состоит в том, что, хотя мы имеем дело с МП-оцениванием в условиях нормальных распределений, выборочные ковариации x не являются оценками теоретических ковариаций. Причиной этому являются ограничения на оценивание, порожденные условиями (43.102)—(43.103). Если бы мы взяли наблюдаемые c в качестве оценок γ, мы бы просто имели

$$c = \hat{l}\hat{l}' + \hat{\Sigma},$$
$$\hat{J} = \hat{l}'\hat{\Sigma}^{-1}\hat{l}.$$

Как мы вскоре увидим, второе уравнение верно, но первое верно лишь для диагональных элементов.

Начнем с логарифма функции правдоподобия:

$$\log L = \text{const} - \frac{1}{2}n\log|\gamma| - \frac{1}{2}n\sum \Gamma_{jk}c_{jk}, \tag{43.107}$$

где Γ — обращение γ. Подстановка выражения для γ из (43.102) дает нам функцию, которую мы будем максимизировать по l и Σ. Там, где это не может вызывать недоразумения, будем опускать «крышечки» сверху, чтобы упростить запись.

Дифференцируя по σ_t^2, после некоторых преобразований получаем, что

$$\Gamma_{tt} - \sum_{j,k} \Gamma_{tj}c_{jk}\Gamma_{kt} = 0. \tag{43.108}$$

Объединим эти соотношения в виде

$$\operatorname{diag}(\gamma^{-1} - \gamma^{-1}c\gamma^{-1}) = 0. \qquad (43.109)$$

Дифференцируя по l_{jk}, после некоторых сокращений получаем, что

$$\sum_t l_{tk}\Gamma_{tj} - \sum_{t,u,v} l_{tk}\Gamma_{tu}c_{uv}\Gamma_{vj} = 0.$$

Последнее есть элемент j-й строки и k-го столбца матричного соотношения

$$l'\gamma^{-1} - l'\gamma^{-1}c\gamma^{-1} = 0. \qquad (43.110)$$

Чтобы быть последовательными, наряду с (43.108), (43.110) мы должны использовать и уравнения (43.102), (43.104), примененные к оценкам МП *):

$$\gamma = ll' + \Sigma. \qquad (43.111)$$

$$J = l'\Sigma^{-1}l. \qquad (43.112)$$

Умножая (43.110) справа на γ, имеем

$$l' - l'\gamma^{-1}c = 0, \qquad (43.113)$$

и, следовательно,

$$l'c^{-1} = l'\gamma^{-1}. \qquad (43.114)$$

Умножая (43.109) слева на $\gamma - ll'$, т. е. на Σ, имеем **)

$$\operatorname{diag}(I - c\gamma^{-1} - ll'\gamma^{-1} + ll'\gamma^{-1}c\gamma^{-1}) = 0.$$

Последнее, в силу (43.113), приводит к тому, что

$$\operatorname{diag}(I - c\gamma^{-1}) = 0 \qquad (43.115)$$

Умножим это справа на $\gamma - ll'$. Аналогично предыдущему получаем, что

$$\operatorname{diag}(\gamma - c) = 0. \qquad (43.116)$$

Последнее эквивалентно уравнениям

$$\hat{\sigma}_j^2 = c_{jj} - \sum_{k=1}^{m} \hat{l}_{jk}^2. \qquad (43.117)$$

Теперь, умножая (43.112) справа на l', получаем

$$Jl' = l'\Sigma^{-1}ll' = l'\Sigma^{-1}(\gamma - \Sigma) = l'\Sigma^{-1}\gamma - l'. \qquad (43.118)$$

Итак,

$$Jl'\gamma^{-1} = l'\Sigma^{-1} - l'\gamma^{-1},$$

*) В (43.111)—(43.112) на самом деле должны присутствовать «крышечки». (*Прим. перев.*)

**) Здесь следует помнить, что Σ диагональна. (*Прим. перев.*)

что в силу (43.114) приводит к тому, что

$$Jl'c^{-1} = l'\Sigma^{-1} - l'c^{-1},$$

откуда

$$\hat{J}\hat{l}' = \hat{l}'(\hat{\Sigma}^{-1}c - \hat{\Sigma}^{-1}\hat{\Sigma}) = \hat{l}'\hat{\Sigma}^{-1}(c - \hat{\Sigma}). \qquad (43.119)$$

43.30 Эти уравнения все еще трудно решить. Вспоминая, что J диагональна, из (43.119) находим, что ее элементы суть собственные числа $\Sigma^{-1}(c - \Sigma)$. Воспользуемся одной итерационной процедурой. Припишем в качестве первого приближения какие-либо значения элементам Σ, определим из (43.119) собственные векторы $\hat{l}$; подставляя их в (43.117), получим улучшенные значения σ_j^2; продолжим итерационный процесс, подставив эти значения снова в (43.119), и т. д. Оценки для γ получаем из (43.111).

Этот процесс, однако, может сходиться довольно медленно (см. Хоув, 1955). В ряде случаев оценки некоторых σ^2 стремятся к нулю. Вообще, сказать, что эта тема хорошо разработана, нельзя.

43.31 После того как получены удовлетворительные оценки, с помощью обычного отношения правдоподобия можно проверить, насколько удачен выбор числа факторов m. Если гипотеза о том, что в действительности имеется m факторов справедлива, логарифм правдоподобия пропорционален

$$-\frac{1}{2}n\log|\hat{\gamma}| - \frac{1}{2}\operatorname{tr}(c\Gamma). \qquad (43.120)$$

Если справедлива гипотеза о том, что с нулевой ошибкой ε x нормальны и независимо распределены, выборочные характеристики рассеяния являются оценками теоретических значений и логарифм правдоподобия пропорционален

$$-\frac{1}{2}n\log|c| - \frac{1}{2}n\operatorname{tr}(cc^{-1}) = -\frac{1}{2}n\log|c| - \frac{1}{2}np.$$

Таким образом, «отношение»

$$-n\left\{\log\frac{|c|}{|\gamma|} - \operatorname{tr}(c\Gamma) + p\right\} \qquad (43.121)$$

распределено приблизительно как χ^2. Число степеней свободы равно числу постоянных, подобранных для второго случая, без числа постоянных для первого случая, что равно, как мы отмечали в (43.105),

$$\{(p - m)^2 - (p + m)\}/2. \qquad (43.122)$$

Бартлетт (1951a) предположил, что, используя вместо n в (43.121) множитель

$$n' = n - \frac{1}{6}(2p + 11) - \frac{2}{3}m, \qquad (43.123)$$

можно получить лучшую аппроксимацию. Это во всяком случае так, если $m = 0$ (**42.12**) и, по-видимому, будет так, если n и p заменить на $n - m$ и $p - m$.

Отвергая гипотезу, мы утверждаем, что нам требуется большее число факторов. Вычисление статистики (43.121) довольно утомительно, а ее значение весьма чувствительно к точности оценивания параметров $\boldsymbol{l}$ и $\boldsymbol{\Sigma}$. Лоли и Максвелл (1963) предложили приближенную форму

$$n' \sum_{j<k} \frac{(c_{jk} - \hat{\gamma}_{jk})^2}{(\hat{\sigma}_j \hat{\sigma}_k)^2}, \qquad (43.124)$$

где n' дано в (43.123).

43.32 До появления электронно-вычислительных машин психологи были вынуждены изобретать различные приемы численного решения задач факторного анализа. Некоторые из этих приемов нельзя назвать иначе, как мерой отчаяния, другие до сих пор могут оказаться полезными, во всяком случае, при нахождении первого приближения, с которого следует начать итерационный процесс решения точных уравнений. Правда, теоретически и хоть в какой-то мере строго обосновать эти приемы весьма трудно. Подробности и численные примеры можно найти у Хармэна (1960), Кендалла (1961 b) и Лоли и Максвелла (1963).

43.33 В факторном анализе, так же как и в анализе компонент, величины x выражаются в виде взвешенных сумм некоторых ненаблюдаемых величин ζ. Обычно последние и невозможно наблюдать непосредственно. Как правило, основная трудность состоит в интерпретации результатов. Психологи обычно связывают величины ζ с некоторыми факторами, которые, как считается, и определяют структуру системы. Применение такого приема к физическим системам очень часто приводит к взвешенным суммам величин, которые уже весьма слабо поддаются интерпретации. По этой и ряду других причин возникает вопрос о возможности полезных модификаций нашей модели.

43.34 Прежде всего вспомним, что для того, чтобы получить единственное решение, на веса факторов мы накладывали условие ортогональности (43.103).

Однако это условие не вытекает из существа модели, и, получив величины $\boldsymbol{l}$, мы вправе как угодно преобразовывать величины ζ в соответствующем им m-мерном пространстве. Короче говоря, мы можем вращать факторы и даже можем преобразовать их в зависимые факторы. На самом деле мы, образно говоря, оцениваем факторное пространство и не привязаны к какой-то особой системе координат в нем. Вариантов бесконечно много, и наш выбор должен определяться в каждом случае соображениями нестатистического характера. Два критерия напрашиваются сами собой.

(а) Вращать так, чтобы как можно больше величин $\boldsymbol{l}$ оказывались нулевыми (или становились в каком-то смысле минимальными). Это равносильно тому, чтобы следовать некоторому закону экономии в представлении x в виде функций от ζ: чем меньше ζ присутствует в соотношениях, тем лучше.

(б) Вращать так, чтобы нагрузка некоторых факторов становилась максимальной. Как правило, такой образ действий сводится к увеличению значений некоторых $\boldsymbol{l}$ за счет оставшихся. Однако это может привести к отождествлению какого-либо фактора с одним из x.

С другой, но столь же разумной, точки зрения, мы можем накладывать условия на $\boldsymbol{l}$ с самого начала. Например, мы можем потребовать, чтобы x_1 и x_2 включали лишь фактор ζ_1, x_3 и x_4 — факторы ζ_1 и ζ_2 и так далее. Это эквивалентно тому, что (приравнивая нулю нагрузки определенных факторов) мы определяем структуру нашей системы.

43.35 Очевидно, что в этом случае задачи оценивания становятся еще более сложными, чем в стандартном случае **43.29**, особенно, если мы допускаем, что некоторые факторы могут быть коррелированы. Однако обсуждать эту тему мы здесь не будем, поскольку уровень развития этой области не настолько высок, чтобы можно было хоть как-то осмыслить возникающие теоретические вопросы. Электронные вычислительные машины снова приходят здесь на помощь психологам, позволяя уточнять критерии с точки зрения определения вращений или других структурных упрощений и решать получающиеся уравнения. Но даже машина порой не может обеспечить нас точной информацией о выборочных распределениях и окончательных значениях оценок.

43.36 Хочется, однако, предостеречь против попытки использования факторного или компонентного анализа при исследовании матриц, которые получены не методом смешанных моментов.

Так, например, элементы корреляционной матрицы можно оценить с помощью тетрахорических или бисериальных коэффициентов (см. **26.27** т. 2). В этом случае матрица может и не быть положительно определенной, так что собственные числа могут оказаться и отрицательными.

УПРАЖНЕНИЯ

43.1 p-мерный комплекс случайных величин имеет следующую корреляционную матрицу:

$$\begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{p-1} \\ \rho & 1 & \rho & \dots & \rho^{p-2} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \rho^{p-1} & \rho^{p-2} & \rho^{p-3} & \dots & 1 \end{pmatrix}.$$

Показать, что определитель этой матрицы равен $(1-\rho^2)^{p-1}$ и, следовательно, «истинная размерность» комплекса не может быть меньше чем p.

43.2 Показать, что если $\rho > 0$, комплекс примера (а не упражнения) 43.1 имеет одно наибольшее собственное число, а остальные числа равны между собой. Убедитесь, что сумма собственных чисел равна p.

43.3 Корреляция между j-й и k-й величинами в p-мерном комплексе равна $1-|j-k|/p$. Показать, что размерность комплекса нельзя понизить. Показать, что при $p=4$ собственные числа равны $(2 \pm \sqrt{2})/4$ и $(6 \pm 2\sqrt{6})/4$.

43.4 *) Показать, что если λ_i — собственные числа матрицы рассеяния A, λ_i^2 — собственные числа матрицы A^2. Пусть λ_1 — наибольшее собственное число A. Показать, что при больших k i-й диагональный элемент матрицы A^k асимптотически эквивалентен λ_1^k, умноженному на квадрат i-й координаты собственного вектора, соответствующего собственному числу λ_1.

43.5 Пусть в обозначениях 43.20

$$A = |\gamma_{a3}|, \qquad B = |\gamma_{ab}|,$$

$$C = \begin{vmatrix} 0, & \gamma_{aa} \\ \gamma_{aa} & \gamma_{ab} \end{vmatrix}, \qquad D = \begin{vmatrix} \gamma_{a3} & \gamma_{aa} \\ \gamma_{aa} & \gamma_{ab} \end{vmatrix}.$$

Показать, *что векторный коэффициент корреляции* K, определенный как

$$K^2 = (-1)^p C/(AB),$$

и квадрат векторого коэффициента **)

$$Z = D/(AB)$$

инвариантны относительно любого линейного преобразования. Показать также, что

$$K = \pm \prod_{j=1}^{p} \rho_j,$$

$$Z = \sum_{j=1}^{p} (1-\rho_j^2),$$

где ρ_j — канонические корреляции.

(Хотеллинг, 1936.)

*) Второе утверждение примера не совсем точно. Сказанное будет верным, если направление соответствующего собственного вектора не перпендикулярно ни к одной из осей координат. (*Прим. ред. и перев.*).

**) В оригинале vector alienation coefficient, в русской литературе такого термина нет. В одномерном случае — это величина $\sqrt{1-r^2}$, где r — смешанный коэффициент корреляции. (*Прим. перев.*)

43.6 Пусть в предыдущем примере k и z — выборочные значения K и Z. Показать, что если канонические корреляции генеральной совокупности различны, то

$$\mathrm{D}k = \frac{1}{n} K^2 \sum_{j=1}^{p} \frac{(1-\rho_j^2)^2}{\rho_j^2},$$

$$\mathrm{D}z = \frac{4}{n} Z^2 \sum_{j=1}^{p} \rho_j^2,$$

$$\mathrm{cov}\,(k, z) = -\frac{2}{n} KZ \sum_{j=1}^{p} (1-\rho_j^2).$$

В частности, если $p=2$, то

$$\mathrm{D}k = \frac{1}{n}\{(1-K^2)^2 - Z(1+K^2)\},$$

$$\mathrm{D}z = \frac{4Z^2}{n}(1-Z+K^2),$$

$$\mathrm{cov}\,(k, z) = -\frac{2}{n} KZ(1+Z-K^2).$$

(Хотеллинг, 1936.)

43.7 Положим в предыдущем примере $p=q=2$. Показать, что при стандартной нормировке

$$k = \frac{r_{13}r_{24} - r_{14}r_{23}}{\{(1-r_{12}^2)(1-r_{34}^2)\}^{1/2}},$$

и получить отсюда критерий для гипотезы о том, что «разность тетрады» $r_{13}r_{24} - r_{14}r_{23}$ равна нулю.

(Хотеллинг, 1936.)

43.8 Показать, что в обозначениях примера 43.6

$$\mathbf{M}(k^\alpha z^\beta) = \prod_{j=1}^{p} \frac{\Gamma\left\{\frac{1}{2}(q+\alpha+1-j)\right\}\Gamma\left\{\frac{1}{2}(n-q+2\beta-j)\right\}\Gamma\left\{\frac{1}{2}(n-j)\right\}}{\Gamma\left\{\frac{1}{2}(q+1-j)\right\}\Gamma\left\{\frac{1}{2}(n-q-j)\right\}\Gamma\left\{\frac{1}{2}(n+\alpha+2\beta-j)\right\}}.$$

(Гиршик, 1939.)

43.9 Показать, что если все собственные числа матрицы рассеяния нормального распределения равны, скажем, единице, то *выбранное случайным образом* выборочное собственное число имеет единичное среднее и дисперсию $(p+1)/n$.

(Гиршик, 1939.)

43.10 Показать, что в случае нулевых теоретических корреляций выборочная «разность тетрады» из примера 43.7, имеет распределение

$$dF = \frac{4(n-2)\,\Gamma^2\left(\frac{1}{2}n\right)}{\pi\Gamma^2\left\{\frac{1}{2}(n-1)\right\}} \int_u^1 \int_{u/t}^1 \frac{(tv-u)^{n-3}\,dt\,dv}{\{(1-t^2)(1-v^2)\}^{1/2}}\,du.$$

(Гиршик, 1939.)

43.11 Используя обозначения **43.29**, показать, что

$$H = l'\Sigma^{-1}(c - \Sigma)\Sigma^{-1}l$$

равно J^2 и, следовательно, диагонально.

(Лоли и Максвелл, 1963.)

43.12 Показать, что если имеется возможность изменять дисперсии «ошибок», присутствующих в схеме факторного анализа, и

$$p \geqslant \frac{1}{2}(p - m)(p - m + 1),$$

то эти дисперсии можно выбрать таким образом, чтобы число требуемых факторов не превышало m.

43.13 Рассматривая факторный анализ при $p = 2$ и $m = 1$, выписать функцию правдоподобия и, используя дифференцирование, показать, что уравнение МП имеет следующий вид:

$$-c_{12}l_2 + \left(c_{11} - \frac{c_{11}\sigma_1}{l_1}\right)l_2 = 0,$$

$$c_{22}\left(1 - \frac{\sigma_2}{l_2}\right) - c_{12}l_2 = 0.$$

Вывести отсюда, что

$$c_{22}l_1^2 = c_{11}l_2^2,$$

$$\sigma_2/l_2 = \sigma_1/l_1.$$

Такой результат недопустим при свободном оценивании четырех параметров. Объяснить его появление.

43.14 Убедиться, что определитель (43.74) действительно равен (43.75).

ГЛАВА 44

ДИСКРИМИНАЦИЯ И КЛАССИФИКАЦИЯ

44.1 В этой главе мы коснемся задач дифференциации двух и более многомерных генеральных совокупностей по результатам измерений. Существуют три различных класса задач, которые часто смешивают:

(а) Дискриминация. Известно о существовании двух или более генеральных совокупностей, и дано по одной выборке из каждой совокупности. Задача заключается в выработке основанного на имеющихся выборках правила, позволяющего приписать некоторый новый элемент к правильной генеральной совокупности в том случае, когда нам заведомо неизвестно, к какой совокупности он принадлежит.

(б) Классификация. Рассматривается некоторая выборка или вся генеральная совокупность. Задача состоит в классификации элементов выборки или совокупности по группам, столь различным между собой, сколь это возможно. При дискриминации существование групп дано, при классификации выявление их и составляет задачу.

(в) Разбиение. Рассматривается некоторая выборка или вся генеральная совокупность. Мы хотим разбить их на группы независимо от того, естественны ли границы разбиения или нет.

Обратимся к примерам. Допустим, что мы располагаем информацией о некотором количестве индивидуумов из двух различных рас. Нам может понадобиться функция, позволяющая приписывать новые индивидуумы к соответствующим им расам. Построение такой функции на основе уже имеющейся информации и составляет задачу дискриминации. Другой пример: дана совокупность элементов неизвестного происхождения и требуется выяснить, возможно ли разбиение этих элементов на группы, естественное в том смысле, что элементы каждой группы сходны между собой, но элементы одной группы значительно отличаются от элементов другой. Это задача классификации. Наконец, допустим, что мы хотим разбить на группы некоторое количество студентов в зависимости от их успехов на экзаменах, понимая при этом, что границы такого разбиения могут быть весьма условны. Это задача разбиения, и она

43.11 Используя обозначения **43.29**, показать, что

$$H = l'\Sigma^{-1}(c - \Sigma)\Sigma^{-1}l$$

равно J^2 и, следовательно, диагонально.

(Лоли и Максвелл, 1963.)

43.12 Показать, что если имеется возможность изменять дисперсии «ошибок», присутствующих в схеме факторного анализа, и

$$p \geqslant \frac{1}{2}(p - m)(p - m + 1),$$

то эти дисперсии можно выбрать таким образом, чтобы число требуемых факторов не превышало m.

43.13 Рассматривая факторный анализ при $p = 2$ и $m = 1$, выписать функцию правдоподобия и, используя дифференцирование, показать, что уравнение МП имеет следующий вид:

$$-c_{12}l_2 + \left(c_{11} - \frac{c_{11}\sigma_1}{l_1}\right)l_2 = 0,$$

$$c_{22}\left(1 - \frac{\sigma_2}{l_2}\right) - c_{12}l_2 = 0.$$

Вывести отсюда, что

$$c_{22}l_1^2 = c_{11}l_2^2,$$

$$\sigma_2/l_2 = \sigma_1/l_1.$$

Такой результат недопустим при свободном оценивании четырех параметров. Объяснить его появление.

43.14 Убедиться, что определитель (43.74) действительно равен (43.75).

ГЛАВА 44

ДИСКРИМИНАЦИЯ И КЛАССИФИКАЦИЯ

44.1 В этой главе мы коснемся задач дифференциации двух и более многомерных генеральных совокупностей по результатам измерений. Существуют три различных класса задач, которые часто смешивают:

(а) Дискриминация. Известно о существовании двух или более генеральных совокупностей, и дано по одной выборке из каждой совокупности. Задача заключается в выработке основанного на имеющихся выборках правила, позволяющего приписать некоторый новый элемент к правильной генеральной совокупности в том случае, когда нам заведомо неизвестно, к какой совокупности он принадлежит.

(б) Классификация. Рассматривается некоторая выборка или вся генеральная совокупность. Задача состоит в классификации элементов выборки или совокупности по группам, столь различным между собой, сколь это возможно. При дискриминации существование групп дано, при классификации выявление их и составляет задачу.

(в) Разбиение. Рассматривается некоторая выборка или вся генеральная совокупность. Мы хотим разбить их на группы независимо от того, естественны ли границы разбиения или нет.

Обратимся к примерам. Допустим, что мы располагаем информацией о некотором количестве индивидуумов из двух различных рас. Нам может понадобиться функция, позволяющая приписывать новые индивидуумы к соответствующим им расам. Построение такой функции на основе уже имеющейся информации и составляет задачу дискриминации. Другой пример: дана совокупность элементов неизвестного происхождения и требуется выяснить, возможно ли разбиение этих элементов на группы, естественное в том смысле, что элементы каждой группы сходны между собой, но элементы одной группы значительно отличаются от элементов другой. Это задача классификации. Наконец, допустим, что мы хотим разбить на группы некоторое количество студентов в зависимости от их успехов на экзаменах, понимая при этом, что границы такого разбиения могут быть весьма условны. Это задача разбиения, и она

возникает и тогда, когда рассматриваемая совокупность однородна.

В этой главе задачу разбиения мы опустим, а обсудим проблемы дискриминации и классификации.

Дискриминация

44.2 Прежде чем перейти к теории вопроса, полезно обсудить, насколько проблема в том виде, в каком мы ее поставили, имеет практический смысл. Нам дано множество элементов, и принадлежность каждого совокупности A или совокупности B точно задана. Если мы обладаем такой информацией о каждом элементе группы, почему мы не можем потребовать ее от новых элементов, которые мы можем встретить? Существуют по крайней мере четыре типа ситуаций, содержащих ответ на этот вопрос.

(а) Потерянная информация. Нам нужно уметь правильно определять пол по набору человеческих костей, найденных на месте археологических раскопок. Если бы обладатель этих костей был жив, не было бы проблемы, но, к сожалению, существеннейшая информация распалась в прах.

(б) Недостижимая информация. Больничные записи обеспечивают нас информацией о внешних симптомах и внутреннем течении болезней. Мы же должны уметь распознавать заболевание по внешним признакам без госпитализации: ведь основная цель может состоять в умении давать диагноз и лечить на ранних стадиях, избегая внутренних обследований.

(в) Предсказание. Основываясь на прошлом опыте, можно научиться различать типы поведения, например экономических систем, в настоящем. Тогда, основываясь на наблюдениях настоящего момента, можно пытаться предсказывать поведение систем в будущем.

(г) Исследование, ведущее к разрушению. В случае, когда исследование объекта влечет его полное разрушение, желательно найти «дискриминатор» неразрушительного характера, позволяющий предсказывать результат исследования.

44.3 Следует отметить, что в первой ситуации проверка правильности наших действий исключена. Мы, как говорится, облечены правом проводить дискриминацию, даже рискуя ошибиться. Приписывая элемент к тому или иному классу, мы можем совершать ошибки двух типов в зависимости от того, к какой именно совокупности мы ошибочно отнесли данный элемент. На первых порах мы будем предполагать эти два типа ошибок равнозначимыми.

Рассмотрим теперь p-мерное пространство W, в котором каждый член выборки представлен точкой, чьи координаты будем обозначать x. Две совокупности можно представить как две грозди точек (или как две непрерывные плотности), которые как-то разделены (иначе мы не могли бы различать их лишь по значениям x), но в то же время имеют и общее пересечение (иначе не было бы проблемы дискриминации). Мы хотим построить границу в пространстве так, чтобы как можно больше точек 1-й совокупности лежало по одну сторону границы и как можно больше точек 2-й совокупности — по другую. Мы потребуем также, чтобы граница имела достаточно простую форму. Пусть f_1 и f_2 — соответствующие плотности вероятностей. Потребуем, чтобы наша граница определяла область R такую, что

$$\int_R f_2\,dx = \int_{W-R} f_1\,dx = 1 - \int_R f_1\,dx, \tag{44.1}$$

или, что эквивалентно,

$$\int_R (f_1 + f_2)\,dx = 1. \tag{44.2}$$

Фактически это означает, что вероятности ошибок двух типов равны. Теперь мы хотим минимизировать вероятность ложной классификации, что эквивалентно минимизации вероятности ошибки одного из типов, скажем, минимизации

$$\int_R f_2\,dx. \tag{44.3}$$

Задача состоит тогда в нахождении безусловного минимума

$$\int_R \{f_2 - \lambda(f_1 + f_2)\}\,dx, \tag{44.4}$$

или, что эквивалентно,

$$\int_R (\beta f_2 - f_1)\,dx, \tag{44.5}$$

где постоянные λ или β определяются условием (44.2). Очевидно, что мы достигнем цели, если к области R отнесем те и только те точки, для которых $\beta f_1 - f_2 < 0$. Граница R таким образом дается формулой

$$f_1/f_2 = \beta, \tag{44.6}$$

т. е. определяется отношением правдоподобия.

44.4 Это безусловно разумный критерий. Мы относим элемент к той или иной совокупности, отталкиваясь от «меры

близости» последних; правда, под «близостью» мы понимаем не расстояние в какой-либо метрике, а близость распределений. Вероятность ложной дискриминации для обоих типов ошибок равна

$$\int\limits_{f_1/f_2>\beta} f_2\,dx = \int\limits_{f_2/f_1>1/\beta} f_1\,dx. \qquad (44.7)$$

Фактически мы передоказали лемму Неймана — Пирсона из 22.10 т. 2.

44.5 Предположим теперь, что две многомерные генеральные совокупности нормальны со средними μ_1 и μ_2 и одной и той же матрицей рассеяния γ. Тогда логарифм отношения правдоподобия равен

$$-\frac{1}{2}\sum_{j,\,k}\Gamma_{jk}\{(x_j-\mu_{1j})(x_k-\mu_{1k})-(x_j-\mu_{2j})(x_k-\mu_{2k})\}=$$

$$=\sum_{j,\,k}\Gamma_{jk}(\mu_{1j}-\mu_{2j})x_k-\frac{1}{2}\sum_{j,\,k}\Gamma_{jk}(\mu_{1j}\mu_{1k}-\mu_{2j}\mu_{2k}), \qquad (44.8)$$

где Γ — матрица, обратная γ. Вторая часть последнего выражения есть величина постоянная и, не ограничивая общности, нашу границу мы можем определить равенством

$$\sum_{j,\,k}\Gamma_{jk}(\mu_{1j}-\mu_{2j})x_k=\text{const}. \qquad (44.9)$$

Это форма, основанная на теоретическом распределении. Если же нам даны выборки со средними $\bar{x}_1$, $\bar{x}_2$ и «объединенной» матрицей рассеяния c_{jk}, выборочная функция, определяющая границу, дается следующей формулой:

$$\sum C_{jk}(\bar{x}_{1j}-\bar{x}_{2j})x_k=\text{const}. \qquad (44.10)$$

44.6 Этот же результат можно получить и другим путем. Определим линейную функцию

$$X=\sum_{j=1}^{p} l_j x_j \qquad (44.11)$$

так, чтобы отношение межклассовой и внутриклассовой дисперсий, иначе, величина

$$\{\sum l_j(\mu_{1j}-\mu_{2j})\}^2/\{\sum l_j l_k \gamma_{jk}\}, \qquad (44.12)$$

была максимальной.

Дифференцируя по l_j, получаем

$$\mu_{1j}-\mu_{2j}=\frac{\left\{\sum\limits_{j=1}^{p} l_j(\mu_{1j}-\mu_{2j})\right\}\sum\limits_k \gamma_{jk}l_k}{2\sum l_j l_k \gamma_{jk}},$$

откуда имеем

$$l_k\sim\sum_j\Gamma_{jk}(\mu_{1j}-\mu_{2j}). \qquad (44.13)$$

Последнее снова приводит к (44.9). Так как наша функция X используется лишь для разделения двух совокупностей, а не измерения расстояний между ними, мы можем умножать ее на любую удобную нам постоянную.

44.7 Как можно ожидать из соображений симметрии, «дискриминирующая» гиперплоскость (44.9) делит пополам отрезок, соединяющий два «центра» с координатами μ_1 и μ_2. Чтобы убедиться в этом, линейным преобразованием сведем наши величины к независимым и имеющим единичные дисперсии. Поскольку у f_1 и f_2 одна матрица рассеяния, их к требуемой форме приводит одно и то же преобразование. Дискриминирующая гиперплоскость дается тогда равенством

$$\sum(\mu_{1j}-\mu_{2j})x_j=\text{const}$$

и перпендикулярна соединяющему центры отрезку, чьи направляющие косинусы пропорциональны косинусам вектора $\mu_1-\mu_2$.

Функции f_1 и f_2 одновременно становятся сферически-симметричными, и без потери общности одну из координатных осей мы можем направить вдоль линии, проходящей через средние. Интегралы, дающие вероятности ошибочной классификации, сводятся к одномерным нормальным интегралам и очевидным образом равны друг другу, если дискриминационная граница делит отрезок, соединяющий средние, пополам.

Это определяет постоянную в (44.10). Если $\bar{X}_1$—среднее левой части, порожденное f_1, а $\bar{X}_2$ — соответствует f_2, то постоянная равна $(\bar{X}_1+\bar{X}_2)/2$. Не ограничивая общности, впредь положим $\bar{X}_1\geqslant\bar{X}_2$.

Пример 44.1 (Фишер, 1936)

В таблице 44.1 приведены результаты четырех типов измерений (в сантиметрах) над 50 цветками от каждой из трех разновидностей ириса, а именно ириса щетинистого (Iris Setosa, обозначим Se), ириса разноцветного (Iris versicolor, обозначим Ve) и ириса вирджиника (Iris virginica, обозначим Vi). Измерялись величины: x_1 — длина чашелистика; x_2 — ширина

Таблица 44.1

Многомерные измерения в задачах таксономии

Ирис щетинистый				Ирис разноцветный				Ирис вирджиника			
длина чашелистика	ширина чашелистика	длина лепестка	ширина лепестка	длина чашелистика	ширина чашелистика	длина лепестка	ширина лепестка	длина чашелистика	ширина чашелистика	длина лепестка	ширина лепестка
5,1	3,5	1,4	0,2	7,0	3,2	4,7	1,4	6,3	3,3	6,0	2,5
4,9	3,0	1,4	0,2	6,4	3,2	4,5	1,5	5,8	2,7	5,1	1,9
4,7	3,2	1,3	0,2	6,9	3,1	4,9	1,5	7,1	3,0	5,9	2,1
4,6	3,1	1,5	0,2	5,5	2,3	4,0	1,3	6,3	2,9	5,6	1,8
5,0	3,6	1,4	0,2	6,5	2,8	4,6	1,5	6,5	3,0	5,8	2,2
5,4	3,9	1,7	0,4	5,7	2,8	4,5	1,3	7,6	3,0	6,6	2,1
4,6	3,4	1,4	0,3	6,3	3,3	4,7	1,6	4,9	2,5	4,5	1,7
5,0	3,4	1,5	0,2	4,9	2,4	3,3	1,0	7,3	2,9	6,3	1,8
4,4	2,9	1,4	0,2	6,6	2,9	4,6	1,3	6,7	2,5	5,8	1,8
4,9	3,1	1,5	0,1	5,2	2,7	3,9	1,4	7,2	3,6	6,1	2,5
5,4	3,7	1,5	0,2	5,0	2,0	3,5	1,0	6,5	3,2	5,1	2,0
4,8	3,4	1,6	0,2	5,9	3,0	4,2	1,5	6,4	2,7	5,3	1,9
4,8	3,0	1,4	0,1	6,0	2,2	4,0	1,0	6,8	3,0	5,5	2,1
4,3	3,0	1,1	0,1	6,1	2,9	4,7	1,4	5,7	2,5	5,0	2,0
5,8	4,0	1,2	0,2	5,6	2,9	3,6	1,3	5,8	2,8	5,1	2,4
5,7	4,4	1,5	0,4	6,7	3,1	4,4	1,4	6,4	3,2	5,3	2,3
5,4	3,9	1,3	0,4	5,6	3,0	4,5	1,5	6,5	3,0	5,5	1,8
5,1	3,5	1,4	0,3	5,8	2,7	4,1	1,0	7,7	3,8	6,7	2,2
5,7	3,8	1,7	0,3	6,2	2,2	4,5	1,5	7,7	2,6	6,9	2,3
5,1	3,8	1,5	0,3	5,6	2,5	3,9	1,1	6,0	2,2	5,0	1,5
5,4	3,4	1,7	0,2	5,9	3,2	4,8	1,8	6,9	3,2	5,7	2,3
5,1	3,7	1,5	0,4	6,1	2,8	4,0	1,3	5,6	2,8	4,9	2,0
4,6	3,6	1,0	0,2	6,3	2,5	4,9	1,5	7,7	2,8	6,7	2,0
5,1	3,3	1,7	0,5	6,1	2,8	4,7	1,2	6,3	2,7	4,9	1,8
4,8	3,4	1,9	0,2	6,4	2,9	4,3	1,3	6,7	3,3	5,7	2,1
5,0	3,0	1,6	0,2	6,6	3,0	4,4	1,4	7,2	3,2	6,0	1,8
5,0	3,4	1,6	0,4	6,8	2,8	4,8	1,4	6,2	2,8	4,8	1,8
5,2	3,5	1,5	0,2	6,7	3,0	5,0	1,7	6,1	3,0	4,9	1,8
5,2	3,4	1,4	0,2	6,0	2,9	4,5	1,5	6,4	2,8	5,6	2,1
4,7	3,2	1,6	0,2	5,7	2,6	3,5	1,0	7,2	3,0	5,8	1,6
4,8	3,1	1,6	0,2	5,5	2,4	3,8	1,1	7,4	2,8	6,1	1,9
5,4	3,4	1,5	0,4	5,5	2,4	3,7	1,0	7,9	3,8	6,4	2,0
5,2	4,1	1,5	0,1	5,8	2,7	3,9	1,2	6,4	2,8	5,6	2,2
5,5	4,2	1,4	0,2	6,0	2,7	5,1	1,6	6,3	2,8	5,1	1,5
4,9	3,1	1,5	0,2	5,4	3,0	4,5	1,5	6,1	2,6	5,6	1,4
5,0	3,2	1,2	0,2	6,0	3,4	4,5	1,6	7,7	3,0	6,1	2,3
5,5	3,5	1,3	0,2	6,7	3,1	4,7	1,5	6,3	3,4	5,6	2,4
4,9	3,6	1,4	0,1	6,3	2,3	4,4	1,3	6,4	3,1	5,5	1,8
4,4	3,0	1,3	0,2	5,6	3,0	4,1	1,3	6,0	3,0	4,8	1,8
5,1	3,4	1,5	0,2	5,5	2,5	4,0	1,3	6,9	3,1	5,4	2,1
5,0	3,5	1,3	0,3	5,5	2,6	4,4	1,2	6,7	3,1	5,6	2,4
4,5	2,3	1,3	0,3	6,1	3,0	4,6	1,4	6,9	3,1	5,1	2,3
4,4	3,2	1,3	0,2	5,8	2,6	4,0	1,2	5,8	2,7	5,1	1,9
5,0	3,5	1,6	0,6	5,0	2,3	3,3	1,0	6,8	3,2	5,9	2,3

Продолжение табл. 44.1

Ирис щетинистый				Ирис разноцветный				Ирис вирджиника			
длина чашелистика	ширина чашелистика	длина лепестка	ширина лепестка	длина чашелистика	ширина чашелистика	длина лепестка	ширина лепестка	длина чашелистика	ширина чашелистика	длина лепестка	ширина лепестка
5,1	3,8	1,9	0,4	5,6	2,7	4,2	1,3	6,7	3,3	5,7	2,5
4,8	3,0	1,4	0,3	5,7	3,0	4,2	1,2	6,7	3,0	5,2	2,3
5,1	3,8	1,6	0,2	5,7	2,9	4,2	1,3	6,3	2,5	5,0	1,9
4,6	3,2	1,4	0,2	6,2	2,9	4,3	1,3	6,5	3,0	5,2	2,0
5,3	3,7	1,5	0,2	5,1	2,5	3,0	1,1	6,2	3,4	5,4	2,3
5,0	3,3	1,4	0,2	5,7	2,8	4,1	1,3	5,9	3,0	5,1	1,8

чашелистика, x_3 — длина лепестка, x_4 — ширина лепестка. Средние (в *см*) таковы:

Величины	Ve	Se	Разность
x_1	5,936	5,006	0,930
x_2	2,770	3,428	—0,658
x_3	4,260	1,462	2,798
x_4	1,326	0,246	1,080

(44.14)

Суммы квадратов и произведений отклонений от средних (в $см^2$) для объединенной выборки таковы:

	x_1	x_2	x_3	x_4
x_1	19,1434	9,0356	9,7634	3,2394
x_2		11,8658	4,6232	2,4746
x_3			12,2978	3,8794
x_4				2,4604

(44.15)

Обратная матрица (в $см^{-2}$) такова:

	x_1	x_2	x_3	x_4
x_1	0,1187161	—0,0668666	—0,0816158	0,0396350
x_2		0,1452736	0,0334101	—0,1107529
x_3			0,2193614	—0,2720206
x_4				0,8945506

(44.16)

В такого рода задачах иногда может возникнуть вопрос о степенях свободы, но поскольку цель дискриминантной функции — разделение, коэффициенты могут быть определены с точностью до произвольной постоянной. В качестве числа степеней свободы в (44.15) мы возьмем общий объем выборок — 100. Тогда, для того чтобы получить обращение матрицы рассеяния, мы должны величины в (44.16) умножить на 100. Используя (44.10), находим коэффициенты:

$$J_1 = (118{,}7161)(0{,}930) - (66{,}8666)(-0{,}658) - (81{,}6158)(2{,}798) + \\ + (39{,}6350)(1{,}080) = -3{,}11511,$$

$$J_2 = -18{,}390\,75, \quad l_3 = 22{,}210\,44, \quad l_4 = 31{,}473\,74. \tag{44.17}$$

Мы можем умножить эти коэффициенты на любую удобную нам постоянную. Потребовав, чтобы l_1 равнялось, например, единице, получаем

$$X = x_1 + 5{,}9037x_2 - 7{,}1299x_3 - 10{,}1036x_4. \tag{44.18}$$

Среднее значение X для ириса разноцветного, полученное путем подстановки средних в (44.18), равно 66,917. Тоже для ириса щетинистого равно — 38,424. Среднее этих двух значений равно 14,247. Таким образом, если значение X для какого-нибудь цветка будет превышать 14,247, мы отнесем последний к ирисам разноцветным, в противном случае — к ирисам щетинистым.

44.8 Мы можем приближенно вычислить вероятность ложной классификации. Имеем

$$\mathsf{D}X = \sum l_j l_k \gamma_{jk} = \sum l_j \gamma_{jk} \Gamma_{km} (\mu_{1m} - \mu_{2m}) = \sum l_j (\mu_{1j} - \mu_{2j}).$$

Используя (44.11), последнее оценим так:

$$\hat{\mathsf{D}}X = \bar{X}_1 - \bar{X}_2. \tag{44.19}$$

Это оценка дисперсии величины X. Если «критическое» значение X равно $(\bar{X}_1 + \bar{X}_2)/2$, вероятность ложной классификации так или иначе равна (приближенно) вероятности отклонения за уровень $\bar{X}_1 - (\bar{X}_1 + \bar{X}_2)/2 = (\bar{X}_1 - \bar{X}_2)/2$ нормальной случайной величины с нулевым средним и дисперсией $(\bar{X}_1 - \bar{X}_2)$.

Пример 44.2

Для данных из примера 44.1

$$\bar{X}_1 = 66{,}917, \quad \bar{X}_2 = -38{,}424.$$

Следовательно, вероятность ложной классификации равна вероятности превышения значения $(66{,}917 + 318{,}442) : 2 = 52{,}67$ нормальной величиной с нулевым средним и дисперсией 105,341 (или среднеквадратичным отклонением $\sqrt{105{,}341} = 10{,}26$). Эта вероятность ничтожно мала.

44.9 Интересен специальный случай, когда все корреляции между x одинаковы. В примерах из биологии не так уж редко встречаются более или менее равные корреляции. Если это так, мы можем провести дискриминацию по двум факторам. Сделать это можно следующим образом.

Как и в примере 43.1, можно показать, что если все корреляции равны ρ, то собственные числа корреляционной матрицы *)

$$\lambda_1 = 1 + (p-1)\rho. \tag{44.20}$$

$$\lambda_2 = \ldots = \lambda_p = 1 - \rho. \tag{44.21}$$

Мы имеем, таким образом, одно основное направление, а среди оставшихся наблюдаем изотропность. Компонента, соответствующая λ_1, равна

$$\zeta_1 = \frac{1}{\sqrt{p}} \sum_{j=1}^{p} x_j. \tag{44.22}$$

Рассмотрим пропорциональную ζ_1 величину, которую будем называть компонентой «объема» —

$$Q = \sum x_j = \sqrt{p}\,\zeta_1, \tag{44.23}$$

так что

$$\mathsf{D}Q = p\lambda_1 = p\{1 + (p-1)\rho\}. \tag{44.24}$$

Среди оставшихся компонент ни одна не имеет преимущества перед остальными. Выберем тогда некоторые веса w_j с ненулевым средним и определим величину

$$P = \sum_{j=1}^{p} \frac{w_j - \bar{w}}{\bar{w}} x_j, \tag{44.25}$$

которую назовем компонентой «формы». Получаем, что

$$\mathsf{D}P = \sum \left(\frac{w_j - \bar{w}}{\bar{w}}\right)^2 (1 - \rho). \tag{44.26}$$

*) Следует обратить внимание читателя на тот факт, что в некоторых дальнейших рассуждениях автор не проводит четкого различия между выборочными и точными числовыми характеристиками распределений (средние, дисперсии, ковариации, собственные числа и т. д.). (*Прим. ред.*)

Далее

$$\operatorname{cov}(Q, P) = \operatorname{cov}\left(\sum \frac{w_j - \bar{w}}{\bar{w}} x_j, \quad \sum x_j\right) =$$

$$= \sum_j \frac{w_j - \bar{w}}{\bar{w}} \mathrm{D}x_j + \sum_{j \neq k}\sum \frac{w_j - \bar{w}}{\bar{w}} \operatorname{cov}(x_j, x_k) =$$

$$= \{1 + (p-1)\rho\} \sum_j \frac{w_j - \bar{w}}{\bar{w}} = 0. \qquad (44.27)$$

Итак, компонента «объема» и компонента «формы» не коррелированы.

44.10 Чтобы получить «дискриминатор», положим

$$w_j = \bar{x}_{1j} - \bar{x}_{2j} \qquad (44.28)$$

и будем искать дискриминатор в форме

$$X = \alpha Q + P. \qquad (44.29)$$

Будем максимизировать $(X_1 - X_2)^2/\mathrm{D}X$. Пусть $\tilde{D}_P = P_1 - P_2$, $\tilde{D}_Q = Q_1 - Q_2$ *). Тогда мы должны максимизировать

$$\frac{(\alpha\tilde{D}_Q + \tilde{D}_P)^2}{\alpha^2 \mathrm{D}Q + 2\alpha \operatorname{cov}(Q, P) + \mathrm{D}P}. \qquad (44.30)$$

Используя (44.27), легко находим решение:

$$\alpha = \frac{\tilde{D}_Q \mathrm{D}P}{\tilde{D}_P \mathrm{D}Q}. \qquad (44.31)$$

Подставляя теперь веса из (44.28) в выражения для $\tilde{D}_P$ и $\tilde{D}_Q$, получаем

$$\tilde{D}_P = \sum \frac{(\bar{x}_{1j} - \bar{x}_{2j})^2}{\bar{x}_1 - \bar{x}_2} - (\bar{x}_1 - \bar{x}_2)p, \qquad (44.32)$$

$$\mathrm{D}P = (1-\rho)\left[\sum \frac{(\bar{x}_{1j} - \bar{x}_{2j})^2}{(\bar{x}_1 - \bar{x}_2)^2} - p\right], \qquad (44.33)$$

$$\tilde{D}_Q = p(\bar{x}_1 - \bar{x}_2), \qquad (44.34)$$

$$\mathrm{D}Q = p\{1 + (p-1)\rho\}. \qquad (44.35)$$

Подстановка в (44.31) дает α, и мы получаем, что

$$X = \frac{1-\rho}{1 + (p-1)\rho} Q + P. \qquad (44.36)$$

*) Случайные величины Q_i, P_i и X_i, $i = 1, 2$, следует понимать как статистики, полученные соответственно по формулам (44.23), (44.25) и (44.29) с заменой x_j их средними выборочными значениями $\bar{x}_j$. (*Прим. ред.*)

Пример 44.3

Рассмотрим снова данные для ирисов в примерах 44.1 и 44.2. Корреляционная матрица такова:

	x_1	x_2	x_3	x_4
x_1	1	0,599513	0,636323	0,472011
x_2		1	0,382719	0,457988
x_3			1	0,705258
x_4				1

(44.37)

Корреляции достаточно близки друг к другу, чтобы использовать предыдущие результаты для приближенного исследования. Приводя величины к величинам, имеющим нулевые средние и единичные дисперсии, получаем

	Ve	S	Ve − S
x_1	1,0628	−1,0628	2,1256
x_2	−0,4551	0,4551	−1,9102
x_3	3,9894	−3,9894	7,9788
x_4	3,4426	−3,4426	6,8852
$\sum = Q$	7,5347	−7,5397	$15{,}0794 = \tilde{D}_Q$

(44.38)

Дисперсия Q вычисляется как сумма 16 элементов (44.37) и равна

$$\mathrm{D}Q = 10{,}5076. \qquad (44.39)$$

Веса для компоненты «формы» получаем из (44.38), деля последний столбец на 15,0794/4 и вычитая единицу. Имеем

$$-0{,}4362; \ -1{,}5067; \ 1{,}1165; \ 0{,}8264. \qquad (44.40)$$

Для оценки P, таким образом, получаем

$$P = (-0{,}4362 \times 1{,}0628) + \ldots = 8{,}2747. \qquad (44.41)$$

Снова используя (44.38), имеем

$$\mathrm{D}P = 3{,}0912. \qquad (44.42)$$

Из (44.38) получаем, что ковариация между Q и P равна 0,36162 (а не нулю). Подстановка в (44.31) дает, что $\alpha = 0{,}2412$,

и наш дискриминатор равен

$$X = 0{,}2412Q + P, \tag{44.43}$$

где, что нужно помнить, Q есть сумма всех стандартно нормированных x, а P представляет взвешенную сумму с весами из (44.40).

Квадратичные дискриминаторы

44.11 Линейный дискриминатор (44.10) построен в предположении, что две сравниваемые генеральные совокупности имеют одинаковые характеристики рассеяния. Если это не так, наш логарифм правдоподобия с точностью до постоянной становится в обычных обозначениях равным

$$\sum \Gamma_{1jk}(x_{1j} - \mu_{1j})(x_{1k} - \mu_{1k}) + \sum \Gamma_{2jk}(x_{2j} - \mu_{2j})(x_{2k} - \mu_{2k}). \tag{44.44}$$

Члены, имеющие второй порядок по x, уже не уничтожаются, и наша граница становится поверхностью второго порядка в p-мерном пространстве. В общем случае эта конструкция слишком громоздка, чтобы иметь с ней дело. Последнее обстоятельство, по-видимому, и объясняет тот факт, что квадратичные дискриминаторы редко встречаются в практике.

Мы, однако, можем преуспеть, если ограничимся рассмотрением частного случая с определенной «формой» и «размером». Пусть $p = 2$, а ковариационные члены равны нулю. Выражение (44.44) сводится тогда к форме типа

$$(y - \nu_1)^2 + (x - \nu_2)^2, \tag{44.45}$$

где y и x — линейные функции от P и Q, а дисперсии поддаются вычислению. Дискриминантная граница становится здесь эллипсом (двумерным), и ситуация поддается изучению. Один пример можно найти у С. Смита (1947).

С. Рао (1966) рассмотрел дискриминацию при наличии сложных гипотез.

Статистическое исследование дискриминантной функции

44.12 Процесс статистического исследования дискриминатора нуждается в небольшом пояснении. Мы можем, например, заподозрить, что генеральные совокупности хотя и различны, но тем не менее столь близки, что дискриминатор перестает быть эффективным. Это проявляется в том, что вероятности ложной классификации, являясь минимальными, все-таки остаются большими. С другой стороны, мы можем также предположить, что существует большее, чем мы наблюдаем, различие между совокупностями, но наш объем выборки недостаточно велик для того, чтобы дискриминатор оказался надежным; фактически здесь требуется построение доверительных интервалов для функции или ее коэффициентов. Или может возникнуть опасение, что на самом деле генеральные совокупности идентичны и дискриминация иллюзорна.

Исследование дискриминантных функций обычно проводится именно в связи с указанными возможностями. Вообще, это не столько исследование дискриминантных функций, сколько проверка однородности с использованием этих функций. Если же неоднородность обнаружена, функция ipso facto *) значима в том смысле, что она различает реальные отличия оптимальным образом (за исключением того, что мы используем оценки средних и характеристик рассеяния, вместо их неизвестных теоретических значений). Правда, путь этот может быть и не слишком хорош, даже будучи наилучшим из всех доступных.

44.13 Допустим, что в действительности две наши генеральные совокупности имеют одинаковые средние. Тогда определяемое дискриминатором различие между выборочными средними характеризуется величиной

$$U = \sum C_{jk}(\bar{x}_{1j} - \bar{x}_{2j})(\bar{x}_{1k} - \bar{x}_{2k}). \tag{44.46}$$

Выборочные средние распределены нормально, так что, если объемы выборок одинаковы, разность выборочных средних также распределена нормально с нулевым средним и дисперсией, равной удвоенной дисперсии выборочного среднего. Отсюда следует (см. **41.17**), что величина U распределена как основанная на $2n$ наблюдениях статистика Хотеллинга T^2, деленная на $2n - 1$. Вследствие (41.84) это эквивалентно распределению множественной корреляции R^2 при $\boldsymbol{R}^2 = 0$, и критерий может быть построен с помощью дисперсионного анализа. Однако, по-видимому, предпочтительнее проверять однородность методом главы 42, который позволяет рассматривать отдельно различия между средними и между характеристиками рассеяния.

44.14 Поскольку в **27.28—9** мы получали «нулевое» распределение R^2, не предполагая многомерной нормальности, не должно вызывать удивления и то, что дискриминация между двумя генеральными совокупностями также не требует такого предположения. Упражнение 44.10 показывает, как границу (44.9) можно получить методом НК.

44.15 Обсудим теперь те случаи, когда ошибки двух типов не равнозначимы по важности. Существуют два типа ситуаций, когда требуется модификация предыдущих результатов:

(а) Возможны ситуации, когда исследователю a priori известно, что у него больше шансов выбрать элемент из одной

*) ipso facto (лат.) — в силу самого факта.

совокупности, чем из другой. Например, при случайном отборе индивидуумов на предмет обнаружения у них активного туберкулеза мы ожидаем встретить гораздо больше здоровых людей, чем больных.

(б) Последствия ложной классификации могут существенно отличаться друг от друга. Например, гораздо менее опасно признать здорового человека больным (так как ошибка скорее всего обнаружится прежде, чем будет нанесен существенный вред), чем признать больного здоровым (так как здесь последствия будут иметь обратный характер).

Предположим, что вероятности принадлежности двум нашим генеральным совокупностям равны соответственно π_1 и $\pi_2(=1-\pi_1)$. Предположим, что мы в состоянии приписать ошибкам некие числа, характеризующие наши потери при ложной классификации; пусть, например, ошибки первого и второго рода стоят нам соответственно c_1 и c_2 (в некоторых единицах измерения). Тогда вместо того, чтобы минимизировать количество ошибок при многократной дискриминации, будем минимизировать наши потери, иначе вместо (44.3) будем минимизировать

$$c_2 \int\limits_R \pi_2 f_2\, dx + c_1 \int\limits_{1-R} \pi_1 f_1\, dx = c_1\pi_1 + \int\limits_R (c_2\pi_2 f_2 - c_1\pi_1 f_1)\, dx. \quad (44.47)$$

Последнее выражение достигает минимума, если область R определяется соотношением

$$\frac{c_2\pi_2 f_2}{c_1\pi_1 f_1} < 1. \quad (44.48)$$

Таким образом, если в качестве дискриминатора выбрать $\log(f_2/f_1)$, эффект введения априорных вероятностей π и постоянных потерь c проявляется лишь в добавлении постоянной к дискриминантной функции, или, что эквивалентно, в замене критического значения последней величиной

$$\log\{(c_2\pi_2)/(c_1\pi_1)\}.$$

Случай k генеральных совокупностей

44.16 Существенно новые вопросы возникают при переходе от дискриминации между двумя совокупностями к дискриминации между многими совокупностями. Как и прежде, мы будем стараться разбить выборочное пространство на непересекающиеся области (по одной для каждой совокупности) и затем приписывать новый элемент той совокупности, в чью область он попал. Но границы областей здесь уже больше не определяются

какой-то одной функцией. Чтобы достигнуть оптимальных свойств, мы должны пользоваться несколькими функциями; если же мы вынуждены пользоваться одной функцией, мы должны смириться с потерей мощности дискриминации.

44.17 Чтобы проиллюстрировать положение, достаточно рассмотреть три генеральные совокупности — обобщение на случай произвольного числа тривиально. Предположим, что вероятности появления трех совокупностей с плотностями f_1, f_2, f_3 равны соответственно π_1, π_2, π_3 $(\pi_1+\pi_2+\pi_3=1)$. Полученное С. Рао обобщение леммы Неймана — Пирсона утверждает, что ошибки ложной классификации минимальны, если области определяются включающими отношения вероятностей соотношениями, представляющими простое обобщение (44.48). А именно, R_1 таково, что $\pi_1 f_1$ одновременно больше или равно $\pi_2 f_2$, $\pi_3 f_3$; R_2 таково, что $\pi_2 f_2 \geqslant \pi_3 f_3$, $\pi_1 f_1$; R_3 таково, что $\pi_3 f_3 \geqslant \pi_1 f_1$, $\pi_2 f_2$.

44.18 В частности, если три совокупности нормальны с общей матрицей рассеяния (γ_{jk}) и средними μ_{1j}, μ_{2j}, μ_{3j}, аналогично 44.5, существуют такие β_2 и β_3, что R_1 определяются неравенствами

$$\sum \Gamma_{jk}(\mu_{1j}-\mu_{2j})x_k \geqslant \beta_2, \quad (44.49)$$

$$\sum \Gamma_{jk}(\mu_{1j}-\mu_{3j})x_k \geqslant \beta_3. \quad (44.50)$$

Так же определяются и остальные области. В выборочном пространстве R_1 определяется как область, лежащая между двумя гиперплоскостями (см. 44.49), (44.50)) и содержащая среднее первой совокупности и т. д. Поверхности постоянства значений взвешенного отношения вероятностей для совокупностей 1 и 2 являются поверхностями постоянства функции

$$\log\{(\pi_1 f_1)/(\pi_2 f_2)\} =$$
$$= \sum \Gamma_{jk}(\mu_{1j}-\mu_{2j})x_k - \frac{1}{2}\sum \Gamma_{jk}(\mu_{1j}\mu_{2k}-\mu_{2j}\mu_{1k}) + \log\frac{\pi_1}{\pi_2}. \quad (44.51)$$

В том частном случае, когда все π одинаковы, мы можем сравнивать три функции:

$$X_1 = \sum \Gamma_{jk}\mu_{1j}x_k - \frac{1}{2}\sum \Gamma_{jk}\mu_{1j}\mu_{1k}, \quad (44.52)$$

$$X_2 = \sum \Gamma_{jk}\mu_{2j}x_k - \frac{1}{2}\sum \Gamma_{jk}\mu_{2j}\mu_{2k}, \quad (44.53)$$

$$X_3 = \sum \Gamma_{jk}\mu_{3j}x_k - \frac{1}{2}\sum \Gamma_{jk}\mu_{3j}\mu_{3k}, \quad (44.54)$$

и приписывать элемент к R_1, R_2 или R_3 соответственно тому, какой из X окажется наибольшим, если в формулы проставить

выборочные значения. Так, если, скажем, X_1 больше X_2 и X_3, из (44.51) следует, что $f_1 > f_2$ и $f_1 > f_3$. Как обычно, чтобы получить приближенный дискриминатор, можно вместо неизвестных параметров использовать их выборочные значения.

Пример 44.4 (С. Рао и Слейтер, 1949)

Ниже приведены осредненные метки трех тестов, которым был подвергнут ряд лиц, разбитых на группы в зависимости от их невротического состояния.

Группа	Размер выборки	Осредненные метки		
		1	2	3
Возбужденное состояние	114	2,9298	1,1667	0,7281
Истерия	33	3,0303	1,2427	0,5455
Психопатия	32	3,8125	1,8438	0,8125
Маниакальный психоз	17	4,7059	1,5882	1,1176
Изменение личности	5	1,40000	0,20000	0,0000
Нормальное состояние	55	0,6000	0,1455	0,2182
	256			

(44.55)

Внутригрупповая матрица рассеяния (250 ст. св.) такова:

	1	2	3
1	2,300851	0,251578	0,474169
2		0,607466	0,035774
3			0,595094

(44.56)

Матрица, обратная ей, такова:

	1	2	3
1	0,543234	−0,200195	−0,420813
2		1,725807	0,055767
3			2,012357

(44.57)

Для целей настоящего примера достаточно предположить, что все π одинаковы. Шесть дискриминантных функций типа (44.49) тогда определятся с помощью следующей таблицы.

	Коэффициенты			Постоянная
	x_1	x_2	x_3	
Нормальное состояние	0,2050	0,1431	0,1947	−0,0931
Изменение личности	0,7204	0,0649	−0,5780	−0,5107
Возбужденное состояние	1,0515	1,4676	0,2974	−2,5047
Истерия	1,1678	1,5679	−0,1081	−2,7139
Психопатия	1,3599	2,4641	0,1336	−4,9182
Маниакальный психоз	1,7680	1,8611	0,3573	−5,8375

(44.58)

Коэффициент при x_1 для нормального состояния вычислялся, например, так:

$$(0{,}543\,234)(0{,}6000) - (0{,}200\,195)(0{,}1455) + \\ + (-0{,}420\,813)\ (0{,}2182) = 0{,}2050.$$

Допустим, что обследование некоего индивидуума привело к меткам 1, 1, 0. Тогда в силу (44.58) значения функций равны 0,2550, 0,2746; 0,0144; 0,0218; —1,0942; —2,2084; и мы должны приписать субъект ко второй группе (изменения личности). Разумеется, на практике такой вывод должен быть подвергнут тщательной проверке, так как значение дискриминатора для нормальной группы здесь весьма близко к такому же значению для группы с изменениями личности, последняя же состояла лишь из пяти членов, по которым мы построили выборочные дискриминаторы.

44.19 С геометрической точки зрения дискриминантные функции определяют некие гиперповерхности в p-мерном пространстве, которые в частном случае **44.18** являются плоскостями. Как мы видели, эти плоскости не ортогональны линиям, соединяющим средние распределений. Если мы имеем более чем две генеральные совокупности, средние в общем случае не колинеарны. Мы можем, однако, построить линию, наиболее близкую ко всем k средним и использовать изменение вдоль направления этой линии в качестве дискриминатора. На самом деле, если мы имеем k совокупностей, мы можем попытаться построить такую функцию X вида

$$X = \sum_{j=1}^{p} l_j x_j,$$

что отклонение межклассовой и внутриклассовой дисперсий будет максимальным. Как и в **44.6**, это аналогично максимизации

отношения межклассовой дисперсии к полной дисперсии. Если $\boldsymbol{A}$ — межклассовая матрица рассеяния, в $\boldsymbol{B}$ — полная матрица, последнее эквивалентно максимизации величины

$$\lambda = \frac{\sum A_{jk} l_j l_k}{\sum B_{jk} l_j l_k}, \tag{44.59}$$

что влечет равенство

$$\sum_k (A_{jk} - \lambda B_{jk}) l_k = 0. \tag{44.60}$$

Таким образом, наибольший корень уравнения $|\boldsymbol{A} - \lambda \boldsymbol{B}| = 0$ дает нам дискриминатор. Подробности можно найти у Бартлетта (1951), Е. Уилльямса (1952) и Блэкиса (1960). Выясняется, что использование для дискриминации между несколькими совокупностями одной функции может привести к довольно ограниченной по возможностям процедуре, если только совокупности не столь различны, что практически любой метод приводит к разумным результатам.

Качественные данные

44.20 До сих пор наши рассуждения касались лишь величин x, которые можно измерять. На практике мы часто имеем дело с ситуациями, где некоторые или все характеристики объекта носят качественный характер. Рассмотрим случай, когда все характеристики являются качественными, и допустим, что число их равно p. Пусть j-я характеристика разбита на категории s_{jk} ($k = 1, 2, \ldots, t$), а выборки из совокупностей 1 и 2 содержат соответственно n_{1jk} и n_{2jk} членов из k-й категории j-й характеристики. Пусть n_1 и n_2 — полные объемы выборок. Тогда, если для нас существенна лишь k-я категория j-й характеристики и у нового элемента эта характеристика имеет k-ю категорию, мы приписываем этот элемент совокупности 1, если

$$\frac{n_{1jk}}{n_1} > \frac{n_{2jk}}{n_2}, \tag{44.61}$$

и совокупности 2 в противном случае. Короче, значения имеют лишь пропорции в «(j, k)-классе», а частоты других классов ничего не говорят о принадлежности элемента к той или иной совокупности.

44.21 Такой образ действия кажется грубым, но мы поступаем в духе критерия, который использовался для измеряемых характеристик. Из (44.61) лишь следует, что процесс дискриминации определяется вероятностями появления того или другого элемента. Если категории s_{jk} упорядочены по k, т. е. если естественным образом задана последовательность $s_{j1}, \ldots, s_{jt}$ (как, например, при упорядоченной категоризации), то появляется возможность использовать информацию и о других «ячейках», кроме «(j, k)-ячейки». Насколько нам известно, до сих пор это делать не пытались.

Если мы подготовлены к тому, чтобы ввести цены за ложную классификацию, можно попытаться выработать несколько более усложненный критерий, по которому часть классов будет распределена по совокупностям так, чтобы минимизировать полную цену. Кокрен и Гопкинс (1961) исследовали эту процедуру. Тому же посвящена работа Линхэрта (1959).

Хиллс (1967) рассмотрел несколько «близлежащих» процедур и многошаговые методы.

44.22 По-видимому, наибольшие трудности возникают в том случае, когда некоторые характеристики объекта качественны, а другие поддаются измерению. Для этого случая пока еще нет удовлетворительной теории. Некоторое эвристическое продвижение заключается в том, чтобы отмечать качественные характеристики «метками» (используя, например, дихотомию (0, 1), трихотомию (—1, 0, 1) и т. д. и проводя усреднение), а затем использовать эти метки как результат измерения наравне с другими измеряемыми величинами. Напротив, если области значений измеряемых величин разбить на подобласти, можно осуществить дискриминацию, имитирующую дискриминацию по качественным характеристикам. Последняя процедура, однако, утомительна и фактически уменьшает объемы выборок, что приводит к снижению надежности. Указанная тематика заслуживает дальнейшего изучения.

44.23 Прежде чем перейти к рассмотрению методов, свободных от распределений, бегло остановимся на нескольких еще не обсуждавшихся вопросах: (а) отказ от решения (или откладывание решения); (б) смещение при оценивании вероятностей ложной классификации; (в) отбрасывание избыточных (излишних) величин.

Отказ от решения

44.24 Во многих, а возможно и в большинстве задач дискриминации, существуют пограничные случаи, когда разумно не настаивать на отнесении элемента к тому или иному классу, а отложить решение. Геометрически это означает, что мы разбиваем выборочное пространство на три области: R_1, R_2 и D_{12}. Если элемент попадает в R_1, мы приписываем его первой генеральной совокупности, если он попадает в R_2 — ко второй. Если же он попадает в D_{12}, мы должны признать, что данных недостаточно

для удовлетворительного заключения. В общем случае эта область будет содержать тесно перемешанные элементы из обеих генеральных совокупностей, и на практике нам пришлось бы искать какой-нибудь другой критерий для их различения.

Нетрудно определить область D_{12}, используя линейный дискриминатор. Мы должны только решить, какие вероятности ложной классификации считать допустимыми, основываясь на последнем, определить R_1 и R_2, а оставшуюся часть выборочного пространства объявить областью D_{12}.

44.25 С увеличением числа совокупностей число областей резко увеличивается. Если совокупностей, например, три, можно по трем дискриминаторам определить «области сомнения» D_{12}, D_{13}, D_{23}, но последние будут пересекаться. Поэтому возможна область D_{123}, при попадании в которую нельзя отнести элемент ни к одной из трех совокупностей; область $D_{12.3}$, где мы отвергаем третью совокупность, но не в состоянии различить первые две и так далее. По крайней мере при линейных дискриминаторах, которые разбивают пространство на области с плоскими границами, особых трудностей здесь не возникает. При использовании квадратичных и кубических форм может возникнуть проблема интерпретации.

Смещение при оценивании вероятностей ложной классификации

44.26 Простейший путь оценивания вероятностей ложной классификации состоит в применении построенного по наблюдениям дискриминатора к каждому члену тех выборок, на основе которых он получен, и к подсчету ошибок в этом случае. Если бы мы были убеждены в нормальности теоретического распределения, равенстве характеристик рассеяния, точности оценок средних и характеристик рассеяния, которые использовались при построении дискриминатора, мы бы смогли оценить вероятности ошибок теоретически, как в примере 44.1. Если же мы не знаем, в какой мере наш дискриминатор отвечает нашим предположениям, лучше установить вероятности ошибок, применяя дискриминатор к наблюденной выборке. В сущности, мы должны в любом случае следовать этой процедуре в качестве меры предосторожности. Однако, в результате такой процедуры возникает небольшое смещение.

44.27 На практике погрешности при эмпирическом определении вероятностей ложной классификации возникают по двум причинам. Прежде всего, нам неизвестны теоретические параметры, и дискриминатор мы строим на основе выборочных оценок. В среднем эмпирическая оценка вероятности ошибки будет больше истинного значения. Во-вторых, наша эмпирическая оценка получена по данным, на основе которых и был построен дискриминатор. Следовательно, в среднем, эмпирическая оценка будет меньше, чем в случае применения дискриминатора к новой выборке. Однако, как мы уже говорили, последняя оценка сама по себе будет больше, чем истинное значение.

Пример 44.5 (Кокрэн и Гопкинс, 1961).

Указанный эффект хорошо виден на следующем простом примере.

Имеются две генеральные совокупности P_1, P_2 и одна случайная переменная, которая может принимать значения a_1 и a_2. Пусть истинные вероятности того, что элемент, принадлежащий совокупности P_1, принимает соответствующие значения, равны $\pi_1(a_1)$ и $\pi_1(a_2) = 1 - \pi_1(a_1)$. Также определяем $\pi_2(a_1)$ и $\pi_2(a_2)$. Если выборка объема n_1 из P_1 содержит r_1 раз a_1, несмещенная оценка $\pi_1(a_1)$ равна r_1/n_1 и так далее.

Дискриминационное правило состоит в том, чтобы относить новый элемент к P_1, если наблюденное значение равно a_1 и $r_1/n_1 > r_2/n_2$ или, если наблюденное значение равно a_2 и $1 - r_1/n_1 > 1 - r_2/n_2$.

Рассмотрим теперь случай, когда истинные вероятности заданы следующим образом:

	P_1	P_2
a_1	0,9	0,05
a_2	0,1	0,95

Предположим, что мы имеем выборки единичного объема из каждой совокупности. (Это довольно тривиально, но достаточно, чтобы пояснить суть дела.) Если элементы из P_1 и P_2 принимают соответственно значения a_1 и a_2, правило на будущее состоит в том, чтобы относить к P_1 любой элемент, принявший значение a_1, и к P_2 любой элемент со значением a_2. Оценка вероятности ложной классификации равна нулю. В действительности эта вероятность равна $0{,}5\ (0{,}1 + 0{,}05) = 0{,}075$. Аналогично, если элементы из P_1 и P_2 принимают соответственно значения a_2 и a_1, мы принимаем противоположное решающее правило, но оценка вероятности ложной классификации снова равняется нулю. В действительности же она равняется $0{,}5\ (0{,}9 + 0{,}95) = 0{,}925$.

Если бы оба элемента приняли значение a_1, мы должны были бы или отказаться от решения, или решать вопрос, бросая монету. В последнем случае истинная вероятность ложной классификации и ее оценка равны 0,5. То же, если оба элемента приняли значение a_2.

Поскольку мы предполагали, что оба элемента выбираются из совокупностей случайным образом, можно эти вероятности усреднить. Результаты таковы:

Появление	Вероятность появления	Вероятность ложной классификации	
		оценка	истинная
$P_1(a_1)\ P_2(a_2)$	$0{,}9\cdot 0{,}95 = 0{,}855$	0,0	0,075
$P_1(a_1)\ P_2(a_1)$	$0{,}9\cdot 0{,}05 = 0{,}045$	0,5	0,500
$P_1(a_2)\ P_2(a_2)$	$0{,}1\cdot 0{,}95 = 0{,}095$	0,5	0,500
$P_1(a_2)\ P_2(a_1)$	$0{,}1\cdot 0{,}05 = 0{,}005$	0,0	0,925
		0,07	0,13875

Это конечно весьма крайний случай. При практически реальных объемах выборок смещение много меньше.

Дальнейшее обсуждение можно найти у Кокрэна и Гопкинса (1961). Смотри также работу Джона (1961). Значительно более полное изложение этой темы содержится у Хилла (1966).

Если первоначальная выборка, по которой построен наш дискриминатор, получена не случайным образом, никакая количественная оценка смещения невозможна.

Избыточные (или излишние) переменные: стандартные ошибки

44.28 Естествен следующий вопрос: все ли величины, от которых зависит наш дискриминатор, необходимы? Отбрасывание величин может ослабить мощность дискриминатора, но можно также ожидать, что потеря окажется незначительной. С геометрической точки зрения это означает, что если наше «созвездие» точек в p-мерном пространстве удачно разделено на два дискриминантной гиперплоскостью, мы можем сохранить положение, проектируя все точки на одну из координатных гиперплоскостей. В этом случае величины, соответствующие направлениям, ортогональным гиперплоскости, окажутся излишними. Существует несколько путей подхода к задаче. Мы бы избежали больших трудностей, если бы отбрасывали бесполезные величины в самом начале, не привнося их в анализ. Это, однако, рискованная операция. Некоторые результаты для нормального случая Кокрэна (1964) приведены в примерах 44.6—9. Дж. Елашов , Р. Елашов, Голдмэн (1967) показали, что простые правила, вытекающие из этих примеров, не работают даже в таких случаях, как выбор лучших двух величин из p биномиальных. Более прямой подход состоял бы в оценивании ошибок ложной

классификации при отбрасывании некоторых величин, но это может оказаться утомительным при большом числе величин. Мы осуществим третий подход, получая стандартные (для больших выборок) ошибки (или отклонения) коэффициентов линейного дискриминатора. Коэффициент $l_k = \sum_j C_{jk}(\bar{x}_{1j} - \bar{x}_{2j})$. Следовательно,

$$dl_k = \sum_j \{(\bar{x}_{1j} - \bar{x}_{2j})\, dC_{jk} + C_{jk} d(\bar{x}_{1j} - \bar{x}_{2j})\}. \qquad (44.62)$$

Точно так же

$$dl_m = \sum_r \{(\bar{x}_{1r} - \bar{x}_{2r})\, dC_{rm} + C_{rm} d(\bar{x}_{1r} - \bar{x}_{2r})\} \qquad (44.63)$$

и, следовательно,

$$\begin{aligned} dl_k\, dl_m = \sum_{j,\,r} \{&(\bar{x}_{1j} - \bar{x}_{2j})(\bar{x}_{1r} - \bar{x}_{2r})\, dC_{jk}\, dC_{rm} + \\ &+ (\bar{x}_{1j} - \bar{x}_{2j})\, C_{rm}\, dC_{jk}\, d(\bar{x}_{1r} - \bar{x}_{2r}) + \\ &+ (\bar{x}_{1r} - \bar{x}_{2r})\, C_{jk}\, dC_{rm}\, d(\bar{x}_{1j} - \bar{x}_{2j}) + \\ &+ C_{jk} C_{rm}\, d(\bar{x}_{1j} - \bar{x}_{2j})\, d(\bar{x}_{1r} - \bar{x}_{2r}). \end{aligned} \qquad (44.64)$$

Вспоминая, что при нормальном распределении средние не зависят от характеристик рассеяния, имеем

$$\begin{aligned} \operatorname{cov}(l_k, l_m) = \sum_{j,\,r} [&(\bar{x}_{1j} - \bar{x}_{2j})(\bar{x}_{1r} - \bar{x}_{2r}) \operatorname{cov}(C_{jk}, C_{rm}) + \\ &+ C_{jk} C_{rm} \operatorname{cov}\{(\bar{x}_{1j} - \bar{x}_{2j})(\bar{x}_{1r} - \bar{x}_{2r})\}]. \end{aligned} \qquad (44.65)$$

Если две выборки состоят из n_1 и n_2 наблюдений, легко получить, что

$$\begin{aligned} &\operatorname{cov}\{(\bar{x}_{1j} - \bar{x}_{2j})(\bar{x}_{1r} - \bar{x}_{2r})\} = \\ &\quad = \operatorname{cov}(\bar{x}_{1j}, \bar{x}_{1r}) + \operatorname{cov}(\bar{x}_{2j}, \bar{x}_{2r}) = \left(\frac{1}{n_1} + \frac{1}{n_2}\right) c_{jr}. \end{aligned} \qquad (44.66)$$

Займемся теперь ковариацией между C_{jk} и C_{rm}. Будем временно обозначать алгебраическое дополнение c_{jk} в $|c|$ символом Γ_{jk}, так что $C_{jk} = \Gamma_{jk}/|c|$. Тогда

$$\begin{aligned} dC_{jk} &= -\frac{\Gamma_{jk}}{|c|^2} d|c| + \frac{1}{|c|} d\Gamma_{jk} = \\ &= -\frac{\Gamma_{jk}}{|c|^2} \sum_{\alpha,\,\beta} \Gamma_{\alpha\beta}\, dc_{\alpha\beta} + \frac{1}{|c|} \sum_{\alpha,\,\beta} \Gamma_{jk,\,\alpha\beta}\, dc_{\alpha\beta}, \end{aligned}$$

где $\Gamma_{jk,\,\alpha\beta}$ — алгебраическое дополнение $c_{\alpha\beta}$ в Γ_{jk}. Итак,

$$dC_{jk} = \frac{1}{|c|^2} \sum_{\alpha,\,\beta} \{-\Gamma_{jk}\Gamma_{\alpha\beta} + |c|\,\Gamma_{jk,\,\alpha\beta}\}\, dc_{\alpha\beta}. \qquad (44.67)$$

Теперь, в силу теоремы Якоби об определителе,

$$|c|\Gamma_{jk.\alpha\beta} = \Gamma_{jk}\Gamma_{\alpha\beta} - \Gamma_{j\beta}\Gamma_{\alpha k}.$$

Следовательно, (44.67) сводится к равенству

$$dC_{jk} = -\frac{1}{|c|^2}\sum_{\alpha,\beta}\Gamma_{j\beta}\Gamma_{k\alpha}\,dc_{\alpha\beta} = -\sum_{\alpha,\beta}C_{j\beta}C_{k\alpha}\,dc_{\alpha\beta}. \qquad (44.68)$$

Используя (41.98), получаем

$$\operatorname{cov}(C_{jk}, C_{rm}) = \frac{1}{n_1+n_2}(C_{jm}C_{kr} + C_{jr}C_{km}). \qquad (44.69)$$

Осуществляя подстановку из (44.66) и (44.69) в (44.65), получаем

$$\operatorname{cov}(l_k, l_m) = \left(\frac{1}{n_1}+\frac{1}{n_2}\right)\sum_{j,r}C_{j,k}C_{r,m}c_{jr} +$$

$$+\frac{1}{n_1+n_2}\sum_{j,r}(\bar{x}_{1j}-\bar{x}_{2j})(\bar{x}_{1r}-\bar{x}_{2r})(C_{jm}C_{kr}+C_{jr}C_{km}) =$$

$$=\left(\frac{1}{n_1}+\frac{1}{n_2}\right)C_{mk} + \frac{1}{n_1+n_2}\sum_{j,r}(\bar{x}_{1j}-\bar{x}_{2j})(\bar{x}_{1r}-\bar{x}_{2r})C_{jm}C_{kr} +$$

$$+\frac{1}{n_1+n_2}C_{km}\sum_{j,r}(\bar{x}_{1j}-\bar{x}_{2j})(\bar{x}_{1r}-\bar{x}_{2r})C_{jr} = \qquad (44.70)$$

$$=\left(\frac{1}{n_1}+\frac{1}{n_2}\right)C_{mk} + \frac{1}{n_1+n_2}l_m l_k + \frac{1}{n_1+n_2}C_{km}\sum l_r(\bar{x}_{1r}-\bar{x}_{2r}) =$$

$$=\left(\frac{1}{n_1}+\frac{1}{n_2}\right)C_{mk} + \frac{1}{n_1+n_2}l_m l_k + \frac{1}{n_1+n_2}C_{km}(\bar{X}_1-\bar{X}_2). \qquad (44.71)$$

В частности, при $k = m = j$

$$\mathrm{D}l_j = \left(\frac{1}{n_1}+\frac{1}{n_2}\right)C_{jj} + \frac{l_j^2}{n_1+n_2} + \frac{1}{n_1+n_2}(\bar{X}_1-\bar{X}_2)C_{jj}. \qquad (44.72)$$

Пример 44.6

Рассмотрим снова данные для ирисов из примера 44.1. Исследуем l_1. Имеем

$$n_1 = n_2 = 50; \qquad l_1 = -3{,}11511;$$

$$\bar{X}_1 - \bar{X}_2 = 105{,}341; \qquad C_{11} = 11{,}87161.$$

Из (44.72) находим

$$\mathrm{D}l_1 = 0{,}4749 + 0{,}0970 + 12{,}5057 = 13{,}0776; \qquad \text{с. о.}(l_1) = 3{,}62.$$

Абсолютное значение меньше, чем стандартная ошибка, и следует подумать о том, нельзя ли отбросить x_1 без серьезной потери в мощности.

Как мы увидим позже, практически хороший дискриминатор может быть основан даже только на одной величине x_4.

Кокрэн и Блисс (1948) рассмотрели случай, когда влияние некоторых величин устраняется ковариационной техникой, а дискриминация основывается на оставшихся величинах. В их статье можно найти описание метода и рабочий пример.

Об использовании статистики D^2 в дискриминационных методах можно прочесть у С. Рао (1952).

Методы, свободные от распределений

44.29 Мы переходим к обсуждению возможностей методов свободных от распределения при дискриминации между k генеральными совокупностями. Над этой темой работали очень мало, и содержание пунктов **44.29—33** следует рассматривать как предложения, которые еще нуждаются в проверке опытом.

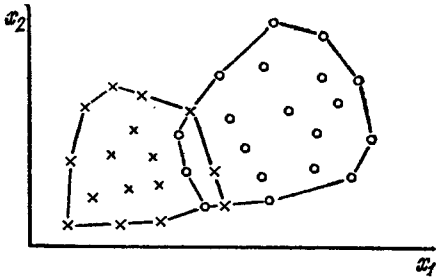

Рис. 44.1 (см. текст).

Вернемся к представлению элементов как точек в p-мерном пространстве, чьи координаты есть значения величин $x_1, \ldots, x_p$. Ограничившись пока двумя совокупностями, будем отмечать элементы одной совокупности (скажем, A) крестиками, а другой (скажем, B) кружочками. Картина в двумерном случае может выглядеть, как на рис. 44.1.

Крестики имеют выпуклую «оболочку», которую мы очертили, то же сделано и для кружочков. В общем случае две области имеют непустое пересечение.

Рассмотрим следующее дискриминационное правило:

(а) Если точка попадает в область A, но не попадает в область B, приписываем ее к A.

(б) Если точка попадает в область B, но не в A, приписываем ее к B.

(в) Если точка попадает в пересечение обеих областей, мы не приписываем ее ни к одной из совокупностей.

Предложение кажется разумным, но мы не можем следовать ему по трем причинам:

(1) Определение выпуклой оболочки есть задача линейного программирования, которая, хотя и разрешима, выводит нас за рамки наших настоящих интересов.

(2) Метод не указывает, как поступить с точками, которые не попадают ни в одну из областей.

(3) Метод в действительности не является свободным от распределения, поскольку нелинейное преобразование переменных не сохраняет границы плоскими.

Подсчет точек в обеих областях и их пересечении тем не менее оказывается полезным, поскольку указывает нам на степень «перемешивания» трех совокупностей или, так сказать, степень трудности задачи дискриминации.

44.30 В качестве вступления к изложению методов, свободных от распределений, рассмотрим снова данные таблицы 44.1 по щетинистым и разноцветным ирисам.

Ширина лепестка щетинистого ириса имеет среднее значение 0,246 и размах от 0,2 до 0,6 (дисперсия равна 0,0109). Та же величина у разноцветного ириса имеет среднее 1,326 и размах от 1,0 до 1,8 (дисперсия равна 0,03). В силу этого ширина лепестка может служить превосходным дискриминатором сама по себе. Если мы будем относить новый элемент к щетинистым или разноцветным ирисам в зависимости от того, будет ли ширина лепестка меньше или больше чем скажем, 0,9, мы будем редко впадать в ошибку, даже если случайные величины нормальны *).

44.31 Метод, который мы предлагаем, можно проиллюстрировать на дискриминации между ирисами разноцветным и вирджиника. Беглый просмотр данных показывает (и это можно подтвердить путем составления специальной таблицы), что указанные сорта больше отличаются по длине и ширине лепестка (ДЛ и ШЛ), чем по длине или ширине чашелистика. Распределение частот для ДЛ и ШЛ дано в таблице 44.2. При измерениях ДЛ в интервале от 4,5 до 5,1 оба сорта имеют ненулевые частоты появления. На внешность этого интервала приходятся 29 случаев ириса разноцветного и 34 — вирджиника. Для ШЛ «общим интервалом» является интервал 1,4—1,8, и на его внешность приходится 28 случаев ириса разноцветного и 34 вирджиника. Общее число случаев, приходящихся на внешность интервала пересечения, равно 63 для ДЛ и 62 для ШЛ. Прежде всего осуществим дискриминацию по ДЛ. Установим следующее правило:

$$\begin{aligned} &\text{ДЛ} \leqslant 4{,}4 && \text{— относим к разноцветным,} \\ &\text{ДЛ} \geqslant 5{,}2 && \text{— относим к вирджинике,} \qquad (44.73) \\ &4{,}5 \leqslant \text{ДЛ} \leqslant 5{,}1 && \text{— переходим к следующей характеристике.} \end{aligned}$$

В 37 случаях ДЛ приходится на общий интервал (4,5—5,1). Исключим эти случаи из таблицы 44.2 и построим для них распределение по ШЛ, как сделано в таблице 44.3.

*) По поводу того, что нормальный случай может оказаться в определенном смысле самым «невыгодным», см. **48.6**. (*Прим. ред.*)

Таблица 44.2

Распределение частот по длине и ширине лепестка ирисов разноцветного и вирджиника

Значения переменной	Длина лепестка разноцв.	Длина лепестка вирджин.	Значения переменной	Ширина лепестка разноцв.	Ширина лепестка вирджин.
4,3	25		1,0	7	
4,4	4		1,1	3	
4,5	7	1	1,2	5	
4,6	3	—	1,3	13	
4,7	5	—	1,4	7	1
4,8	2	2	1,5	10	2
4,9	2	3	1,6	3	1
5,0	1	3	1,7	1	1
5,1	1	7	1,8	1	11
5,2		2	1,9		5
5,3		2	2,0		6
5,4		2	2,1		6
5,5		3	2,2		3
5,6		6	2,3		8
5,7		3	2,4		3
5,8		3	2,5		3
5,9		13			
	50	50		50	50

Таблица 44.3

Распределение частот для 37 случаев, не различимых по ДЛ

Значения переменной	Ширина лепестка разноцв.	Ширина лепестка вирджиника
1,2	1	
1,3	2	
1,4	4	
1,5	9	2
1,6	3	—
1,7	1	1
1,8	1	5
1,9		3
2,0		3
2,1		—
2,2		—
2,3		1
2,4		1
	21	16

Поступая, как прежде, мы видим, что общий интервал для ШЛ лежит от 1,5 до 1,8. Таким образом, к (44.73) добавляем следующее правило:

$$\left.\begin{array}{l} \text{Если } 4{,}5 \leqslant \text{ДЛ} \leqslant 5{,}1 \text{ и ШЛ} \leqslant 1{,}4 \text{ — относим к разноцветным,} \\ \qquad \text{ШЛ} \geqslant 1{,}9 \text{ — относим к вирджинике,} \\ \qquad 1{,}5 \leqslant \text{ШЛ} \leqslant 1{,}8 \text{ — переходим к следующей характеристике} \end{array}\right\} \tag{44.74}$$

По ДЛ было дискриминировано 63 случая, по ШЛ — 15. Осталось 22 спорных случая. У оставшихся объектов рассмотрим длину чашелистика (ДЧ) и ширину чашелистика (ШЧ).

Исследуя ДЧ, мы видим, что только пять случаев лежат вне общего интервала. По ШЧ таких случаев шесть. Поэтому в качестве следующего дискриминатора выберем ШЧ и к (44.74) добавим следующее правило.

$$\left.\begin{array}{l} \text{Если } 4{,}5 \leqslant \text{ДЛ} \leqslant 5{,}1, \quad \text{ШЧ} \geqslant 3{,}1 \text{ — относим к разноцветным,} \\ \qquad 1{,}5 \leqslant \text{ШЛ} \leqslant 1{,}8 \text{ и ШЧ} < 3{,}1 \text{ — переходим к следующей характеристике.} \end{array}\right\} \tag{44.75}$$

По последней характеристике дискриминировано еще шесть случаев. Всего разобрано 84 случая, осталось неразобранными 16. Для этих 16 распределение по ДЧ дано в таблице 44.5.

Анализируя последнее, добавляем к (44.75) правило:

$$\left.\begin{array}{l} \text{Если } 4{,}5 \leqslant \text{ДЛ} \leqslant 5{,}1, \quad \text{ДЧ} \geqslant 6{,}4 \text{ — относим к разноцветным,} \\ \qquad 1{,}5 \leqslant \text{ШЛ} \leqslant 1{,}8, \quad \text{ДЧ} \leqslant 5{,}3 \text{ — относим к вирджинике,} \\ \qquad \text{ШЧ} < 3{,}1 \text{ и } 5{,}4 \leqslant \text{ДЧ} \leqslant 6{,}3 \text{ — отказываемся от решения.} \end{array}\right\} \tag{44.76}$$

Итак 87 случаев разобрано, 13 нет. Никакая дальнейшая дискриминация невозможна.

44.32 Теперь можно обратиться и к общему методу. Он полностью свободен от распределения и основывается лишь на рангах значений случайных величин. Мы просматриваем характеристики объекта одну за другой по степени важности их для дискриминации, при этом вычисления сводятся к самому элементарному счету.

С другой стороны, предлагаемый метод дискриминации может оказаться и не оптимальным. С геометрической точки зрения, вместо того чтобы устанавливать плоскую границу, как

Таблица 44.4

Распределение частот для 22 случаев, не различимых по ДЛ и ШЛ

Значения переменной	Длина чашелистика		Значения переменной	Ширина чашелистика	
	разноцв.	вирджиника		разноцв.	вирджиника
4,9		1	2,2	1	1
5,4	—	—	2,3	—	—
5,5	1	—	2,4	—	—
5,6	—	—	2,5	1	1
5,7	—	—	2,6	—	—
5,8	—	—	2,7	1	1
5,9	1	1	2,8	1	2
6,0	3	2	2,9	1	—
6,1	—	1	3,0	3	3
6,2	1	1	3,1	2	
6,3	2	2	3,2	2	
6,4	1	—	3,3	1	
6,5	1	—	3,4	1	
6,6	—	—			
6,7	2	—			
6,8	—	—			
6,9	1	—			
	14	8		14	8

Таблица 44.5

Распределения частот для 16 случаев, не различимых по ДЛ, ШЛ, ШЧ

Значения переменной	Длина чашелистика	
	разноцв.	вирджиника
4,9	—	1
—	—	—
5,4	1	—
5,5	—	—
5,6	1	—
5,7	—	—
5,8	—	—
5,9	—	1
6,0	2	2
6,1	—	1
6,2	1	1
6,3	1	2
6,4	—	—
6,5	1	—
6,6	—	—
6,7	1	—
	8	8

в примере 44.1, мы строим границу поэтапно. Дискриминируя по первой величине, мы разбиваем пространство на три области гиперплоскостями, ортогональными этой величине. Вторая величина порождает аналогичные разбиения уже в области, оставшейся после дискриминации по первой величине, и т. д. Возможно, что основанный на распределениях оптимальный метод приводит к меньшему числу неразобранных случаев, но это может произойти лишь за счет отказа от тех преимуществ, которые дает нам процедура, свободная от распределений.

Различия в характеристиках рассеяния

44.33 Заканчивая изложение проблемы дискриминации, остановимся на том случае, когда генеральные совокупности отличаются не средними, а характеристиками рассеяния. Эту задачу легче поставить, чем решить. Рассмотрим, например,

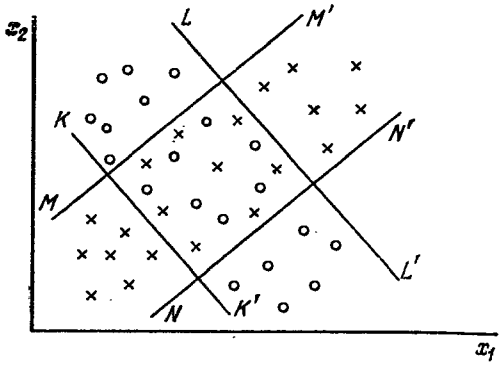

Рис. 44.2 (см. текст).

рис. 44.2, где совокупности имеют одинаковые средние, но различные характеристики рассеяния. Области, где дискриминация возможна, видны довольно ясно, но обнаружить их предшествующими методами не удается.

Если, сохраняя конфигурацию, повернуть рисунок на 45 градусов, ранговыми методами, можно прийти к значимым результатам. Линии KK', LL' ограничивали бы область, вне которой крестики, так сказать, доминировали бы. MM', NN' определяли бы аналогичные области для кружочков. Прямоугольник посередине стал бы зоной отказа от решения, причем неизбежно большой в силу характера данных.

Эвристическая процедура в таких случаях сводилась бы к вращению осей координат (если мы имеем дело с измеряемыми величинами), скажем, к преобразованию к главным компонен-

там Любишев (1962) обсуждал эту задачу в связи с задачами из биологии. См. также работу Бартлетта и Плиза (1963).

44.34 Одним из недостатков изложенного метода является его чувствительность к удаленным значениям. Как мы уже объясняли, для дискриминации пригодны лишь области, где не смешиваются элементы различных совокупностей. Если же величины могут в принципе принимать сколь угодно большие значения, а объем выборки растет, возрастает и тенденция к перемешиванию. Поэтому иногда предпочтительнее с самого начала пойти на некоторую ложную классификацию, разрешая перемешивание до определенного предела. Можно также, используя соответствующие одномерные распределения, урезать случайные величины, устанавливая границы, за которыми перемешивание допускается. Вообще говоря, в этой области осталось сделать значительно больше, чем уже сделано.

Классификация

44.35 Проблема классификации, как видно из самого термина, состоит в выяснении по эмпирическим данным, насколько элементы «группируются» или распадаются на изолированные «скопления», «кластеры» (clusters). Два различных подхода к проблеме соответствуют двум различным типам пространств, в которых мы представляем данные:

(а) Как обычно, задана $(p \times n)$-матрица наблюдений. Рассмотрим n выборочных точек в p-мерном евклидовом пространстве, порожденном p переменными. Если эти точки, соответственно некоторому приемлемому определению, разбиваются на ясно различимые группы, можно сказать, что n элементов классифицированы по этим группам. «Близость» элементов должна определяться некоторой функцией от их значений.

(б) Если p-мерное пространство вложить в n-мерное, значения по каждой компоненте будут представлены n-мерными векторами. Как мы видели в каноническом анализе, представляет интерес, насколько эти векторы группируются. Отметим, что в этом случае мы имеем дело с тем, насколько группируются компоненты, а не элементы выборки.

Хотя это и не общепринято, было бы удобно назвать первый тип классификационным анализом, а второй — классификацией по компонентам *). В первом случае компоненты вектора мы принимаем такими, какие они есть, и пытаемся классифицировать элементы выборки. Во втором случае, который логически

*) В подлиннике — classification analysis and cluster analysis. В литературе обычно как раз первый тип классификации часто называют кластер-анализом. Чтобы избежать путаницы, мы пошли на несколько вольный перевод. (*Прим. перев.*)

предшествует первому, мы интересуемся компонентами случайного вектора; например, насколько все они необходимы и которая из них наиболее важна для целей исследования.

44.36 В любом случае начальная трудность состоит в определении того, что мы понимаем под «классом» или «группой». Сделать это можно разными способами, но все они основаны на понятии «близости» или «расстояния». Следовательно, мы должны ввести некоторую метрику, определяющую расстояние между двумя точками, а затем уже установить расстояние, при котором две точки считаются «близкими». В классификации по компонентам естественным показателем расстояния между компонентами x_j и x_k является корреляция ρ_{jk}. Ее можно рассматривать или как косинус угла между векторами или как косинус расстояния по большой окружности единичной сферы между концами векторов. Таким образом, корреляционная матрица ρ представляет функцию расстояний. Мы должны лишь решить, какие значения означают близость и как использовать последнее при определении групп.

44.37 Предположим, что мы решили, что при $\rho \geqslant 0{,}7$ точки (компоненты) близки. Можно тогда, например, поступать следующим образом. По корреляционной матрице найдем пары с корреляцией $\geqslant 0{,}7$. Если таковых нет, невозможна и группировка компонент. В противном случае рассмотрим какую-нибудь пару с корреляцией $\geqslant 0{,}7$, скажем x_j, x_k. Исследуем, как коррелированы остальные случайные компоненты с этими двумя. Если существует такое x_l, что средняя корреляция (по значениям) между x_j, x_k, x_l окажется $\geqslant 0{,}7$, добавим x_l к группе. Начнем теперь, если это возможно, искать четвертую компоненту такую, что среднее ρ (между шестью значениями) $\geqslant 0{,}7$ и так далее, пока процесс не оборвется. Выделенные компоненты составляют группу. Обратимся теперь к оставшимся компонентам, повторим с ними ту же процедуру и будем так поступать до тех пор, пока не исчерпаем все компоненты. Такую процедуру довольно просто осуществить, если число компонент не превышает разумных пределов, — на практике это число редко бывает больше 50. Однако процедура не обладает свойством единственности в том смысле, что окончательный результат зависит от того, с какой именно пары мы начинаем наш процесс отбора. Если вычислительные возможности позволяют, можно, конечно, разбить p векторов на группы всевозможными способами (по числу представлений p в виде не содержащей единиц суммы целых чисел) и исследовать группы в каждом разбиении. Однако с этим вряд ли справится даже самая мощная вычислительная машина.

Дальнейшие исследования в этом направлении можно найти у Трайона (1939) и Фортье и Соломона (1966). Классификация по компонентам не часто привлекала внимание статистиков, и затронутые здесь вопросы нуждаются в дальнейшем исследовании в связи с отбрасыванием излишних величин, регрессионным анализом, структурным анализом, дискриминантным анализом и многомерным анализом вообще. Нужно также помнить, что коэффициенты корреляции являются характеристиками, носящими суммарный характер, и во всех этих случаях, чтобы получить представление об общем характере явления, разумно предварительно начертить несколько двумерных диаграмм рассеяния.

Кинг (1967) разработал два многошаговых метода формирования групп, один из которых минимизирует (42.45) на каждом шаге образования групп. (Эмпирические соображения позволяют предполагать, что оба метода работают не оптимально, но хорошо.) Х. Фридмэн и Рубин (1967) при определении групп минимизировали (42.17).

44.38 Метод классификации по компонентам с использованием корреляций имеет то преимущество, что не зависит от масштаба, в котором измеряются отдельные x_j. Мы, образно говоря, ориентируемся здесь на число компонент p, а не на способ, с помощью которого каждая из них измеряется. Такой метод не свободен от распределения, но если мы чувствуем некоторое неудобство от того, что распределения не являются нормальными, мы можем переставить компоненты по их рангам, и корреляционная процедура останется применимой. В действительности мы можем распространить метод и на качественные данные, используя тот факт, что категории можно упорядочить (см., например, **33.36** т. 2). В этих случаях можно ввести естественную метрику.

Если же мы рассматриваем проблему классификации n точек в p-мерном пространстве, такие соображения уже неприменимы. «Расстояния» здесь весьма чувствительны к изменению масштаба какой-либо переменной и могут принимать практически любые значения при соответствующем изменении масштаба. В ряде случаев указанная трудность, по крайней мере частично, преодолевается с помощью стандартной нормировки. Нам, однако, кажется предпочтительнее свободный от распределения ранговый метод.

44.39 Введем теперь функцию расстояния для элементов, а не для компонент. Пусть $x_{1j}, \ldots, x_{pj}$ и $x_{1k}, \ldots, x_{pk}$ — соответственно значения j-го и k-го случайных элементов. Нам нужно измерить степень корреляции между ними. Однако вычислять корреляцию, основываясь на самих значениях элементов, было бы весьма неразумно; ведь даже изменение знака одного из векторов повлекло бы изменение значений смешанных моментов.

Поэтому мы заменим n значений каждой компоненты x_j на последовательность рангов (от 1 до n). Ранги могут совпадать, если несколько элементов имеют одно значение x_j. В частности, качественные данные, распределенные по упорядоченным категориям, могут рассматриваться как частично ранжированные, так что наш метод применим в весьма общем случае. Первона-

чальная матрица заменяется $(p \times n)$-матрицей рангов, которые обозначаются $r_{\alpha\beta}$, $\alpha = 1, \ldots, p$; $\beta = 1, \ldots, n$.

Для каждой пары элементов выборки вычислим функцию, аналогичную статистике хи-квадрат:

$$\tilde{D}_{jk} = \sum_{\alpha=1}^{p} \frac{(r_{\alpha j} - r_{\alpha k})^2}{\mathrm{D}r_\alpha}. \qquad (44.77)$$

Дисперсия по множеству n рангов зависит от чисел совпадений. Если имеются группы из t_1, t_2, ... и т. д. совпадающих элементов, имеем (см. упражнение 44.4)

$$\mathrm{D}r_\alpha = \frac{1}{12n}\left\{(n^3 - n) - \sum (t^3 - t)\right\}. \qquad (44.78)$$

Заметим, что, хотя сами ранги и не инвариантны относительно изменения ранжирования на обратное, выражение (44.77) инвариантно.

44.40 Как и в большинстве процедур классификации, практическая трудность возникает в связи с тем, что число пар, которые можно выбрать из последовательности n элементов, может быть велико. Это число равно $n(n-1)/2$ и, если, скажем, размер выборки равен 100, число пар, каждой из которых приписана величина $\tilde{D}_{jk}$, равно 4950. Осуществить процедуру в духе **44.37** и формировать группы, добавляя по одному элементу на каждом этапе, достаточно трудно, и приходится обращаться к вычислительной машине, но теоретических проблем здесь нет.

Пример 44.7

Одна эвристическая процедура, отражающая по крайней мере некую предварительную идею разбиения по группам, может быть проиллюстрирована на части данных из таблицы 44.1 (Кендалл, 1966). Была построена таблица, использующая данные по ирисам разноцветному и вирджиника, в которой эти данные рассматриваются как принадлежащие одной выборке неизвестного происхождения, имеющей объем 100. Первые две компоненты — длина и ширина чашелистика были ранжированы обычным образом, от 1 до 100. Длины лепестков были разбиты на 4 категории, соответственно по значениям <4; $\geqslant 4$ и <5; $\geqslant 5$ и <6; $\geqslant 6$. Ширине лепестка соответствовали две категории: <2 и $\geqslant 2$ (ранжирование с большим числом совпадений). Было вычислено 4950 значений $\tilde{D}_{jk}$, которые использовались для сортировки элементов.

По данным было выявлено два четко определяемых, непересекающихся класса A и B, содержащих $58(A)$ и $25(B)$ элементов. Что касается оставшихся 17, то была выявлена еще одна группа, основанная на второй паре компонент. Однако фактически эта группа содержала 30 членов, 9 новых из оставшихся 17 и 21 из A. Таким образом, было решено объединить эти 9 с A и сформировать класс A из 67 элементов. B по-прежнему содержал 25 элементов. Оставшиеся 8 нельзя было отнести ни к одному четко определяемому классу.

Таким образом, удалось четко выявить два и только два класса. Но в то время как B содержал лишь разноцветные ирисы, A содержал как 48 вирджиники, так и 19 разноцветных. Поэтому мы правильно определили лишь число классов, но ложно классифицировали 19 элементов и определили как сомнительные 8. В аналогичной задаче дискриминации было принято решение для 87 случаев и оставлены как сомнительные 13. Однако в настоящем примере мы пожертвовали доброй долей информации, разбивая данные по длине и ширине лепестка на категории, так что результат не столь уж плох. Подробности можно найти у Кендалла (1966).

44.41 Наша тема далеко не исчерпана. Есть несколько способов разбиения элементов на группы даже в том случае, когда пригодная метрика уже введена. Например, можно рассмотреть классификацию, основанную на соотношениях между внутригрупповыми и внегрупповыми расстояниями. Вальд (1944) предложил одну статистику такого типа, которая оказалась весьма сродни дискриминантной функции.

В заключение хотелось бы отметить, что под влиянием работ Фишера (1936) и Вальда (1944) статистики в основном тяготели к такому решению задач дискриминации и классификации, где бы использовалась только одна функция от соответствующих величин. Нам такие процедуры по многим обстоятельствам представляются слишком жесткими. На самом деле мы нуждаемся в некотором правиле различения элементов или системе правил. В принципе последние могут использовать как линейную так и нелинейную функцию, как одну так и многофункций от случайных величин.

УПРАЖНЕНИЯ

44.1 Пусть x_1 и x_2 — соответственно длина и ширина чашелистика. Используя данные для ирисов щетинистого и разноцветного в примере 44.1, показать, что линейная дискриминантная функция равна $x_1 - 1{,}236x_2$ и что последняя почти столь же хороша, как и зависящий от четырех переменных дискриминатор соответствующего примера.

44.2 Показать, что дискриминантная граница, определяемая в (44.9), может быть также определена равенством

$$U \equiv \boldsymbol{x}'\boldsymbol{V}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)'\boldsymbol{V}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = 0,$$

где V — теоретическая матрица рассеяния. Показать, что если x распределено по закону $N(\mu_1, V)$, то U имеет среднее

$$\frac{1}{2}\alpha = \frac{1}{2}(\mu_1 - \mu_2)' V^{-1} (\mu_1 - \mu_2)$$

и дисперсию α, а если x распределено по закону $N(\mu_2, V)$, то U имеет среднее — $\alpha/2$ и дисперсию α.

(Т. Андерсон, 1958. Распределение в случае, когда μ и V заменены выборочными оценками, оказывается очень сложным, но асимптотически имеет тот же вид. См. Вальд (1944), Ситгривз (1952), Т. Андерсон (1951) и Джои (1961).)

44.3 Рассмотреть две гипотезы. Первая состоит в том, что $x, x_{11}, x_{21}, \ldots, x_{n_1 1}$ взяты из генеральной совокупности 1 с законом $N(\mu_1, V)$, а $x_{12}, \ldots, x_{n_2 2}$ взяты из генеральной совокупности 2 с законом $N(\mu_2, V)$. Вторая гипотеза состоит в том, что $x_{11}, x_{21}, \ldots, x_{n_1 1}$ взяты из совокупности 1, а $x, x_{12}, \ldots, x_{n_2 2}$ — из совокупности 2. Показать, что отношение правдоподобия для проверки этих сложных гипотез равно

$$\frac{1 + \frac{n_2}{n_2+1}(x - \bar{x}_2)' V^{-1} (x - \bar{x}_2)}{1 + \frac{n_1}{n_1+1}(x - \bar{x}_1)' V^{-1} (x - \bar{x}_1)}.$$

(Т. Андерсон, 1958.)

44.4 В последовательности n рангов ранги $p_{k+1}, \ldots, p_{k+t}$ совпадают и заменены рангом $p_k + \frac{1}{2}(t+1)$. Показать, что сумма квадратов рангов уменьшается на $(t^3 - t)/12$ Вывести формулу для дисперсии ранжирования с $t_1, \ldots, t_m$ совпадениями:

$$\mathrm{D}r = \frac{1}{12n}\left\{(n^3 - n) - \sum_{j=1}^{m}(t_j^3 - t_j)\right\}.$$

44.5 Данные для ирисов щетинистого и разноцветного могут быть классифицированы по длине и ширине лепестка так, как показано в следующей упорядоченной таблице (ширина лепестка названа «маленькой», если она $< 1{,}5$; длина лепестка названа «маленькой», если она $< 4{,}0$; «средней», если она $\geqslant 4$, < 5; «большой», если она $\geqslant 5$. Левые числа в колонке относятся к разноцветным, правые — к вирджинике).

		Ширина лепестка: маленькие	Ширина лепестка: большие	Всего
Длина лепестка	маленькие	11 : 0	0 : 0	11 : 0
	средние	24 : 0	13 : 6	37 : 6
	большие	0 : 1	2 : 43	2 : 44
Всего		35 : 1	15 : 49	50 : 50

Показать, что если мы новый элемент будем относить к тому или другому виду в зависимости от того, у какого вида в ячейке, куда попал элемент, частота больше, вероятность ложной классификации может быть оценена как 8 процентов. Исследовать значимость и надежность такой таблицы.

44.6 Пусть x есть p-мерный нормальный вектор, а выборки взяты из двух генеральных совокупностей с одинаковыми матрицами рассеяния. Каждая переменная в каждой совокупности нормирована так, что она имеет единичную дисперсию. Переменные в первой совокупности имеют нулевое среднее, а во второй положительные средние δ_j, $j = 1, \ldots, p$ (положительности можно добиться изменением знака, если это необходимо).

Показать, что если в качестве дискриминатора взять только x_1, а ошибки обоих типов признать одинаково важными, вероятность ложной классификации будет равна

$$\frac{1}{\sqrt{2\pi}} \int_{\delta_1/2}^{\infty} \exp\left\{-\frac{1}{2}x^2\right\} dx$$

(Кокрен (1964). Было высказано предположение, что если указанная вероятность велика, скажем $\delta_1 < 1/2$, соответствующая величина, как «бедный» дискриминатор, должна быть отброшена.)

44.7 Показать, что если все p переменных в предыдущем примере независимы, нормированный до единичной дисперсии наилучший полный дискриминатор равен

$$\sum_{j=1}^{p} \delta_j x_j \Big/ \sqrt{\sum \delta_j^2}.$$

Пусть x_1 и x_2 — независимые величины со средними δ_1 и δ_2 $(\delta_1 > \delta_2)$. Показать, что одно наблюдение по первой компоненте эквивалентно m наблюдениям по второй, если $m = \delta_1^2/\delta_2^2$.

44.8 Пусть величины x_1, x_2 из предыдущего примера имеют корреляцию ρ. Пусть $\delta_2 = f\delta_1$ $(0 \leqslant f \leqslant 1)$ и

$$(f - \rho)^2/(1 - \rho^2) > f^2.$$

Рассматривая независимые величины $x_2 - \rho x_1$ и x_1, показать, что в этом случае корреляция только улучшает дискриминацию по сравнению со случаем независимости. Вывести отсюда, что отрицательная корреляция помогает дискриминации всегда, а положительная, лишь если $\rho > 2f/(1 + f^2)$.

(Кокрэн, 1964.)

44.9 Пусть теперь в том же примере корреляции между парами x_j, x_k одинаковы и равны ρ. Показать, что если ρ отрицательно, дискриминатор, основанный на всех p компонентах, оказывается более надежным по сравнению с тем случаем, когда компоненты независимы. Если же ρ положительно, это верно в случае

$$\rho > \frac{\left(\sum_{j=1}^{p} \delta_j\right)^2 - \sum_{j=1}^{p} \delta_j}{(p-1)\sum_{j=1}^{p} \delta_j^2}.$$

(Кокрэн, 1964.)

44.10 Показать, что дискриминация между двумя генеральными совокупностями может формально рассматриваться как основанный на методе НК регрессионный анализ, в котором зависимая величина y может принимать два значения, а именно $m = n_2/(n_1 + n_2)$ для n_1 элементов из первой совокупности и $m - 1$ для n_2 элементов из второй. Границу в упражнении 44.2 можно получить тогда без предположения о многомерной нормальности.

44.11 Выяснить, почему подход, указанный в примере 44.10, не годится для случая более чем двух генеральных совокупностей.

ГЛАВА 45

ВРЕМЕННЫЕ РЯДЫ. ОБЩАЯ ЧАСТЬ

45.1 Наблюдения над некоторым явлением, характер которого меняется во времени, порождает упорядоченную последовательность, которую обычно называют временным рядом. Значения, которые принимает некоторая величина в момент t, могут как быть случайными, так и не быть таковыми. В большинстве ситуаций, с которыми мы будем иметь дело, элемент случайности связан лишь с ошибками наблюдений. Множество значений $u(t_1)$, $u(t_2)$, ..., $u(t_n)$ можно рассматривать как совокупность наблюдений над некоторым многомерным комплексом. Однако характерной чертой здесь является то, что порядок в последовательности t_1, ..., t_n существен, в то время как в случайной выборке x_1, ..., x_n индексы служат лишь для удобства идентификации.

45.2 Хотя переменную t обычно называют временным параметром и думают о ней как о времени, теория, которую мы собираемся развить, применима и при изучении пространственных изменений. Например, если мы исследуем изменение толщины нити вдоль ее длины l или зависимость интенсивности заражения сельскохозяйственного угодья жуком-щелкуном *) вдоль некоторого направления от расстояния по этому направлению d, то переменные l и d можно интерпретировать в том же духе, что и t. Возможны обобщения и на тот случай, когда исследуемые величины зависят от более чем одного параметра указанного типа. Мы, однако, таких обобщений касаться не будем.

45.3 Совокупность величин $u(t)$ отличается от многомерного комплекса еще тем, что переменная t может быть непрерывной. Мы можем, таким образом, оказаться вынужденными рассматривать бесконечную последовательность величин $u(t)$. Привычно и удобно (хотя и не очень точно) говорить о непрерывных временных рядах, когда мы хотим сказать, что t непрерывно. Последнее отнюдь не означает, что при фиксированном t $u(t)$ является непрерывной величиной. Точно так же, если $u(t)$

*) В оригинале — wire-worm, переводится так же как проволочник Elateridae. (*Прим. перев.*)

задано лишь на некотором конечном множестве точек t_1, ..., t_n, мы будем говорить о дискретном временном ряде. Разумеется само u может быть непрерывной величиной, такой, как, например, длина или вес.

В большей части наших исследований мы будем иметь дело с дискретными рядами, но там где это необходимо, будем говорить и о модификациях для непрерывного случая. Как правило, мы будем сталкиваться с рядами, определенными в равноотстоящих друг от друга моментах времени. Выбрав в качестве единицы времени интервал между двумя соседними моментами, можно обозначить члены ряда символами u_0, u_1, u_2, ... и т. д. Если нам требуются значения ряда в моменты, предшествующие начальному, можно ввести обозначения u_{-1} u_{-2}, ... и т. д.

Несколько примеров временных рядов

45.4 Читатель несомненно знаком с многочисленными примерами временных рядов, появляющихся в реальной действительности: кривая потребления товаров в течение многих лет, запись температуры и атмосферного давления в некотором районе, осуществленная неким самопишущим прибором, данные о населении какой-либо страны, полученные в дни переписей и т. д. Мы начнем с того, что приведем еще несколько примеров, которые укажут нам на область применений и будут служить числовыми иллюстрациями тех теоретических результатов, которые мы получим в дальнейшем.

В таблице 45.1 (данные проиллюстрированы на рис. 45.1) приведены данные о годовой урожайности ячменя (на один акр) в Англии и Уэльсе в период с 1884 по 1939 г. В таблице 45.2 (рис. 45.2) приведены данные о годовых осадках в Лондоне с 1813 по 1912 г. В таблице 45.3 (рис. 45.3) приведены данные о ежемесячном количестве яиц, приходящихся на одну несушку в США в период с 1938 по 1940 г. В таблице 45.4 (рис. 45.4) приведены данные о поголовье овец в Англии и Уэльсе на 4 июня с 1867 по 1939 г. Наконец, таблица 45.5 (рис. 45.5) показывает прирост населения в Англии и Уэльсе за каждое десятилетие в период с 1811 по 1961 г.

45.5 Эти ряды весьма типичны для того материала, с которым имеет дело наша теория. Данные из таблицы 45.1 (урожайность ячменя) указывают на очень нерегулярные колебания, и, насколько можно видеть, здесь нет никакой системы или тенденции к увеличению или уменьшению урожайности. Таблица 45.2 также отражает некоторые колебания, правда несколько более регулярные. Данные из таблицы 45.3 содержат в себе явную периодичность. В данных из таблицы 45.4 (поголовье

Таблица 45.1

Годовая урожайность ячменя (в *цн* на 1 акр) в Англии и Уэльсе с 1884 по 1939 г. (Данные из сельскохозяйственного статистического справочника)

Год	Урожай с акра (в *цн*)	Год	Урожай с акра (в *цн*)	Год	Урожай с акра (в *цн*)	Год	Урожай с акра (в *цн*)
1884	15,2	1898	16,9	1912	14,2	1926	16,0
85	16,9	99	16,4	13	15,8	27	16,4
86	15,3	1900	14,9	14	15,7	28	17,2
87	14,9	01	14,5	15	14,1	29	17,8
88	15,7	02	16,6	16	14,8	30	14,4
89	15,1	03	15,1	17	14,4	31	15,0
90	16,7	04	14,6	18	15,6	32	16,0
91	16,3	05	16,0	19	13,9	33	16,8
92	16,5	06	16,8	20	14,7	34	16,9
93	13,3	07	16,8	21	14,3	35	16,6
94	16,5	08	15,5	22	14,0	36	16,2
95	15,0	09	17,3	23	14,5	37	14,0
96	15,9	10	15,5	24	15,4	38	18,1
97	15,5	11	15,5	25	15,3	39	17,5

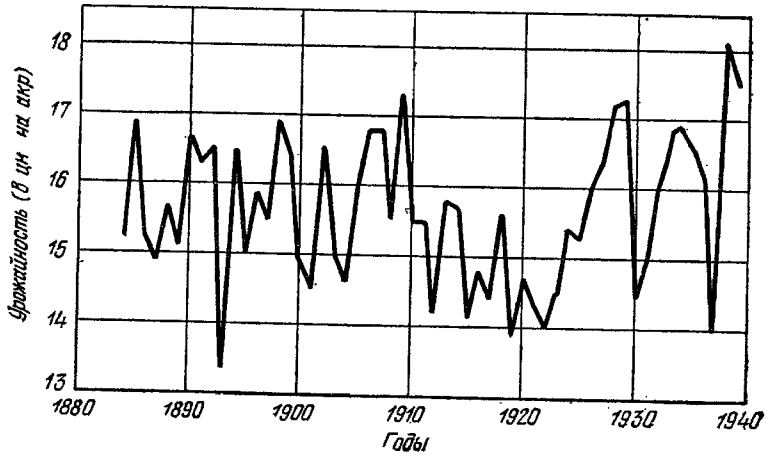

Рис. 45.1. Графяк значений ряда из таблицы 45.1 (урожайность ячменя на 1 акр).

Таблица 45.2

Годовое количество осадков (в дюймах) в Лондоне с 1813 по 1912 г. (Данные Д. Бранта, Phil. Trans., A, 247, 1925)

Год	Осадки (в дюймах)	Год	Осадки (в дюймах)	Год	Осадки (в дюймах)	Год	Осадки (в дюймах)
1813	23,56	1838	21,63	1863	21,59	1888	27,74
14	26,07	39	27,49	64	16,93	89	23,85
15	21,85	40	19,43	65	29,48	90	21,23
16	31,24	41	31,13	66	31,60	91	28,15
17	23,65	42	23,09	67	26,25	92	22,61
18	23,88	43	25,85	68	23,40	93	19,80
19	26,41	44	22,65	69	25,42	94	27,94
20	22,67	45	22,75	70	21,32	95	21,47
21	31,69	46	26,36	71	25,02	96	23,52
22	23,86	47	17,70	72	33,86	97	22,86
23	24,11	48	29,81	73	22,67	98	17,69
24	32,43	49	22,93	74	18,82	99	22,54
25	23,26	50	19,22	75	28,44	1900	23,28
26	22,57	51	20,63	76	26,16	01	22,17
27	23,00	52	35,34	77	28,17	02	20,84
28	27,88	53	25,89	78	34,08	03	38,10
29	25,32	54	18,65	79	33,82	04	20,65
30	25,08	55	23,06	80	30,28	05	22,97
31	27,76	56	22,21	81	27,92	06	24,26
32	19,82	57	22,18	82	27,14	07	23,01
33	24,78	58	18,77	83	24,40	08	23,67
34	20,12	59	28,21	84	20,35	09	26,75
35	24,34	60	32,24	85	26,64	10	25,36
36	27,42	61	22,27	86	27,01	11	24,79
37	19,44	62	27,57	87	19,21	12	27,88

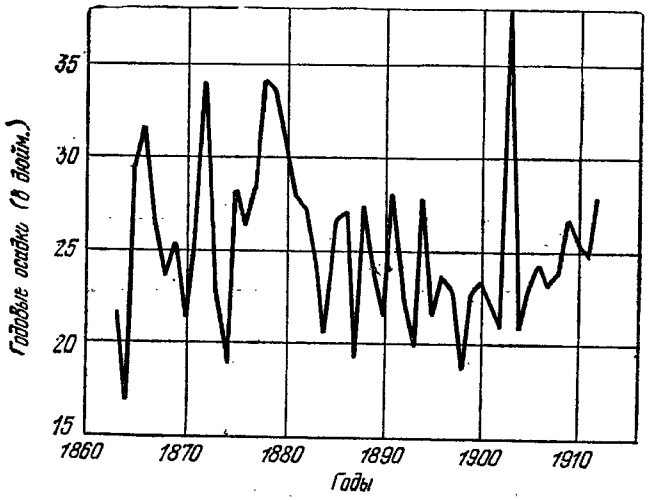

Рис. 45.2. График для последних 50 членов ряда из таблицы 45.2 (осадки).

Таблица 45.3

Среднее число яиц на несушку на каждый месяц по США с 1938 по 1940 г.
Данные из доклада Бюро по сельскохозяйственной экономике; США, Сельскохозяйственный департамент. (Report of the Bureau of Agricultural Economics, U. S. Dept. of Agriculture, on the Poultry and Egg Situation, March 1941)

Год	Январь	Февраль	Март	Апрель	Май	Июнь
1938	7,9	9,9	15,4	17,5	17,3	14,9
1939	8,0	9,7	14,9	17,0	17,0	14,6
1940	7,2	9,0	14,4	16,5	17,0	14,8

Год	Июль	Август	Сентябрь	Октябрь	Ноябрь	Декабрь
1938	13,6	11,8	9,4	7,5	5,9	6,4
1939	13,2	11,7	9,3	7,4	6,0	6,8
1940	13,4	11,8	9,7	7,9	6,2	6,8

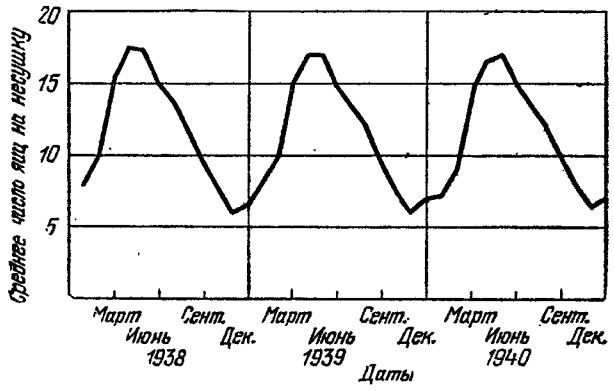

Рис. 45.3. График значений ряда из таблицы 45.3 (производство яиц).

Таблица 45.4

Поголовье овец в Англии и Уэльсе на каждый год с 1867 по 1939. (Данные из сельскохозяйственного справочника)

Год	Поголовье (10 000)	Год	Поголовье (10 000)	Год	Поголовье (10 000)	Год	Поголовье (10 000)
1867	2203	1886	1892	1904	1824	1922	1344
68	2360	87	1919	05	1823	23	1384
69	2254	88	1853	06	1843	24	1484
70	2165	89	1868	07	1880	25	1597
71	2024	90	1991	08	1968	26	1686
72	2078	91	2111	09	2029	27	1707
73	2214	92	2119	10	1996	28	1640
74	2292	93	1991	11	1933	29	1611
75	2207	94	1859	12	1805	30	1632
76	2119	95	1856	13	1713	31	1775
77	2119	96	1924	14	1726	32	1850
78	2137	97	1892	15	1752	33	1809
79	2132	98	1916	16	1795	34	1653
80	1955	99	1938	17	1717	35	1648
81	1785	1900	1928	18	1648	36	1665
82	1747	01	1898	19	1512	37	1627
83	1818	02	1850	20	1338	38	1791
84	1909	03	1841	21	1383	39	1797
85	1958						

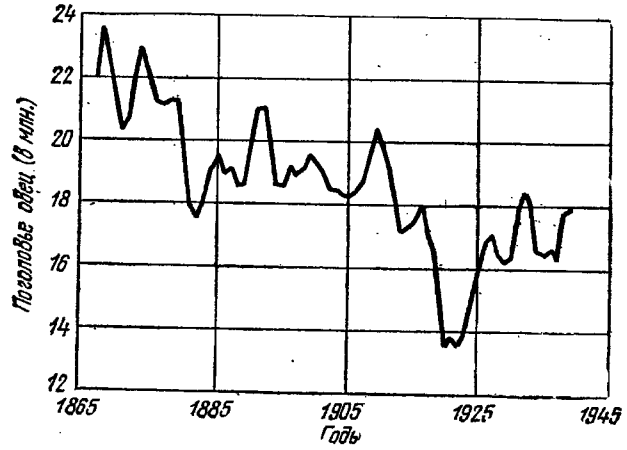

Рис. 45.4. График значений ряда из таблицы 45.4 (поголовье овец).

Таблица 45.5

Население Англии и Уэльса через десятилетия с 1811 по 1961 г. (Данные из Registrar-General's Statistical Review)

Год	Население (в млн.)	Год	Население (в млн.)	Год	Население (в млн.)	Год	Население (в млн.)
1811	10,16	1851	17,93	1891	29,00	1931	39,95
21	12,00	61	20,07	1901	32,53	41	—
31	13,90	71	22,71	11	36,07	51	43,76
41	15,91	81	25,97	21	37,89	61	46,07

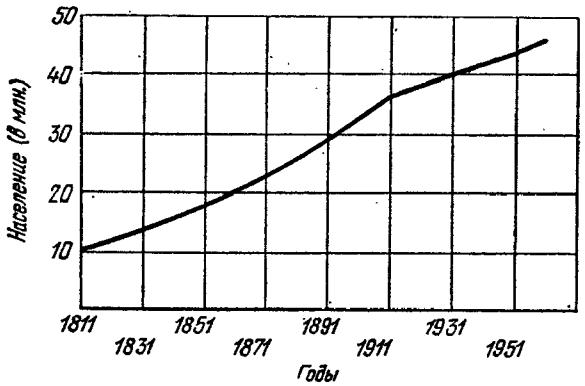

Рис. 45.5. График значений ряда из таблицы 45.5 (население Англии и Уэльса).

овец) колебания накладываются на явное систематическое смещение. Наконец, таблица 45.5 (численность населения) указывает на регулярный рост без каких-либо видимых флуктуаций.

Типы дискретности

45.6 Приведенные выше примеры временных рядов хорошо иллюстрируют различные типы дискретности.

(а) Временной ряд урожайности ячменя представляет случай существенной дискретности. Каждый год на одном акре получают один и только один урожай. Действительное время уборки урожая, правда, может меняться год от года, но, огрубляя, можно считать, что интервалы между последовательными наблюдениями одинаковы.

(б) Дискретность ряда численности населения обусловлена тем фактом, что переписи проводятся раз в десять лет. Исследуемая величина, однако, принимает определенные значения в течение всего периода и теоретически ее можно наблюдать в

любой момент времени. То же можно сказать и о ряде данных о поголовье овец.

(в) В данных об осадках дискретность обусловлена агрегацией. Осадки не выпадают каждый момент времени, и мы вычисляем суммарное количество за некоторый конечный интервал времени, представляющий для нас интерес. Длина интервала, конечно, определяется нашим выбором. Можно в качестве такого интервала выбрать год, месяц, день и даже час. Интервалы могут даже пересекаться, например, когда из недельных осадков мы составляем месячные. Такие данные тем не менее представляют дискретный в нашем смысле временной ряд.

(г) Если же мы имеем некую непрерывную запись, например запись давления на барографе, мы не можем затабулировать данные в стиле таблиц 45.1—45.5. Можно, разумеется, использовать показания прибора в любой отдельный момент времени, но учесть показания во все моменты времени, конечно, нельзя. Следовательно, анализ таких рядов численными методами, являющимися, по существу, дискретными, возможен лишь в качестве приближения. Можно, однако, анализировать такие ряды методами, включающими графическое интегрирование, например, с помощью планиметра или более современных устройств.

45.7 На практике, особенно в экономике, моменты времени, в которые можно проводить наблюдения, часто даны заранее. В ситуациях, носящих более экспериментальный характер, мы можем решить этот вопрос сами до того, как данные уже собраны, или даже после, если осуществлено наиболее полное из всех возможных исследование. Вопрос о наилучшем интервале между наблюдениями решается в зависимости от обстоятельств, свойственных рассматриваемому случаю. Здесь мы этого вопроса касаться не будем. (Тем более, что существует очень мало теоретических работ по этому поводу). Отметим, однако, что проводя наблюдения через равные фиксированные интервалы (что само по себе удобно), можно пройти мимо колебательных процессов, происходящих с периодом, равным длине этих интервалов или кратным ей. Так, например, ежегодные наблюдения за поголовьем овец не учитывают сезонные изменения, связанные с убоем и окотом, данные о годовых осадках не отражают тот факт, что выпадение осадков носит до некоторой степени сезонный характер даже в Лондоне.

«Календарные неприятности»

45.8 Равны временные интервалы или нет, но очевидным образом желательно, чтобы результаты наблюдений были сравнимы между собой. В том случае, когда временными интервалами

служат дни или месяцы, чтобы обеспечить такую сравнимость, приходится устранять специфические мешающие эффекты, связанные со свойствами календаря. В некоторых трудностях такого рода остается винить лишь саму природу; так, ничего не поделаешь с тем, что год не содержит целого числа астрономических суток. Но большинство трудностей связано с календарем, изобретенным самим человеком. Месяцы, например, имеют разную продолжительность, общественные праздники влияют на сравнимость экономических и социологических данных, торговые и обменные операции прекращаются на время уикэнда и т. д. Проводя эксперименты, обычно можно позаботиться заранее о том, чтобы такого рода трудности исключить, но при анализе рядов, связанных с реальным производством, указанные препятствия могут играть определенную роль, как серьезную (например, остановки в связи с забастовками), так и менее серьезную (например, обеденные перерывы). В дальнейшем мы будем предполагать, что наши данные подправлены так, что указанные эффекты оказались устраненными.

Задача анализа временных рядов

45.9 Двигает ли нами чистое любопытство или желание предвидеть будущее, но конечной целью анализа временных рядов, как и статистического анализа вообще, является достижение более глубокого понимания тех причинных механизмов, которые обусловливают появление этих рядов. Отсюда, однако, не следует, что такого понимания можно достигнуть, рассматривая лишь какой-то один ряд, поскольку последний может отражать лишь одну грань сложного явления, порождающего, вообще говоря, несколько различных рядов. Мы вернемся к этому вопросу в главе 50, где будут обсуждаться многомерные системы, а пока несколько ограничимся в наших стремлениях и обратимся к изучению различных типов поведения отдельных рядов и построению моделей, которые могли бы эти типы поведения объяснить. При этом мы будем сознавать, что такие модели сами по себе могут являться лишь частями структурно более сложных систем. Позднее будет видно, что такой подход не влечет за собой никакой логической противоречивости.

45.10 Читатель, изучая реальные ситуации (как те, о которых мы уже говорили, так и другие, известные ему самому), придет к выводу, что типичные временные ряды могут складываться из четырех составляющих:

(а) тренд, или систематическое движение;

(б) колебания относительно тренда с большей или меньшей регулярностью;

(в) эффект сезонности;

(г) «случайная», «несистематическая» или «нерегулярная» компонента.

Что касается математического описания, то мы всегда будем рассматривать ряд как одну из таких составляющих или сумму нескольких из них. Фактически большая часть традиционной теории временных рядов посвящена анализу данных, основанному на разложении данных на указанные компоненты и дальнейшем отдельном изучении последних. Мы должны здесь, однако, помнить следующее. Из того, что мы можем представить ряд как сумму указанных компонент, отнюдь не следует, что последние соответствуют независимо действующим системам. Разложение ряда часто оказывается очень полезным, но может и ввести в заблуждение. Во всяком случае такая операция не должна являться конечной целью статистического анализа.

45.11 Пожалуй, наиболее легок для обнаружения, выделения и изучения эффект сезонности. Здесь мы имеет дело с изменениями, накладываемыми на систему неким циклическим механизмом, внешним по отношению к основным механизмам, определяющим поведение системы. Например, колебания в производстве яиц в таблице 45.3, отражают ритм процесса воспроизводства, определяемый в конечном счете тем фактом, что земля совершает оборот вокруг солнца один раз в год. Мы будем относить термин «сезонность» к эффектам, которые происходят с периодом в год, но те же идеи могут быть использованы при изучении любых явлений, обусловленных строго периодическими процессами в природе; такими, например, как изменения уровней приливов и отливов или изменения температуры в течение дня. Мы, однако, должны быть осторожны, распространяя понятие сезонности на явления, чья строгая периодичность не проявляется явным образом и потому подвергается сомнению. Например, при современном уровне знаний мы зашли бы слишком далеко, утверждая, что в изменениях пятен на солнце есть в нашем смысле сезонность, и уж было бы совсем смело говорить о том, что сезонность в урожайности определяется солнечными пятнами, даже если некоторая связь здесь и установлена. Мы еще вернемся к этому вопросу ниже, когда будем различать понятия «цикла» и «колебания».

45.12 Определить понятие тренда труднее. Вообще под трендом понимают некое устойчивое, систематическое изменение в течение долгого периода. Отметим, что понятие «долгий» здесь относительно, и то, что с одной точки зрения является долгим, с другой таковым не будет. Например, если мы исследуем величины осадков в течение сотен лет, медленное увеличение в течение всего периода может быть воспринято как тренд, но если мы обладаем данными за два тысячелетия (по кольцам некоторых красных деревьев можно судить об изменении климатических

условий в течение приблизительно такого периода), рост количества осадков в течение некоторого столетия может оказаться лишь частью некоторого медленного колебательного процесса. Поэтому, исходя из «тренда» в течение столетия, приходить к выводу, что погода становится все более и более влажной, было бы ошибкой. Практически вывод, который мы делаем, должен зависеть от того, что мы хотим. Если мы являемся инженерами, проектирующими систему водоснабжения и хотим принять меры против возможного переполнения водохранилищ, мы можем считать, что тренд будет длиться столько, сколько будет работать система, и поступать соответствующим образом. Но если бы мы пытались изучать климатические изменения в масштабе всей земли в течение геологических периодов, мы или приняли бы факт существования тренда с весьма большими оговорками, или, что скорее всего, отвергли бы его на основании дополнительных соображений.

45.13 Как бы ни был велик ряд, мы никогда не можем быть уверенными, что тренд не является частью медленного колебательного процесса, за исключением, конечно, того случая, когда ряд, по существу, конечен. (Например, в случае, когда мы рассматриваем продолжительности правлений римских императоров.) Таким образом, говоря о тренде, мы должны иметь в виду длину ряда, к которому относится наше утверждение. Может быть, было бы более точно говорить о медленном или быстром движении, чем о тренде или колебаниях, но даже поступая так, мы не сможем исключить из наших рассуждений некоторый элемент субъективности.

45.14 Выделив тренд и сезонные (периодические) изменения, мы столкнемся с рядами, представляющими флуктуации более или менее регулярного типа. На рис. 45.1 представлен ряд такого типа, поскольку в этом примере отсутствуют тренд и сезонность. Возникает вопрос: являются ли остаточные ряды систематическими в том смысле, что их значения можно представить в виде функции от времени? Или напротив эти значения случайны в том смысле, что представляют собой случайную выборку из некоторой однородной генеральной совокупности? Или же мы имеем дело с чем-то средним между точной функциональной зависимостью и обыкновенной случайной выборкой? Исследование систематических эффектов в остаточных флуктуациях позволяет нам разработать технику анализа, цель которого — выяснить, подчинены ли ряды некоторому закону и, следовательно, предсказуемы, или любая часть их абсолютно случайна. Флуктуации первого типа мы будем называть систематическими и относить их к колебаниям (а не циклам, которые, как мы позже увидим, представляют очень частный случай колебаний). Флуктуацию второго типа мы назовем несистематической компонентой и будем трактовать ее как случайную. В том случае, когда ряд представляет смесь колебательного и случайного движения, и мы еще не разложили его на составляющие, движение, заключающееся в последовательных «взлетах» и «падениях», можно трактовать как флуктуацию вообще. Иначе, можно говорить о флуктуациях без ущерба для дальнейшей возможности обнаружения колебательного движения.

Критерии случайности

45.15 Некоторые из рядов, которые мы разбирали, заведомо не случайны. Сомневаться в присутствии систематических эффектов в данных из таблиц 45.3 и 45.5 не приходится. В других же случаях, как, например, в таблицах 45.1 (урожайность ячменя) или 45.2 (осадки), не очевидно, есть ли в данных некая система или нет. Оставшуюся часть главы мы в основном посвятим обсуждению критериев случайности для рядов. Более точно мы будем отвечать на следующий вопрос. Пусть дана упорядоченная последовательность наблюдений $u_1, \ldots, u_n$. Представляют ли эти данные результат независимых наблюдений над некоторой случайной величиной, т. е. имеем ли мы дело с выборкой из генеральной совокупности с неизвестными нам характеристиками?

45.16 Нет предела числу критериев, которые можно построить для этой цели. Выбирая наиболее подходящие, мы должны руководствоваться следующими требованиями:

(а) Критерий, если это возможно, должен быть «свободен от распределения».

(б) Так как мы хотим проверять довольно длинные ряды, вычисления должны быть сведены к минимуму.

(в) Хотя мы и не в состоянии определить точно альтернативную гипотезу, мы можем иметь некоторые соображения о ее природе и выбирать критерий, который был бы близок к тому, чтобы иметь наибольшую мощность относительно альтернативы. Например, если мы подозреваем, что присутствует тренд, более полезным может оказаться иной критерий, чем тот, что мы используем, имея в качестве альтернативы периодичность.

Экстремальные точки

45.17 Один из наиболее простых для применения критериев состоит в подсчете «пиков» или «ям» в ряде. Пик есть значение, которое больше двух соседних. Если есть два и более равных значения, которые больше предшествующих и последующих, мы будем рассматривать их как один пик. Точно так же «ямой» назовем значение, которое меньше двух соседних. Наш первый

вопрос таков: каково распределение пиков в случайном ряде? (Распределение ям, очевидно, такое же. Для доказательства достаточно изменить знак у случайной величины.)

В действительности нам будет удобнее изучать сразу и пики и ямы, иначе говоря «экстремальные точки» ряда (или «точки поворота»). Ясно, что число экстремальных точек на единицу меньше, чем число интервалов монотонности. Интервал между двумя экстремальными точками назовем «фазой».

45.18 Чтобы определить экстремальную точку, требуются три последовательных наблюдения. Если бы ряд был случаен, эти три величины могли бы появиться в любом порядке: всего шесть вариантов. Только в четырех вариантах есть экстремальная точка (когда наибольшее или наименьшее значение в середине). Следовательно, вероятность появления экстремальной точки в случае трех величин равна 2/3.

Рассмотрим теперь последовательность $u_1, \ldots, u_n$ и определим «индикатор» X_i таким образом:

$$\begin{aligned} &X_i = 1, \quad \text{если} \quad u_i < u_{i+1} > u_{i+2} \quad \text{или} \quad u_i > u_{i+1} < u_{i+2}, \\ &X_i = 0 \quad \text{в противном случае}, \quad i = 1, \ldots, n-2. \end{aligned} \tag{45.1}$$

Число экстремальных точек тогда просто равно

$$p = \sum_{i=1}^{n-2} X_i. \tag{45.2}$$

Сразу же имеем

$$\mathsf{M}p = \sum_{i=1}^{n-2} \mathsf{M}X_i = 2(n-2)/3. \tag{45.3}$$

Кроме того,

$$\begin{aligned} \mathsf{M}p^2 &= \mathsf{M}\left(\sum_1^{n-2} X_i\right)^2 = \\ &= \mathsf{M}\left\{\sum_{n-2} X_i^2 + 2\sum_{n-3} X_iX_{i+1} + 2\sum_{n-4} X_iX_{i+2} + \sum_{(n-4)(n-5)} X_iX_{i+k}\right\}. \end{aligned} \tag{45.4}$$

$$k \neq 0, 1, 2,$$

Индексы под знаком $\sum$ указывают на число слагаемых в сумме. В качестве проверки заметим, что

$$(n-2)^2 = (n-2) + 2(n-3) + 2(n-4) + (n-4)(n-5).$$

Отсюда имеем

$$\begin{aligned} \mathsf{M}p^2 = (n-2)\,\mathsf{M}X_i^2 + 2(n-3)\,\mathsf{M}(X_iX_{i+1}) + \\ + 2(n-4)\,\mathsf{M}(X_iX_{i+2}) + (n-4)(n-5)\,\mathsf{M}(X_iX_{i+k}), \end{aligned} \tag{45.5}$$

Поскольку $X_i^2 = X_i$,

$$\mathsf{M}X_i^2 = \frac{2}{3}. \tag{45.6}$$

При $k > 2$ X_i и X_{i+k} независимы, поскольку определяются различными u. Отсюда

$$\mathsf{M}(X_iX_{i+k}) = \mathsf{M}X_i\mathsf{M}X_{i+k} = \frac{4}{9}. \tag{45.7}$$

Нам осталось вычислить $\mathsf{M}(X_iX_{i+1})$ и $\mathsf{M}(X_iX_{i+2})$. Прежде всего рассмотрим четыре последовательных члена ряда. Наибольшему припишем номер 1, следующему по величине номер 2 и т. д. Величина X_iX_{i+1} не будет равняться нулю лишь тогда, когда экстремальные точки будут находиться на втором и третьем месте. Если читатель выпишет все 24 перестановки чисел 1, 2, 3, 4, он убедится, что только при 10 перестановках величина X_iX_{i+1} не будет равняться нулю. Эти перестановки таковы:

1324	2143	3142	4132
1423	2314	3241	4231
	2413	3412	

Итак,

$$\mathsf{M}(X_iX_{i+1}) = \frac{10}{24} = \frac{5}{12}. \tag{45.8}$$

Для X_iX_{i+2} мы должны выписать 120 перестановок чисел 1, 2, 3, 4, 5 и подсчитать те перестановки, где экстремальные точки находятся одновременно на втором и четвертом месте. Таких перестановок 54 и, следовательно,

$$\mathsf{M}(X_iX_{i+2}) = \frac{54}{120} = \frac{9}{20}. \tag{45.9}$$

Осуществляя подстановку в (45.5), после упрощений находим

$$\mathsf{M}p^2 = \frac{40n^2 - 144n + 131}{90}.$$

Используя (45.3), получаем

$$\mathsf{D}p = \frac{16n - 29}{90}. \tag{45.10}$$

Моменты более высокого порядка получаются сходным образом. Имеем

$$\varkappa_3(p) = \frac{-16(n+1)}{945}. \tag{45.11}$$

$$\varkappa_4(p) = \frac{-1408n + 3317}{18\,900}. \tag{45.12}$$

Таким образом, после стандартной нормировки $\varkappa_3(p)$ приблизительно равен $-0{,}2n^{-1/2}$, а $\varkappa_4(p)$ приблизительно равен $-2{,}4n^{-1}$. Последнее говорит о быстром стремлении к нормальности при увеличении n.

45.19 Рассмотрим теперь распределение длины фазы. Чтобы установить наличие фазы длины d (скажем, в случае роста), нам нужно обнаружить $d+3$ члена, где первый больше второго, второй меньше третьего, третий меньше четвертого, $(d+1)$-й меньше $(d+2)$-го, а $(d+2)$-й больше $(d+3)$-го. Рассмотрим $d+3$ значения, расположенные в порядке возрастания. Если мы возьмем любые два, кроме первого и последнего, значения u, поставим одно на первое место, а второе на последнее, мы получим случай фазы длины d. Существует $d(d+1)/2$ способов такого выбора пар. Каждый член пары может быть поставлен на любой конец, и мы имеем, таким образом, $d(d+1)$ фазы длины d. Но, кроме того, мы можем поставить первый член в конец, а любой другой, за исключением второго, в начало, получая при этом еще $d+1$ случай. Можно также последний ставить в начале, и любой, за исключением предпоследнего, в конец, что дает еще $d+1$ случай. Правда мы должны теперь из общего числа вычесть тот случай, когда первый становится последним, а последний первым, так как этот случай мы посчитали дважды. Итак, существует

$$(d+1)d+(d+1)+(d+1)-1=d^2+3d+1$$

фазы для случая роста. Вероятность фазы, следовательно, в случае, когда имеется или рост или снижение, равна

$$\frac{2(d^2+3d+1)}{(d+3)!}. \tag{45.13}$$

Далее для ряда длины n аналогично (45.2) число фаз длины d будет равно сумме $n-d-2$ соответствующих индикаторов (по числу возможных фаз). Следовательно, математическое ожидание числа фаз длины d равно

$$N_d=\frac{2(n-d-2)(d^2+3d+1)}{(d+3)!}. \tag{45.14}$$

Из (45.14) следует, что математическое ожидание полного числа фаз равно

$$N=2\sum_{d=1}^{n-3}\frac{(n-d-2)(d^2+3d+1)}{(d+3)!}.$$

Поскольку

$$(n-d-2)(d^2+3d+1)=$$
$$=-(d+3)(d+2)(d+1)+(n+1)(d+3)(d+2)-$$
$$-(2n+1)(d+3)+(n+1),$$

имеем

$$N=2\sum_{d=1}^{n-3}\left\{-\frac{1}{d!}+\frac{n+1}{(d+1)!}-\frac{2n+1}{(d+2)!}+\frac{n+1}{(d+3)!}\right\}=$$
$$=2\left(\frac{2n-7}{6}+\frac{1}{n!}\right). \tag{45.15}$$

На практике можно пренебречь вторым членом в (45.15) и, следовательно,

$$N\doteqdot\frac{1}{3}(2n-7). \tag{45.16}$$

Поскольку число фаз на единицу меньше числа экстремальных точек, за исключением двух из $n!$ случаев, когда оба числа равны нулю, (45.15) согласуется с (45.3). Далее

$$N_d/N\doteqdot\frac{6(n-d-2)(d^2+3d+1)}{(d+3)!(2n-7)}. \tag{45.17}$$

Можно довольно легко вычислить моменты этого отношения. Например,

$$\mu_1'=\frac{6}{2n-7}\sum_{1}^{n-3}\frac{d(n-d-2)(d^2+3d+1)}{(d+3)!}=$$
$$=\frac{6}{2n-7}\sum_{1}^{n-3}\left\{-\frac{1}{(d-1)!}+\frac{n+1}{d!}-\frac{3n+2}{(d+1)!}+\frac{5n+3}{(d+2)!}-\frac{3(n+1)}{(d+3)!}\right\}.$$

Вспоминая о быстрой сходимости $\sum_0^n 1/m!$ к e, мы можем с высокой степенью точности утверждать, что

$$\mu_1'=\frac{6}{2n-7}\left\{-e+(n+1)(e-1)-(3n+2)(e-2)+\right.$$
$$\left.+(5n+3)\left(e-\frac{5}{2}\right)-3(n+1)\left(e-\frac{8}{3}\right)\right\}=\frac{3(n+7-4e)}{2n-7}\doteqdot\frac{3}{2}. \tag{45.18}$$

Аналогично находим, что

$$\mu_2=\frac{3}{(2n-7)^2}\{(8e-21)n^2+(4e-17)n-(48e^2-140e+14)\}\doteqdot 0{,}560. \tag{45.19}$$

45.20 Распределение, моменты которого мы вычисляли, не сходится к нормальному с ростом n (см. упражнение 45.1). Естественная процедура проверки случайности состоит в сравнении наблюдаемого распределения с ожидаемым, которое дается в (45.14). Однако в случае небольших рядов возникает теоретическая трудность, связанная с тем, что длины фаз не незави-

симы, так что критерий согласия χ^2 здесь непосредственно использовать нельзя. Этот вопрос изучался Уоллисом и Муром (1941), которые пришли к заключению, что при разбиении длин фазы на три группы, $d = 1, 2, \geqslant 3$ (две степени свободы), статистика *) X^2 может быть использована в обычной форме с $\nu = 2{,}5$ при $X^2 \geqslant 6{,}3$. Для небольших значений можно использовать $6X^2/7$ в той же форме с $\nu = 2$.

Волфовиц (1944) и Левен (1952) показали, что распределение числа фаз стремится к нормальному, а Глейзберг (1945) затабулировал распределение этого числа для $n \leqslant 25$.

Пример 45.1

Рассмотрим таблицу 45.1 (урожайность ячменя). Данный ряд содержит 56 значений, но дважды (1906, 1907 и 1910, 1911) значения для соседних годов оказываются равными. Поскольку нас интересуют экстремальные точки и фазы, будем считать каждую такую пару точек за одну точку и, следовательно, понизим число членов до 54.

Если читатель отметит пики и ямы в таблице или подсчитает их на рис. 45.1, он найдет, что существует 35 экстремальных точек. Математическое ожидание в силу (45.3) равно $\frac{2}{3}(52) = 34\frac{2}{3}$. Это весьма близко к наблюденному значению, так что никакого другого критерия не требуется.

Распределение фаз оказывается следующим:

Длина фазы	Число наблюденных фаз	Число фаз, вычисленных теоретически по (45.14), (45.16)
1	23	21,25
2	7	9,17
3	4	2,59
Всего	34	33,67

Вновь дальнейшая проверка вряд ли является необходимой. Следуя этим критериям, мы должны были бы прийти к выводу, что изменения урожайности от года к году носят чисто случайный характер.

45.21 Если в качестве альтернативы выступает наличие тренда, характеристики критерия, основанного на экстремальных

*) Следуя практике последнего времени, автор, в отличие от более ранних традиционных обозначений, использует символ X^2 для обозначения статистики критерия χ^2. (*Прим. ред.*)

точках, оказываются довольно плохими. Позже (упражнение 45.4) мы увидим, что в некоторых случаях этот критерий будет обладать сравнительно с другими критериями нулевой эффективностью. Интуитивно это понятно, поскольку наличие экстремальных точек есть локальное свойство и на нем не сильно отразились бы последствия легкого тренда, если таковому был бы подвержен наш ряд. Если в качестве альтернативы взять цикличность, критерий становится много лучше. В случайных рядах средний интервал между двумя экстремальными точками (см. (45.10) и (10.14)) равен приблизительно 1,5 с дисперсией приблизительно $9/10n$. Так что критерий во всяком случае предписывает дальнейшее исследование, если ряд состоит из более чем 10 членов, а средний интервал между экстремальными точками больше или равен двум.

Мощность критерия при определенных альтернативах, касающихся «роста» и «снижения», исследовал Левен (1952).

Критерий, основанный на знаках разностей

45.22 Несколько более трудоемкий критерий состоит в подсчете числа положительных разностей первого порядка в ряде, иначе говоря, числа точек роста ряда. (Как и прежде, мы игнорируем точки, где не происходит ни увеличения, ни уменьшения значений.) Для ряда из n членов получаем $n - 1$ разностей. Как и прежде, определим величины

$$\begin{aligned} X_i &= 1, \quad u_{i+1} > u_i, \\ X_i &= 0, \quad u_{i+1} < u_i; \qquad i = 1, \ldots, (n-1). \end{aligned} \tag{45.20}$$

Тогда число точек роста, скажем c, равно

$$c = \sum_{i=1}^{n-1} X_i.$$

Для случайного ряда немедленно получаем, что

$$\mathsf{M}c = (n-1)\,\mathsf{M}X_i = \frac{1}{2}(n-1). \tag{45.21}$$

Точно так же

$$\begin{aligned} \mathsf{M}c^2 &= \mathsf{M}\Big\{\sum_{n-1} X_i^2 + 2\sum_{n-2} X_iX_{i+1} + \sum_{(n-2)(n-3)} X_iX_j\Big\} = \\ &= (n-1)\,\mathsf{M}X_i^2 + 2(n-2)\,\mathsf{M}(X_iX_{i+1}) + (n-2)(n-3)\,\mathsf{M}(X_iX_j) = \\ &= \frac{1}{2}(n-1) + 2(n-2)\,\mathsf{M}(X_iX_{i+1}) + \frac{1}{4}(n-2)(n-3) \qquad (45.22) \\ &\qquad (j \neq i,\ i+1). \end{aligned}$$

Чтобы вычислить $\mathbf{M}(X_iX_{i+1})$, рассмотрим перестановки трех элементов. Только в одном случае из шести получаем ненулевое значение X_iX_{i+1}. Отсюда и из (45.22) и (45.21) получаем, что

$$\mathbf{D}c = \frac{1}{2}(n-1) + \frac{1}{3}(n-2) + \frac{1}{4}(n-2)(n-3) - \frac{1}{4}(n-1)^2 = \\ = \frac{1}{12}(n+1). \qquad (45.23)$$

Распределение довольно быстро сходится к нормальному (см. упражнение 45.3). Оно было затабулировано Муром и Уоллисом (1943).

45.23 Очевидно, что критерий совершенно бесполезен, если альтернативой являются симметрические колебания, где число движений «вверх» приблизительно равно числу движений «вниз». Этот критерий в основном считается полезным при такой альтернативе, как тренд, особенно линейный тренд. Здесь он значительно лучше, чем критерий по экстремальным точкам, но много хуже других критериев, основанных на ранговых соотношениях. Эти критерии мы обсудим ниже.

Рассмотрим ряд с линейным трендом и случайным остаточным членом:

$$u_t = \alpha + \beta t + \varepsilon_t, \qquad (45.24)$$

где ε_t — нормальная величина с нулевым средним и единичной дисперсией. Последнее мы можем рассматривать как регрессию u_t по t в том специальном случае, когда t принимает равноотстоящие друг от друга значения 1, 2, ..., n. В ситуации регрессии нам следовало бы оценить β величиной

$$b = \frac{\sum (u_t - \bar{u}_t)(t - \bar{t})}{\sum (t - \bar{t})^2}. \qquad (45.25)$$

Эта оценка является несмещенной и имеет дисперсию

$$\mathbf{D}b = \frac{1}{\sum (t-\bar{t})^2} = \frac{12}{n(n^2-1)} \doteqdot \frac{12}{n^3}. \qquad (45.26)$$

Используем теперь асимптотическую относительную эффективность (см. 25.5—6) для того, чтобы сравнить другие состоятельные критерии с тем, что основан на b. Так как b несмещена, $[\partial \mathbf{M} b/\partial\beta]_{\beta=0} = 1$, то (25.16) переходит ($m = 1$) в соотношение

$$[\mathbf{M}'(b)]_{\beta=0}/(\mathbf{D}b)^{1/2} \sim (n^3/12)^{1/2}. \qquad (45.27)$$

Таким образом, для статистик b, величина δ, определенная в (25.16), принимает значение

$$\delta_b = \frac{3}{2}. \qquad (45.28)$$

Вычислим теперь δ для статистики, основанной на знаках разностей.

Рассмотрим индикаторы

$$\begin{aligned} H_{ij} &= 1, \quad u_i > u_j, \\ H_{ij} &= 0, \quad u_i < u_j; \end{aligned} \qquad (45.29)$$

причем будем полагать $H_{ii} = 1$. Математическое ожидание H_{ij} равно вероятности того, что H_{ij} равно единице и, поскольку $u_i - u_j$ представляет собой нормальную величину со средним $\beta(i-j)$ и дисперсией 2, указанная вероятность равна

$$\int_0^\infty \frac{1}{\sqrt{2\pi}\, 2^{1/2}} \exp\left\{-\frac{1}{4}[x - \beta(i-j)]^2\right\} dx = \\ = \frac{1}{\sqrt{2\pi}} \int_{\frac{1}{\sqrt{2}}\beta(i-j)}^\infty \exp\left\{-\frac{1}{2}y^2\right\} dy.$$

Следовательно,

$$\left[\frac{\partial}{\partial\beta}\mathbf{M}H_{ij}\right]_{\beta=0} = \frac{i-j}{\sqrt{2}} \cdot \frac{1}{\sqrt{2\pi}} = \frac{i-j}{2\sqrt{\pi}}. \qquad (45.30)$$

Для получения АОЭ настоящего критерия нам больше ничего не нужно, но для дальнейших целей мы сейчас вычислим еще несколько величин такого же типа.

Так как H_{ij} и H_{kl} независимы, если i, j, k, l не совпадают, имеем

$$\left[\frac{\partial}{\partial\beta}\mathbf{M}(H_{ij}H_{kl})\right]_{\beta=0} = \left[\mathbf{M}H_{ij}\frac{\partial}{\partial\beta}\mathbf{M}H_{kl}\right]_{\beta=0} + \\ + \left[\mathbf{M}H_{kl}\frac{\partial}{\partial\beta}\mathbf{M}H_{ij}\right]_{\beta=0} = \frac{1}{4\sqrt{\pi}}(i-j+k-l). \qquad (45.31)$$

Рассмотрим теперь $H_{ij}H_{jk}$. Поскольку $u_i - u_j$ и $u_j - u_k$ имеют совместное нормальное распределение с корреляцией $(-1/2)$, имеем

$$\mathbf{M}(H_{ij}H_{jk}) = \mathbf{P}(H_{ij} = 1, H_{jk} = 1) = \\ = \int_{-\beta(i-j)/\sqrt{2}}^\infty \int_{-\beta(j-k)/\sqrt{2}}^\infty \frac{1}{2\pi\sqrt{3/4}} \exp\left\{-\frac{2}{3}(x^2 - xy + y^2)\right\} dx\, dy,$$

и

$$\left[\frac{\partial}{\partial\beta}\mathbf{M}(H_{ij}H_{jk})\right]_{\beta=0} = \frac{i-j}{\sqrt{2}} + \frac{j-k}{\sqrt{2}}\int_0^\infty \frac{1}{2\pi}\exp\left\{-\frac{x^2}{2}\right\} dx = \frac{i-k}{4\sqrt{\pi}}. \qquad (45.32)$$

Аналогично

$$\left[\frac{\partial}{\partial\beta}\mathsf{M}(H_{ik}H_{jk})\right]_{\beta=0}=\frac{i+j-2k}{4\sqrt{\pi}}. \qquad (45.33)$$

Наконец, нам понадобится аналогичное выражение для случая, когда $u_{i+2}-u_{i+1}$ и $u_{i+1}-u_i$ имеют разные знаки. Оно совпадает с вероятностью того, что $(u_{i+2}-u_{i+1})\times(u_{i+1}-u_i)$ отрицательно, и, как легко видеть, равно

$$M=\int\limits_{-\infty}^{-\beta/\sqrt{2}}\int\limits_{-\beta/\sqrt{2}}^{\infty}\frac{1}{\sqrt{3/4}\,2\pi}\exp\left\{-\frac{2}{3}(x^2-xy+y^2)\right\}dx\,dy.$$

Отсюда получаем, что

$$\left[\frac{\partial M}{\partial\beta}\right]_{\beta=0}=0. \qquad (45.34)$$

Возвращаясь к критерию по знакам, получаем, что число положительных приращений в ряде (см. 45.20)) равно

$$c=\sum_{i=1}^{n-1}H_{i,\,i+1}.$$

Таким образом, из (45.30) получаем

$$\left[\frac{\partial\mathsf{M}c}{\partial\beta}\right]_{\beta=0}=\sum_{i=1}^{n-1}\frac{1}{2\sqrt{\pi}}=\frac{n-1}{2\sqrt{\pi}}. \qquad (45.35)$$

Вспоминая, что в силу (45.23), дисперсия равна $(n+1)/12$, имеем

$$\left[\frac{\partial\mathsf{M}c}{\partial\beta}\right]_{\beta=0}\Big/(\mathsf{D}c)^{1/2}\sim\left(\frac{3n}{\pi}\right)^{1/2}. \qquad (45.36)$$

Таким образом, для критерия по знакам величина δ, определенная в (25.16) т. 2, принимает значение

$$\delta_c=\frac{1}{2}. \qquad (45.37)$$

Сравнение (45.37) и (45.28) показывает, что c имеет нулевую асимптотическую относительную эффективность (см. 25.24)). Манн (1945а) дал нижнюю границу для мощности критерия. Стьюарт (1952) для 95-процентного уровня значимости получил таблицу мощности критерия с нормальной регрессией в качестве альтернативы.

Критерии, основанные на ранговой корреляции

45.24 Существует веское предположение, что мы улучшим наш критерий, если будем сравнивать все пары, а не только соседние, как в критерии по знакам. Рассмотрим для этого последовательность $u_1, \ldots, u_n$ и подсчитаем число пар, для которых $u_j>u_i$, $j>i$. Пусть это число равно P. Заметим, что всего существует $n(n-1)/2$ пар и математическое ожидание числа P для случайного ряда равно $n(n-1)/4$. Если P превышает последнее число, то это указывает на тенденцию к положительному тренду, в противном случае мы сталкиваемся с отрицательным трендом. Фактически исследуемая величина является простой линейной функцией от определенного в (31.23) т. 2 коэффициента ранговой корелляции*) τ между порядком величин во времени и их порядком по величине u. Для случайного ряда дисперсия τ известна. Если Q — дополнение к P, а именно, число пар у которых $u_j<u_i$, $j>i$, из (31.23) получаем, что

$$\tau=1-\frac{4Q}{n(n-1)}, \qquad (45.38)$$

и, учитывая (31.33—4), имеем

$$\mathsf{M}\tau=0, \qquad (45.39)$$

$$\mathsf{D}\tau=\frac{2(2n+5)}{9n(n-1)}. \qquad (45.40)$$

Распределение τ быстро сходится к нормальному (см. **31.26**).

45.25 Используя обозначения **45.23**, можно написать, что

$$Q=\sum_{i<j}^{n}H_{ij}.$$

В случае, когда альтернативой является линейный тренд (45.24) с нормальным остаточным членом, из (45.30) получаем

$$\left[\frac{\partial}{\partial\beta}\mathsf{M}Q\right]_{\beta=0}=\frac{1}{2\sqrt{\pi}}\sum_{i<j}^{n}(i-j)=-\frac{1}{2\sqrt{\pi}}\sum_{j=1}^{n}\left\{\frac{1}{2}j(j+1)-j\right\}=$$
$$=-\frac{1}{2\sqrt{\pi}}\cdot\frac{n(n^2-1)}{6}.$$

Из (45.40) имеем также, что

$$\mathsf{D}Q\sim\frac{n^3}{36},$$

так что

$$\left[\frac{\partial}{\partial\beta}\mathsf{M}Q\right]_{\beta=0}\Big/(\mathsf{D}Q)^{1/2}\sim\{n^3/(4\pi)\}^{1/2}. \qquad (45.41)$$

*) В другом контексте вместо статистики τ мы пишем t, но здесь это вызвало бы путаницу с временным параметром. Мы откажемся временно от обычного соглашения не использовать греческие буквы для выборочных значений.

Осуществляя подстановку из (45.41) и (45.27) в (25.27) для АОЭ τ (или Q) относительно регрессионной оценки имеем

$$A_{t,b} = \left\{ \frac{[\mathsf{M}'(t)]_{\beta=0}/(\mathsf{D}t)^{1/2}}{[\mathsf{M}'(\beta)]_{\beta=0}/(\mathsf{D}b)^{1/2}} \right\}^{\frac{1}{3/2}} = (3/\pi)^{1/3} = 0{,}98. \qquad (45.42)$$

Результат ранее получен в **31.38**. Статистика τ, таким образом, в этом случае весьма эффективна.

Манн (1945b), рассматривая τ-критерий для величин u_i, для которых $\mathsf{P}(u_i > u_j) = \frac{1}{2} + \varepsilon_{ij}$ при $i < j$ и которые подчинены еще ряду условий, получил (см. упражнение 31.8) условия несмещенности критерия и привел пример, в котором этот критерий является наиболее мощным.

45.26 Мы можем также для нашего случая вычислить ранговую корреляцию Спирмена r_S (**31.22** т. 2). Имеем (см. (31.40))

$$r_S = 1 - \frac{12V}{n(n^2-1)}, \qquad (45.43)$$

где

$$V = \sum_{i<j}^{n} (j-i) H_{ij}. \qquad (45.44)$$

Как и в (31.22), для случайных рядов получаем

$$\mathsf{M}r_S = 0,$$

$$\mathsf{D}r_S = \frac{1}{n-1},$$

так что

$$\mathsf{M}V = \sum_{i<j} (j-i) MH_{ij}, \qquad (45.45)$$

$$\mathsf{D}V \doteqdot \frac{n^5}{144}. \qquad (45.46)$$

Таким образом,

$$\left[\frac{\partial}{\partial\beta} \mathsf{M}V\right]_{\beta=0} = -\frac{1}{2\sqrt{\pi}} \sum_{i<j} (j-i)^2 = -\frac{n^2(n^2-1)}{24\sqrt{\pi}}, \qquad (45.47)$$

и, точно так же как и в (45.42), получаем

$$A_{r_S,\,b} = \left(\frac{3}{\pi}\right)^{1/3} = 0{,}98. \qquad (45.48)$$

Коэффициент Спирмена, таким образом, имеет ту же АОЭ, что и τ.

45.27 По сравнению со статистиками, основанными на знаках разностей или экстремальных точках, вычисление значений r_S и τ весьма утомительно. На практике значение r_S вычислять легче, чем τ, если использовать форму (31.21).

Пример 45.2

Не перегружая себя вычислениями, возьмем первые 25 членов из таблицы 45.2 (1813—1837 гг.). Сравнительные характеристики значений таковы:

Номер	Ранг	Квадрат разности	Номер	Ранг	Квадрат разности	Номер	Ранг	Квадрат разности
1	9	64	10	11	1	19	21	4
2	18	256	11	13	4	20	2	324
3	4	1	12	25	169	21	15	36
4	23	361	13	8	25	22	3	361
5	10	25	14	5	81	23	14	81
6	12	36	15	7	64	24	20	16
7	19	144	16	22	36	25	1	576
8	6	4	17	17	0			
9	24	225	18	16	4			2898

Итак,

$$V = 2898/2 = 1449,$$

$$r_S = 1 - (12)(1449)/15\,600 = -0{,}114.$$

Корреляция мала. Стандартная ошибка величины r_S равна $(n-1)^{-1/2}$, что приблизительно равно 0,2, так что наблюденное значение ничего не говорит нам о наличии тренда.

Кроме того, имеется 17 экстремальных точек, а соответствующее математическое ожидание равно $2 \cdot 23/3 = 15{,}7$, так что согласие почти полное.

45.28 Не вдаваясь в детали, можно упомянуть еще о двух критериях, представляющих некоторый интерес.

(а) *Критерий, основанный на «рекордных значениях»*. Наблюдение u_t будем называть верхним (нижним) рекордным значением *) (короче, рекордом), если он больше (меньше), чем все предыдущие наблюдения этого ряда. Число рекордов, появляющихся в процессе того, как мы наблюдаем значение ряда, дает нам статистику, которую можно сравнивать с соответствующей характеристикой случайного ряда. Прием использовался Фостером и Стьюартом (1954). Некоторые их результаты приведены в упражнениях 45.8—9. Если альтернативой является

*) Имеет хождение также термин «верхняя» или «нижняя», лестничная высота. (*Прим. перев.*)

тренд, критерий, основанный на рекордных значениях, оказывается более мощным, чем критерии, использующие знаки разностей или экстремальные точки, и значительно менее мощным, чем τ и r_S-критерии (Стьюарт, 1956, 1957).

(б) *Критерий, основанный на ранговой сериальной корреляции.* Сериальные коэффициенты корреляции, которые мы определим ниже в **45.32**, представляют весьма специальный тип статистик. Если ранги последовательности n величин, центрированные средним рангом $(n+1)/2$, обозначать d_i $(i=1, \ldots, n)$, коэффициент порядка k определяется как

$$r_k = \frac{\frac{1}{n-k}\sum_{i=1}^{n-k} d_i d_{i+k}}{\frac{1}{12}(n^2-1)}. \qquad (45.49)$$

Поскольку нас интересует лишь статистика критерия, можно использовать просто

$$W_k = \sum_{i=1}^{n-k} d_i d_{i+k}. \qquad (45.50)$$

Коэффициент W_k равен (умноженной на $(n-k)$) ковариации рангов тех членов, что отстоят друг от друга на k единиц времени. Для случайных рядов соответствующее математическое ожидание равно нулю. При проверке наличия тренда коэффициент имеет нулевую эффективность, если в качестве альтернативы взята нормальная линейная регрессия (см. Нётер (1950), Стьюарт (1954, 1956), хотя для проверки отсутствия тренда этот критерий предлагался еще Вальдом и Волфовицем (1943)).

45.29 Очевидно, что если ряд случаен, то мы можем его столь же успешно анализировать, если будем опускать некоторые члены, например, рассматривая лишь каждый второй член или каждый двенадцатый. С несколько иной точки зрения это означает, что если значения ряда регистрировались лишь периодически, через некоторые интервалы времени, то, несмотря на очевидную потерю информации, наши критерии и в этом случае остаются пригодными. То же верно и для составных (или агрегированных) рядов. Например, если на каждый год приходятся двенадцать месячных данных, причем каждое из месячных данных является членом случайного ряда, то и ряд годовых данных также случаен. С другой стороны, случайность ряда годовых данных не исключает наличия сезонности в рядах месячных данных.

45.30 Некоторые ряды (представьте, например, лезвие бритвы под микроскопом) содержат нерегулярную флуктуацию, которая выглядит примерно как флуктуация случайного ряда *) в том случае, когда интервал между наблюдениями мал. С другой стороны, у таких рядов «время» непрерывно, и возникает следующий вопрос: возможны ли случайные, но непрерывные ряды? В нашем понимании нет. Нам кажется, что некая существенная «дискретность» заложена в самой идее независимости последовательных наблюдений; непрерывность времени исключает независимость. Можно представить последовательность моментов «времени», каждому из которых соответствует значение некоторой случайной величины, и считать, что эти моменты времени сближаются, а дисперсия случайной величины уменьшается, так что полная вариация остается конечной. Однако не представляется возможным переход к пределу в том смысле, как это делается в математике при переходе от счетного множества точек ко всем точкам прямой.

45.31 Большинство рядов, если не все ряды, встречающиеся на практике, можно трактовать как непрерывные на первый взгляд, но дискретные при более детальном, или, образно говоря, мелком рассмотрении. Так, например, давление может мыслиться заданным в каждой точке, но с более строгой точки зрения различать сколь угодно близко отстоящие точки на самом деле нельзя, так как давление есть усредненный результат столкновений между молекулами. Длина нити может восприниматься как непрерывная величина, но на самом деле нить состоит из отдельных неделимых частиц. Учитывая это, мы можем говорить о непрерывных случайных рядах, но лишь понимая условность сказанного и всегда соблюдая при этом осторожность. Чтобы проверить такие ряды на «случайность», можно использовать наблюдения, произведенные через какой-то приемлемый интервал времени и применить к результатам один из обсуждавшихся выше критериев. Более тонкие методы мы обсудим позже, когда подойдем к спектральному анализу и коррелограммам.

Сериальные корреляции

45.32 В неслучайных рядах должен существовать тот или иной тип зависимости между членами. Одной из очень полезных характеристик такой зависимости является корреляция между последовательными членами. Заданы n значений $u_1, \ldots, u_n$. Се-

*) Напомним читателю, что под *случайным* рядом автор, как правило, подразумевает ряд, образованный последовательностью *независимых* наблюдений над некоторой случайной величиной, см. **45.15**. (*Прим. ред.*)

риальной корреляцией с единичным запаздыванием (или лагом 1) назовем величину

$$r_1=\frac{\frac{1}{n-1}\sum_{i=1}^{n-1}\left\{\left(u_i-\frac{1}{n-1}\sum_{i=1}^{n-1}u_i\right)\left(u_{i+1}-\frac{1}{n-1}\sum_{i=1}^{n-1}u_{i+1}\right)\right\}}{\left[\frac{1}{n-1}\sum_{i=1}^{n-1}\left\{u_i-\frac{1}{n-1}\sum_{i=1}^{n-1}u_i\right\}^2\frac{1}{n-1}\sum_{i=1}^{n-1}\left\{u_{i+1}-\frac{1}{n-1}\sum_{i=1}^{n-1}u_{i+1}\right\}^2\right]^{1/2}}. \tag{45.51}$$

Аналогично сериальной корреляцией с запаздыванием k (или лагом k) назовем корреляцию между отстоящими друг от друга на k единиц членами:

$$r_k=\frac{\frac{1}{n-k}\sum_{i=1}^{n-k}\left\{\left(u_i-\frac{1}{n-k}\sum_{i=1}^{n-k}u_i\right)\left(u_{i+k}-\frac{1}{n-k}\sum_{i=1}^{n-k}u_{i+k}\right)\right\}}{\left[\frac{1}{n-k}\sum_{i=1}^{n-k}\left\{u_i-\frac{1}{n-k}\sum_{i=1}^{n-k}u_i\right\}^2\frac{1}{n-k}\sum_{i=1}^{n-k}\left\{u_{i+k}-\frac{1}{n-k}\sum_{i=1}^{n-k}u_{i+k}\right\}^2\right]^{1/2}}. \tag{45.52}$$

На практике (а также для упрощения теоретических рассуждений) удобно эти определения до некоторой степени модифицировать. Вместо того чтобы центрировать первые $n-k$ величин u их средним, будем центрировать их средним всех n значений; аналогично поступим и с последними $n-k$ величинами. Точно так же, вместо отдельных дисперсий в знаменателе (45.52), будем использовать полную дисперсию по всему ряду. Итак, положив $\bar{u}=\sum_{i=1}^{n}u_i/n$, имеем

$$r_k=\frac{\frac{1}{n-k}\sum_{i=1}^{n-k}(u_i-\bar{u})(u_{i+k}-\bar{u})}{\frac{1}{n}\sum_{i=1}^{n}(u_i-\bar{u})^2}. \tag{45.53}$$

Это та форма, которую мы в основном будем использовать. Для рядов с не слишком малой длиной разница между (45.52) и (45.53) незначительна. В целях осторожности мы не будем использовать (45.53) для небольших рядов, так как здесь точность в оценивании необходима, иначе, например, могут появиться значения r_k большие единицы

45.33 Последовательность коэффициентов $r_0(=1)$, r_1, r_2, ... говорит нам многое о природе внутренней зависимости членов ряда. Совокупность коэффициентов называют *коррелограммой*: этот термин также употребляется для обозначения графика r_k как функции от k. Для случайных рядов (с точностью до выборочных ошибок) все коэффициенты, кроме r_0, равны нулю. Свойства коэффициентов для рядов других типов мы будем изучать в последующих главах. Отметим, что по определению $r_{-k}=r_k$.

45.34 Для определенных теоретических нужд, а также для облегчения вычислений малыми средствами определение (45.53) можно еще более упростить. Для коэффициента k-го порядка в числителе имеется $n-k$ членов. Предположим, что

$$u_{n+1}=u_1,\quad u_{n+2}=u_2,\ \ldots,\ u_{n+k}=u_k. \tag{45.54}$$

Возьмем теперь сумму произведений в числителе по всем членам. Тогда мы получим, что

$$r_k=\frac{\sum_{i=1}^{n}(u_i-\bar{u})(u_{i+k}-\bar{u})}{\sum_{i=1}^{n}(u_i-\bar{u})^2}. \tag{45.55}$$

Конечно, мы исказили данные, предполагая справедливость (45.54). Но если k мало по сравнению с n, искажения не столь велики. Коэффициент r_k, определенный в (45.55), называют *циклической* (*круговой*) сериальной корреляцией. Цель введения такого понятия станет ясной в **47.30** и главе 48.

45.35 В заключение этой главы, чтобы исключить путаницу, мы коснемся объектов, которые, в некотором смысле, также являются временными рядами, хотя и не того типа, что мы до сих пор рассматривали. Такие события, как несчастные случаи, остановки автомобилей перед светофорами, эпидемические вспышки и т. п., происходят время от времени и, таким образом, в течение некоторого периода составляют ряд событий. Интервалы между ними, как правило, не регулярны, но тем не менее имеют некоторую функцию распределения. Такие типы поведения изучаются теорией случайных процессов, причем в основном представляют интерес как раз интервалы между событиями, а не сами события. Отметим, что эта тематика резко отличается от тематики временных рядов, где изучаются изменения во времени и наблюдения происходят через определенные интервалы.

УПРАЖНЕНИЯ

45.1 Показать, что производящая функция моментов величины N_d/N (см. распределение фаз в **45.19**) при больших n дается выражением

$$3(e^{-\theta}-2e^{-2\theta}+e^{-3\theta})(g-1-e^{\theta})+\frac{5}{2}-\frac{3}{2}e^{-\theta},$$

где $g=\exp(e^{\theta})$.

Воспользовавшись этим, убедиться, что $\mu_1' = 1{,}5$, $\varkappa_2 = 0{,}56$, и показать, что $\varkappa_3 = 0{,}677$, $\varkappa_4 = 0{,}904$. Распределение, таким образом, отлично от нормального.

45.2 Показать, что если ряд «случаен», у распределения числа положительных разностей первого порядка из 45.22 нечетные семиинварианты равны нулю и

$$\varkappa_4 = -(n+1)/120.$$

45.3 Пусть $P_n(S)$ — число таких перестановок среди $n!$ перестановок n чисел, которые содержат S положительных разностей. Показать, что

$$P_n(S) = (S+1)\,P_{n-1}(S) + (n-S)\,P_{n-1}(S-1).$$

Получить отсюда рекуррентные соотношения для моментов:

$$\mathsf{M}_n(x^{2i+1}) = 0,$$

$$\mathsf{M}_n(x^{2i}) = \frac{n+1}{n}\,\mathsf{M}_{n-1}\left(x+\frac{1}{2}\right)^{2i} - \frac{2}{n}\,\mathsf{M}_{n-1}\left(x+\frac{1}{2}\right)^{2i+1},$$

где $x = S - \mathsf{M}S$.

Получить отсюда индукцией, что

$$\lim_{n\to\infty} \frac{\mu_{2i}(x)}{\{\mu_2(x)\}^i} = (2i-1)(2i-3)\ldots 3\cdot 1$$

и, следовательно, распределение, асимптотически нормально.

(Манн 1945а.)

45.4 Показать, что если в качестве альтернативы взять нормальную регрессию, у критерия, основанного на экстремальных точках со статистикой (45.2)

$$\{\mathsf{M}'p\}_{\beta=0} = 0, \qquad \{\mathsf{M}''p\}_{\beta=0} \neq 0,$$

и δ, определенное в (25.16), равно 1/4, так что АОЭ по сравнению с критерием, основанным на коэффициенте регрессии, равна нулю. К тому же приводит сравнение с критерием по знакам.

(Стюарт, 1954, 1956.)

45.5 Заметив, что в обозначениях 45.23, ранг r_i можно выразить в форме

$$r_i = \sum_{j=1}^{n} H_{ij},$$

показать, что если альтернативой является нормальная регрессия, то

$$\left[\frac{\partial}{\partial\beta}\,\mathsf{M}(r_i r_k)\right]_{\beta=0} = \frac{n(n+1)}{4\sqrt{n}}\{i+k-(n+1)\},$$

Следовательно, для статистики W из (15.50)

$$\left[\frac{\partial}{\partial\beta}\,\mathsf{M}W\right]_{\beta=0} = 0 \qquad \left[\frac{\partial^2}{\partial\beta^2}\,\mathsf{M}W\right]_{\beta=0} \sim \frac{n^5}{24\pi},$$

и

$$\mathsf{D}W \sim n^5/144,$$

так что $\delta_W = 5/4$ и АОЭ по сравнению с регрессионной оценкой равна нулю.

(Стьюарт, 1954, 1956.)

45.6 Показать, что если u_i, $i = 1, \ldots, n$ — значения случайных величин с плотностями $f_i(u)$, математическое ожидание числа c — числа точек роста ряда $u_1, \ldots, u_n$ равно

$$\mathsf{M}c \doteq \sum_{i=1}^{n-1} \int_{-\infty}^{\infty} f_{i+1}(u_{i+1}) \int_{-\infty}^{u_{i+1}} f_i(u_i)\,du_i\,du_{i+1}.$$

Пусть распределения равномерны с линейным трендом:

$$f_i(u_i) = 1, \qquad i\theta \leqslant u_i \leqslant i\theta + 1,$$

$$f_i(u_i) = 0 \qquad \text{в противном случае.}$$

Показать, что тогда

$$\lim_{n\to\infty} \frac{1}{n}\,\mathsf{M}c = \begin{cases} \frac{1}{2}\{1+\theta(2-|\theta|)\}, & -1 \leqslant \theta \leqslant 1, \\ 0, & \theta < -1, \\ 1, & \theta > 1. \end{cases}$$

(Левен, 1952.)

45.7 Как и в упражнении 45.6, показать, что для нормального распределения

$$f_i(u_i) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(u_i-\mu_i)^2\right.,$$

$$\lim_{n\to\infty} \frac{1}{n}\,\mathsf{M}c = \Phi(\theta/\sqrt{2}),$$

где

$$\Phi(t) = (2\pi)^{-1/2} \int_{-\infty}^{t} e^{-x^2/2}\,dx,$$

а тренд дается соотношением $\mu_i = (i-1)\theta$

(Левен, 1952, там же есть таблица.)

45.8. Для случайного ряда $x_1, \ldots, x_n$ определим

$$u_r = \begin{cases} 1, & \text{если } r\text{-е наблюдение является верхним рекордным значением,} \\ 0 & \text{в противном случае;} \end{cases}$$

$$l_r = \begin{cases} 1, & \text{если } r\text{-е наблюдение является нижним рекордным значением,} \\ 0 & \text{в противном случае;} \end{cases}$$

и $s_r = u_r + l_r$, $\quad d_r = u_r - l_r$. Пусть также

$$s = \sum_{r=2}^{n} s_r,$$

$$d = \sum_{r=2}^{n} d_r.$$

Суммирование начинается со второго элемента.

Показать, что если $p^{(r)}(s, d)$ — совместная функция частоты s и d для ряда из r наблюдений, то

$$p^{(r)}(s, d) = \left(1 - \frac{2}{r}\right) p^{(r-1)}(s, d) + \frac{1}{r} p^{(r-1)}(s-1, d-1) + \\ + \frac{1}{r} p^{(r-1)}(s-1, d+1); \quad p^{(1)}(0, 0) = 1.$$

Обозначая $g(\theta_1, \theta_2)$ производящую функцию, показать, что

$$g^{(n)}(\theta_1, \theta_2) = \frac{1}{n!} \prod_{r=0}^{n-2} (r + \theta_1\theta_2 - \theta_1/\theta_2).$$

Исследуя характеристические функции величин s и d, получить следующие формулы:

$$\varphi_s^{(n)}(t) = \frac{1}{n!} \prod_{r=0}^{n-2} (r + 2e^{it}),$$

$$\varphi_d^{(n)}(t) = \frac{1}{n!} \prod_{r=0}^{n-2} (r + 2\cos t).$$

Получить также совместную х. ф. и, используя формулы обращения, показать, что

$$p^{(n)}(s, d) = p_1^{(n)}(s)\, 2^{-s} \binom{s}{\frac{1}{2}(s+d)},$$

где p_1 — функция частоты величины s, определяемая соотношением

$$p_1^{(n)}(s) = \frac{1}{n!} 2^s u^{(n-2)}(n-s-1),$$

а $u^{(n)}(r)$ обозначает сумму произведений элементов всех выборок порядка r из последовательности $1, \ldots, n$.

(Фостер и Стьюарт, 1954.)

45.9 Показать, что в предыдущем примере

$$\mathsf{M}d = 0, \qquad \mathsf{D}d = 2\sum_{r=2}^{n} \frac{1}{r},$$

$$\mathsf{M}s = 2\sum_{r=2}^{n} \frac{1}{r}, \qquad \mathsf{D}s = 2\sum_{r=2}^{n} \frac{1}{r} - 4\sum_{r=2}^{n} \frac{1}{r^2}.$$

(Фостер и Стюарт, 1954.)

45.10 Показать, что r_{kc}, определенное в (45.55), не может быть больше единицы, а r_k, определенное в (45.53), может.

45.11 Показать, что для случайного ряда у коэффициента, определенного в (45.53),

$$\mathsf{M}r_k = -\frac{1}{n-1},$$

и, следовательно, r_k — смещенная оценка теоретической сериальной корреляции.

ГЛАВА 46

ВРЕМЕННЫЕ РЯДЫ. ТРЕНД И СЕЗОННОСТЬ

Определение тренда

46.1 В понятие тренда заложено то существенное обстоятельство, что движение, рассмотренное в течение достаточно долгого периода, представляется нам как бы сглаженным. Последнее означает, что, по крайней мере локально, компоненту, соответствующую тренду, мы можем представить в виде полинома от времени t. Таким образом, если задан ряд u_t, мы можем сначала стараться представить u_t в виде

$$u_t = a_0 + a_1 t + a_2 t^2 + \ldots + a_p t^p, \tag{46.1}$$

что даст нам представление о тренде. Беря p достаточно большим, мы можем получить такое точное представление любого конечного ряда, какое нам нужно. Вопрос, сколь большим выбирать p, решается отдельно для каждого случая.

Хотя полиномы и являются наиболее удобными с математической точки зрения функциями, ограничиваться только ими не следует. В принципе можно выбрать любую подходящую функцию времени, лишь бы она хорошо отражала тренд. Так, например, кривые, отражающие «чистый прирост», могут быть описаны экспоненциальными функциями, а кривые, отражающие рост численности населения и сходные с кривой на рис. 45.5, иногда представляют в виде «логистической» кривой типа

$$u(t) = \frac{k}{1 + e^{-\lambda t}}. \tag{46.2}$$

46.2 Используя метод наименьших квадратов, можно получить полином, наиболее хорошо (среди полиномов) отражающий эволюцию членов ряда (иначе говоря, осуществить «выравнивание» членов ряда методом наименьших квадратов). Такой полином, очевидно, дает кривую регрессии u_t по t. В то же время ясно, что для того, чтобы получить удовлетворительную кривую тренда, например для таких данных, как из таблицы 45.4 (поголовье овец), мы должны взять полином весьма высокого порядка или даже несколько более сложную функцию, чем полином. Однако повышение точности за счет усложнения функции

Показать, что если $p^{(r)}(s, d)$ — совместная функция частоты s и d для ряда из r наблюдений, то

$$p^{(r)}(s, d) = \left(1 - \frac{2}{r}\right) p^{(r-1)}(s, d) + \frac{1}{r} p^{(r-1)}(s-1, d-1) + \\ + \frac{1}{r} p^{(r-1)}(s-1, d+1); \quad p^{(1)}(0, 0) = 1.$$

Обозначая $g(\theta_1, \theta_2)$ производящую функцию, показать, что

$$g^{(n)}(\theta_1, \theta_2) = \frac{1}{n!} \prod_{r=0}^{n-2} (r + \theta_1\theta_2 - \theta_1/\theta_2).$$

Исследуя характеристические функции величин s и d, получить следующие формулы:

$$\varphi_s^{(n)}(t) = \frac{1}{n!} \prod_{r=0}^{n-2} (r + 2e^{it}),$$

$$\varphi_d^{(n)}(t) = \frac{1}{n!} \prod_{r=0}^{n-2} (r + 2\cos t).$$

Получить также совместную х. ф. и, используя формулы обращения, показать, что

$$p^{(n)}(s, d) = p_1^{(n)}(s)\, 2^{-s} \binom{s}{\frac{1}{2}(s+d)},$$

где p_1 — функция частоты величины s, определяемая соотношением

$$p_1^{(n)}(s) = \frac{1}{n!} 2^s u^{(n-2)}(n-s-1),$$

а $u^{(n)}(r)$ обозначает сумму произведений элементов всех выборок порядка r из последовательности $1, \ldots, n$.

(Фостер и Стьюарт, 1954.)

45.9 Показать, что в предыдущем примере

$$\mathsf{M}d = 0, \qquad \mathsf{D}d = 2\sum_{r=2}^{n} \frac{1}{r},$$

$$\mathsf{M}s = 2\sum_{r=2}^{n} \frac{1}{r}, \qquad \mathsf{D}s = 2\sum_{r=2}^{n} \frac{1}{r} - 4\sum_{r=2}^{n} \frac{1}{r^2}.$$

(Фостер и Стюарт, 1954.)

45.10 Показать, что r_{kc}, определенное в (45.55), не может быть больше единицы, а r_k, определенное в (45.53), может.

45.11 Показать, что для случайного ряда у коэффициента, определенного в (45.53),

$$\mathsf{M}r_k = -\frac{1}{n-1},$$

и, следовательно, r_k — смещенная оценка теоретической сериальной корреляции.

ГЛАВА 46

ВРЕМЕННЫЕ РЯДЫ. ТРЕНД И СЕЗОННОСТЬ

Определение тренда

46.1 В понятие тренда заложено то существенное обстоятельство, что движение, рассмотренное в течение достаточно долгого периода, представляется нам как бы сглаженным. Последнее означает, что, по крайней мере локально, компоненту, соответствующую тренду, мы можем представить в виде полинома от времени t. Таким образом, если задан ряд u_t, мы можем сначала стараться представить u_t в виде

$$u_t = a_0 + a_1 t + a_2 t^2 + \ldots + a_p t^p, \tag{46.1}$$

что даст нам представление о тренде. Беря p достаточно большим, мы можем получить такое точное представление любого конечного ряда, какое нам нужно. Вопрос, сколь большим выбирать p, решается отдельно для каждого случая.

Хотя полиномы и являются наиболее удобными с математической точки зрения функциями, ограничиваться только ими не следует. В принципе можно выбрать любую подходящую функцию времени, лишь бы она хорошо отражала тренд. Так, например, кривые, отражающие «чистый прирост», могут быть описаны экспоненциальными функциями, а кривые, отражающие рост численности населения и сходные с кривой на рис. 45.5, иногда представляют в виде «логистической» кривой типа

$$u(t) = \frac{k}{1 + e^{-\lambda t}}. \tag{46.2}$$

46.2 Используя метод наименьших квадратов, можно получить полином, наиболее хорошо (среди полиномов) отражающий эволюцию членов ряда (иначе говоря, осуществить «выравнивание» членов ряда методом наименьших квадратов). Такой полином, очевидно, дает кривую регрессии u_t по t. В то же время ясно, что для того, чтобы получить удовлетворительную кривую тренда, например для таких данных, как из таблицы 45.4 (поголовье овец), мы должны взять полином весьма высокого порядка или даже несколько более сложную функцию, чем полином. Однако повышение точности за счет усложнения функции

представляется нам несколько искусственным. В любом случае коэффициенты полинома, будучи основаны на моментах высокого порядка, были бы очень неустойчивы с точки зрения выборочных ошибок. Более практическое возражение, хотя никоим образом не маловажное, состоит в том, что добавление к ряду нового члена (допустим, например, что мы каждый год добавляем к ряду неких годовых данных новое данное) вновь влечет за собой работу по аппроксимации. Более того, характер тренда может меняться со временем. Поэтому, когда ряд не обладает ярко выраженным трендом, удобно пользоваться более простыми методами, которые мы опишем ниже.

Скользящие средние

46.3 Альтернативой к нахождению полинома, который представлял бы ряд целиком, служит поиск полинома, представляющего некоторую часть ряда и использование различных полиномов на различных этапах. Простейший метод, который к тому же служит основой для большинства других методов исследования тренда, состоит в следующем. Берутся первые n членов (n может быть любым), по этим членам строится полином степени p, не большей, чем $n-1$, и находится значение полинома в середине его области определения. Затем берутся n членов, начиная со второго и до $(n+1)$-го, процедура повторяется, и так далее, сдвигаясь на каждом этапе на один член вправо. Будем выбирать n нечетным пока для того, чтобы середина интервала была тем моментом времени, где наблюдение действительно производится. Иначе средняя точка попадет между двумя наблюдениями, и мы будем вынуждены использовать значение подогнанного под соответствующие члены полинома, отличное от значения в средней точке, что приведет к потере полезной для нас симметрии.

46.4 Итак, предположим, что число выбираемых членов ряда нечетно и равно $2m+1$. Не теряя в общности, можно обозначить члены символами $u_{-m}, u_{-(m-1)}, \ldots, u_0, \ldots, u_m$. Если мы решим производить выравнивание полиномом степени p, мы можем обычным образом определить коэффициенты полинома методом наименьших квадратов, решая уравнения

$$\frac{\partial}{\partial a_j}\sum_{t=-m}^{m}(u_t-a_0-a_1t-\ldots-a_pt^p)^2=0, \qquad j=0,\ 1,\ \ldots,\ p, \tag{46.3}$$

которые сводятся к уравнениям

$$\sum_{t=-m}^{m}t^ju_t-a_0\sum_{t=-m}^{m}t^j-a_1\sum_{t=-m}^{m}t^{j+1}-\ldots-a_p\sum_{t=-m}^{m}t^{j+p}=0. \tag{46.4}$$

Суммы $\sum_{t=-m}^{m}t^j$ зависят только от m. Следовательно, решая (46.4) относительно a_0, получим равенство

$$a_0=c_0+c_1u_{-m}+c_2u_{-(m-1)}+\ldots+c_{2m+1}u_m, \tag{46.5}$$

где все c зависят только от m и p и не зависят от u.

Поскольку, в соответствии с (46.1), u_0 положено равным величине a_0, которая в свою очередь задается формулой (46.5), то, следовательно, требуемое значение полинома в точке $t=0$ равно взвешенному среднему наблюденных значений. Веса зависят от того, какую часть ряда мы рассмотрели.

Таким образом, наш процесс определения линии тренда состоит в определении постоянных c (которые зависят от m и p и, следовательно, представляют нам две степени свободы при их выборе и вычислении с помощью (46.5) величин a_0), соответствующих всем кускам ряда, состоящим из $2m+1$ последовательных членов.

Если рассматриваются члены $u_x, \ldots, u_{2m+x}$, вычисляемое значение соответствует $t=m+x$. Для получения значений, соответствующих первым m и последним m членам, требуется дополнительная процедура.

Пример 46.1

Предположим, что нам дан некий ряд, и мы хотим построить кривую, дающую наилучшее приближение по семи точкам. Допустим, что нас вполне удовлетворит аппроксимация кубическим полиномом. Чему равны веса для скользящих средних?

У нас $m=3$, $p=3$ и полином равен

$$u_t=a_0+a_1t+a_2t^2+a_3t^3.$$

Выбрав в качестве начала координат $t=0$, из уравнений (46.4), учитывая, что при нечетных k $\sum t^k=0$, получаем, что

$$\left.\begin{aligned}\sum u &= 7a_0 + 28a_2,\\ \sum tu &= 28a_1 + 196a_3,\\ \sum t^2u &= 28a_0 + 196a_2,\\ \sum t^3u &= 196a_1 + 1588a_3.\end{aligned}\right\} \tag{46.6}$$

Отсюда для a_0 получаем

$$a_0=\frac{1}{21}\left\{7\sum u-\sum t^2u\right\}=$$

$$=\frac{1}{21}\{-2u_{-3}+3u_{-2}+6u_{-1}+7u_0+6u_1+3u_2-2u_3\}.$$

Правую часть последнего равенства можно символически записать в виде

$$\frac{1}{21}[-2,\ 3,\ 6,\ 7,\ 6,\ 3,\ -2],$$

или, когда, как в нашем случае, формула симметрична, в виде

$$\frac{1}{21}[-2,\ 3,\ 6,\ 7],$$

отмечая средний член полужирным шрифтом.

В качестве иллюстрации приведем простой пример. Предположим, что нам дан следующий временной ряд:

t:	1	2	3	4	5	6	7	8	9	10
u_t:	0	1	8	27	64	125	216	343	512	729

Значение тренда при $t = 4$ равно

$$a_0 = \frac{1}{21}\{(-2\times 0) + (3\times 1) + (6\times 8) + \ \ldots\ - (2\times 216)\} =$$
$$= \frac{1}{21}\cdot 567 = 27.$$

Значение тренда, таким образом, совпадает с действительным значением ряда в этой точке. Очевидно, что это и должно было быть, поскольку мы использовали кубический многочлен для выравнивания ряда

$$u_t = (t-1)^3.$$

Легко понять, что мы получили бы то же самое значение для a_0, если бы использовали квадратичные полиномы вместо кубических, так как в равенствах (46.6) a_0 не зависит от a_3. Вообще, случай нечетного p как бы включает в себя случай предыдущего (четного) p, так что нет нужды получать отдельные формулы для четных p.

46.5 Пусть $a_0[k]$ — вычисленное с помощью описанной выше процедуры значение скользящего среднего k последовательных членов. Выпишем соответствующие формулы вплоть до $p = 5$. Читатель может проверить их справедливость в качестве упражнения. Очевидно, что сумма коэффициентов в любой формуле равняется единице, ибо, если мы будем искать тренд у ряда, состоящего из единиц, то и получим единицу.

Второй и третий порядок

$$\left.\begin{array}{ll}
[5] & \frac{1}{35}\ [-3,\ 12,\ \mathbf{17}],\\
[7] & \frac{1}{21}\ [-2,\ 3,\ 6,\ 7],\\
[9] & \frac{1}{231}\ [-21,\ 14,\ 39,\ 54,\ \mathbf{59}],\\
[11] & \frac{1}{429}\ [-36,\ 9,\ 44,\ 69,\ 84,\ \mathbf{89}],\\
[13] & \frac{1}{143}\ [-11,\ 0,\ 9,\ 16,\ 21,\ 24,\ \mathbf{25}],\\
[15] & \frac{1}{1105}\ [-78,\ -13,\ 42,\ 87,\ 122,\ 147,\ 162,\ \mathbf{167}],\\
[17] & \frac{1}{323}\ [-21,\ -6,\ 7,\ 18,\ 27,\ 34,\ 29,\ 42,\ \mathbf{43}],\\
[19] & \frac{1}{2261}\ [-136,\ -51,\ 24,\ 89,\ 144,\ 189,\ 224,\ 249,\ 264,\ \mathbf{269}],\\
[21] & \frac{1}{3059}\ [-171,\ -76,\ 9,\ 84,\ 149,\ 204,\ 249,\ 284,\ 309,\ 324,\ \mathbf{329}].
\end{array}\right\}\quad (46.7)$$

Четвертый и пятый порядок

$$\left.\begin{array}{ll}
[7] & \frac{1}{231}\ [5,\ -30,\ 75,\ \mathbf{131}],\\
[9] & \frac{1}{429}\ [15,\ -55,\ 30,\ 135,\ \mathbf{179}],\\
[11] & \frac{1}{429}\ [18,\ -45,\ -10,\ 60,\ 120,\ \mathbf{143}],\\
[13] & \frac{1}{2431}\ [110,\ -198,\ -135,\ 110,\ 390,\ 600,\ \mathbf{677}],\\
[15] & \frac{1}{46189}\ [2145,\ -2860,\ -2937,\ -165,\ 3755,\ 7500,\ 10\,125,\ \mathbf{11\,063}],\\
[17] & \frac{1}{4199}\ [195,\ -195,\ -260,\ -177,\ 135,\ 415,\ 660,\ 825,\ \mathbf{883}],\\
[19] & \frac{1}{7429}\ [340,\ -225,\ -420,\ -290,\ 18,\ 405,\ 750,\ 1110,\ 1320,\ \mathbf{1393}],\\
[21] & \frac{1}{260\,015}\ [11\,628,\ -6460,\ -13\,005,\ -11\,220,\ -3940,\ 6378,\ 17\,655,\ 28\,190,\ 36\,660,\ 42\,120,\ \mathbf{44\,003}].
\end{array}\right\}\quad (46.8)$$

46.6 Иногда более удобно выписывать такие формулы на языке разностей рядов вида $\Delta^r u_t$, где

$$\Delta u_t = u_{t+1} - u_t. \tag{46.9}$$

Так, например,

$$\frac{1}{21}[-2, 3, 6, 7, 6, 3, -2] = u_t - \frac{1}{21}(9\Delta^4 + 9\Delta^5 + 2\Delta^6)\, u_{t-3}, \tag{46.10}$$

что, в частности, немедленно обнаруживает тот факт, что наше приближение тренда является точным, если члены ряда являются значениями кубического полинома.

Столь же хорошо можно представлять процесс в виде скользящих средних от разностей, что удобно для вычислений в том случае, когда величины разностей убывают с ростом их порядков или когда разности меньше, чем сами значения ряда. Например,

$$\frac{1}{21}[-2, 3, 6, 7] = u_t + \frac{1}{21}\{2\Delta^3 u_{t-3} + 3\Delta^3 u_{t-2} - 3\Delta^3 u_{t-1} - 2\Delta^3 u_t\} = \tag{46.11}$$

$$= u_t + \frac{1}{21}[2, 3, -3, 2]\, \Delta^3 u_{t-3} = \tag{46.12}$$

$$= u_t - \frac{1}{21}[2, 5, 2]\, \Delta^4 u_{t-3}. \tag{46.13}$$

Очевидно, мы можем выписывать подобные соотношения множеством различных способов; в частности, формула (46.13) удобна потому, что дает нам непосредственное выражение для «остатков».

Пример 46.2

Допустим, что мы хотим одну из формул (46.8), например относящуюся к выравниванию 11 точек полиномом пятой степени, переписать указанным выше способом.

Прежде всего из среднего члена формулы

$$\frac{1}{429}[18, -45, -10, 60, 120, \mathbf{143}]$$

вычтем единицу, так что получим

$$\frac{1}{429}[18, -45, -10, 60, 120, -\mathbf{286}]. \tag{46.14}$$

Сумма коэффициентов теперь должна равняться нулю. Обозначим U оператор сдвига такой, что

$$Uu_t = u_{t+1}. \tag{46.15}$$

Тогда

$$\Delta = U - 1. \tag{46.16}$$

Опуская делитель 429, можно записать скользящее среднее (46.14) в виде

$$18 - 45U - 10U^2 + 60U^3 + 120U^4 - 286U^5 + 120U^6 + \\ + 60U^7 - 10U^8 - 45U^9 + 18U^{10}.$$

Мы знаем, что это выражение является точным вплоть до разностей пятого порядка и, следовательно, $\Delta^6 = (U-1)^6$ можно вынести за скобку. Получаем

$$(U-1)^6(18U^4 + 63U^3 + 98U^2 + 63U + 18).$$

Первоначальный процесс можно (так как $U - 1 = \Delta$) записать в виде

$$u_t + \frac{1}{429}[18, 63, 98, 63, 18]\, \Delta^6 u_{t-5}. \tag{46.17}$$

46.7 Выпишем формулы (46.7) и (46.8) *) на языке разностей **).

Второй и третий порядок

$$\left.\begin{array}{ll}
[5] & u_3 - \frac{3}{35}[1]\,\Delta^4 u_1, \\
[7] & u_4 - \frac{1}{21}[2, 5, 2]\,\Delta^4 u_2, \\
[9] & u_5 - \frac{1}{231}[21, 70, 115, 70, 21]\,\Delta^4 u_3, \\
[11] & u_6 - \frac{1}{429}[36, 135, 280, \mathbf{385}]\,\Delta^4 u_4, \\
[13] & u_7 - \frac{1}{143}[11, 44, 101, 168, \mathbf{210}]\,\Delta^4 u_5, \\
[15] & u_8 - \frac{1}{1105}[78, 325, 790, 1435, 2100, \mathbf{2478}]\,\Delta^4 u_6, \\
[17] & u_9 - \frac{1}{323}[21, 90, 227, 434, 686, 924, \mathbf{1050}]\,\Delta^4 u_7, \\
[19] & u_{10} - \frac{1}{2261}[136, 595, 1540, 3045, 5040, 7266, \\
 & \qquad 9240, \mathbf{10\,230}]\,\Delta^4 u_8, \\
[21] & u_{11} - \frac{1}{3059}[171, 76, 2005, 4060, 6930, 10\,416, \\
 & \qquad 14\,070, 17\,160, \mathbf{18\,645}]\,\Delta^4 u_9.
\end{array}\right\} \tag{46.18}$$

*) Кендалл (1961а) показал, что при увеличении n и p числа в квадратных скобках правых частей формул стремятся к соответствующим значениям нормальной плотности.

**) Обозначения в формулах (46.18), (46.19) не согласуются с обозначениями, использованными ранее автором в формулах (46.10)—(46.13). В частности, в формулах (46.10)—(46.13) индекс времени при значении члена временного ряда, стоящего после операторов разностей Δ^r, определял начальный момент в соответствующей операции суммирования, а в формулах (46.18), (46.19) — среднюю временную точку. (*Прим. ред.*)

Четвертый и пятый порядок

$$
\left.\begin{aligned}
&[7]\quad u_4+\frac{5}{231}[1]\,\Delta^6u_1,\\
&[9]\quad u_5+\frac{5}{429}[3,\ 7]\,\Delta^6u_2,\\
&[11]\quad u_6+\frac{1}{429}[18,\ 63,\ \mathbf{98}]\,\Delta^6u_3,\\
&[13]\quad u_7+\frac{1}{2431}[110,\ 462,\ 987,\ \mathbf{1032}]\,\Delta^6u_4,\\
&[15]\quad u_8+\frac{1}{4689}[2145,\ 10\,010,\ 24\,948,\ 42\,273,\\
&\qquad\qquad \mathbf{51\,198}]\,\Delta^6u_5,\\
&[17]\quad u_9+\frac{1}{4199}[195,\ 975,\ 2665,\ 5148,\ 7623,\ \mathbf{8778}]\,\Delta^6u_6,\\
&[19]\quad u_{10}+\frac{1}{7429}[340,\ 1785,\ 5190,\ 10\,875,\ 18\,018,\\
&\qquad\qquad 24\,453,\ \mathbf{27\,258}]\,\Delta^6u_7,\\
&[21]\quad u_{11}+\frac{1}{260\,015}[11{,}628,\ 63\,308,\ 192\,423,\ 426\,258,\\
&\qquad\qquad 759\,003,\ 1\,135\,134,\ 1\,450\,449,\\
&\qquad\qquad \mathbf{1\,581\,294}]\,\Delta^6u_8.
\end{aligned}\right\}\quad(46.19)
$$

46.8 Поскольку наличие больших чисел в некоторых выражениях из (46.7) и (46.8) приводило к значительным трудностям при непосредственном использовании этих формул, было предложено несколько методов, упрощающих вычисления при определении линии тренда с помощью скользящих средних. Простейшим из этих методов, по-видимому, является метод итерированных усреднений.

Допустим, что мы вычисляем средние по четырем значениям при равных весах (очень простая операция), а затем вычисляем такие же средние по полученным ранее средним. Если первоначальный ряд обозначить u_t, первая операция приводит нас к ряду с членами типа

$$v_1=\frac{1}{4}(u_1+u_2+u_3+u_4),$$

а вторая операция — к ряду с членами типа

$$w_1=\frac{1}{4}(v_1+v_2+v_3+v_4)=$$
$$=\frac{1}{16}(u_1+2u_2+3u_3+4u_4+3u_5+2u_6+u_7).\qquad(46.20)$$

Символически это можно записать в виде

$$\left\{\frac{1}{4}[1,\ 1,\ 1,\ 1]\right\}^2=\frac{1}{16}[1,\ 2,\ 3,\ \mathbf{4}],\qquad(46.21)$$

или, вводя символ $[k]/k$ для простого арифметического среднего, в виде

$$\frac{1}{16}[4]^2=\frac{1}{16}[1,\ 2,\ 3,\ \mathbf{4}].\qquad(46.22)$$

Сравним теперь результаты с весами для среднего, полученными в примере 46.1 при выравнивании семи точек кубическим полиномом. Переходя к десятичным дробям, получаем для последних следующие значения:

$$-0{,}0952;\ 0{,}1429;\ 0{,}2857;\ \mathbf{0{,}3333},$$

а для весов из (46.20):

$$0{,}0625;\ 0{,}1250;\ 0{,}1875;\ \mathbf{0{,}2500}.$$

Эти две группы неодинаковы, но обнаруживают одинаковую тенденцию в переходе от одного веса к другому. Так что последняя группа может рассматриваться как приближение к первой. (Вскоре мы получим лучшую аппроксимацию, но она будет служить нам лишь для иллюстрации.)

Процедуру последовательных операций, которая приводит к (46.21), осуществить значительно легче, чем однократный процесс вычисления взвешенных средних из примера 46.1. Вообще, если можно найти средние с простыми целыми весами, предпочтительно единичными, которые затем приведут нас к аппроксимациям для более сложных весов в однократных средних, как правило, лучше выбрать итерационный процесс.

46.9 Используя обозначения конечных разностей, запишем:

$$\delta u_t=u_{t+\frac{1}{2}}-u_{t-\frac{1}{2}}.\qquad(46.23)$$

Для второй «центральной» разности δ^2u_t имеем

$$\delta^2u_t=(u_{t+1}-u_t)-(u_t-u_{t-1})=\left(U-2+U^{-1}\right)u_t.\qquad(46.24)$$

Записывая U в виде

$$U=\exp(2i\varphi),\qquad(46.25)$$

получаем следующую символическую запись:

$$\delta^2=\exp(2i\varphi)-2+\exp(-2i\varphi)=-4\sin^2\varphi.\qquad(46.26)$$

Тогда, учитывая, что члены с $\sin 2j\varphi$ пропадают, имеем

$$\sum_{j=-m}^{m}u_j=\sum_{j=-m}^{m}U^ju_0=\left(1+2\sum_{j=1}^{m}\cos 2j\varphi\right)u_0=\frac{\sin(2m+1)\varphi}{\sin\varphi}u_0.\qquad(46.27)$$

Таким образом,

$$\frac{1}{k}[k]u_0 = \frac{1}{k}\frac{\sin k\varphi}{\sin\varphi}u_0 =$$
$$= \frac{1}{k}\left\{k - \frac{k(k^2-1)}{3!}\sin^2\varphi + \frac{k(k^2-1)(k^2-3^2)}{5!}\sin^4\varphi - \ldots\right\}u_0 =$$
$$= u_0 + \frac{k^2-1}{2^2 3!}\delta^2 u_0 + \frac{(k^2-1)(k^2-3^2)}{2^4 5!}\delta^4 u_0 + \ldots \qquad (46.28)$$

В этой интересной формуле арифметическое среднее представлено как функция от среднего члена u_0 и его центральных разностей.

Если теперь ряд приближенно представить в виде значений кубического полинома, так что четвертые разности становятся равными нулю, и взять u_0 в качестве среднего члена, можно получить, что

$$\frac{1}{k}[k]u_0 = u_0 + \frac{k^2-1}{24}\delta^2 u_0. \qquad (46.29)$$

Это равенство в любом случае будет верно с точностью до третьих разностей. Аналогично для двухкратной итерации с той же степенью точности получаем, что

$$\frac{1}{k_1 k_2}[k_1][k_2]u_0 = u_0 + \frac{1}{24}(k_1^2 + k_2^2 - 2)\delta^2 u_0, \qquad (46.30)$$

и т. д. Мы используем эти результаты, чтобы получить две формулы, весьма широко используемые в статистике страхования для «градации» рядов. Последний процесс весьма сходен с выделением линии тренда.

Пример 46.3. Формула Спенсера для 15 точек

Рассмотрим три последовательных усреднения с равными весами:

$$\frac{1}{80}[4][4][5]u_0 = u_0 + \frac{1}{24}(4^2 + 4^2 + 5^2 - 3)\delta^2 u_0 = u_0 + \frac{9}{4}\delta^2 u_0.$$

Умножая на $1 - 9\delta^2/4$, получаем с точностью до разностей третьего порядка, что

$$u_0 = \frac{1}{80}[4]^2[5]\left(1 - \frac{9}{4}\delta^2\right)u_0.$$

Следуя (46.24), заменяем δ^2 выражением [1, —2, 1] и получаем, что

$$u_0 = \frac{1}{320}[4]^2[5][-9, 22, -9].$$

Теперь, не меняя порядок аппроксимации, можно добавить члены, содержащие δ^4 или центральные разности более высокого порядка, так, чтобы упростить до некоторой степени числовые коэффициенты. Добавим к множителю [—9, 22, 9] член $-3\delta^4 =$ $= [-3, 12, -18, 12, -3]$. В результате получим [—3, 3, 4, 3, —3] и, следовательно,

$$u_0 = \frac{1}{320}[4]^2[5][-3, 3, 4, 3, -3]. \qquad (46.31)$$

Это формула Спенсера для 15 точек. Она охватывает 15 последовательных членов; окончательно веса таковы:

$$\frac{1}{320}[-3, -6, -5, 3, 21, 46, 67, \mathbf{74}]. \qquad (46.32)$$

Пример 46.4. Формула Спенсера для 21-й точки

Сходным образом находим, что

$$\frac{1}{175}[5]^2[7] = 1 + 4\delta^2,$$

что дает нам с точностью до разностей третьего порядка

$$u_0 = \frac{1}{175}[5]^2[7][-4, 9, -4].$$

Добавим к множителю [—4, 9, —4] выражение

$$-3\delta^4 - \frac{1}{2}\delta^6 = [-3, 12, -18] + \left[-\frac{1}{2}, 3, -7\frac{1}{2}, \mathbf{10}\right],$$

что даст нам представление

$$u_0 = \frac{1}{175}[5]^2[7]\left[-\frac{1}{2}, 0, \frac{1}{2}, \mathbf{1}\right] = \frac{1}{350}[5]^2[7][-1, 0, 1, \mathbf{2}]. \qquad (46.33)$$

Это формула Спенсера для 21-й точки.

46.10 В наши дни простота вычислений не столь важна, как раньше, и в определенных ситуациях таких приближений следует избегать. Первоначальные формулы (46.7)—(46.8) дают нам наилучшее выравнивание при заданном числе охватываемых точек и порядке полинома. Иначе говоря, при заданных m и p сумма квадратов весовых коэффициентов из этих формул будет минимальной. При вычислении скользящих средних ряда, состоящего из полиномиального тренда степени p и случайного остаточного члена ε_t (причем распределение ε_t одно и то же для всех t) остаточная сумма квадратов равна

$$\sum(c_0\varepsilon_j + c_1\varepsilon_{j+1} + \ldots + c_{2m}\varepsilon_{j+2m})^2.$$

Поэтому математическое ожидание выборочной дисперсии остатков пропорциональна

$$\sum_{j=0}^{2m} c_j^2, \tag{46.34}$$

и, следовательно, минимальна в классе весов, соответствующих одному и тому же порядку полинома p и числу m.

Краевые эффекты

46.11 Метод скользящих средних в том виде, в каком мы его изложили, обладает явными свойствами симметрии. Недостатком его является то, что он не дает значения тренда для первых m и последних m членов ряда. Как правило, то, что мы отказываемся от нескольких значений, в начале не ведет к большим потерям, но отсутствие значений тренда в конце представляет большую помеху, особенно если мы хотим провести экстраполяцию на будущее. Понимая, что значения тренда в конце не столь устойчивы как в середине, мы можем заполнять образовавшуюся брешь различными способами. Метод, иллюстрируемый в следующем примере, по-видимому, не сложнее, чем остальные.

Пример 46.5

Снова рассмотрим формулу, используемую в примере 46.1:

$$\frac{1}{21}[-2, 3, 6, 7].$$

Мы получили ее при выравнивании кубическим полиномом, но использовали последний лишь для определения среднего по семи членам. Непонятно, почему бы нам не использовать его и для трех последних членов крайней семерки. Чтобы сделать это, нам, однако, надо решать уравнения (46.6) относительно a_1, a_2, a_3 так же, как относительно a_0. Находим

$$a_1 = \frac{1}{1512}\left\{397 \sum tu - 49 \sum t^3 u\right\},$$
$$a_2 = \frac{1}{84}\left\{-4 \sum u + \sum t^2 u\right\},$$
$$a_3 = \frac{1}{216}\left\{-7 \sum tu + \sum t^3 u\right\}.$$

Подставляя эти результаты в полином, получаем в обычных обозначениях

$$u_t = \frac{1}{21}[-2, 3, 6, 7, 6, 3, -2] + \frac{t}{252}[22, -67, -58, 0, 58, 67, -22] + \frac{t^2}{84}[5, 0, -3, -4, -3, 0, 5] + \frac{t^3}{36}[-1, 1, 1, 0, -1, -1, 1]. \tag{46.35}$$

Таким образом, например, при $t = 1, 2, 3$ выражения сводятся к следующим:

$$u_1 = \frac{1}{42}[1, -4, 2, 12, 19, 16, -4], \tag{46.36}$$

$$u_2 = \frac{1}{42}[4, -7, -4, 6, 16, 19, 8], \tag{46.37}$$

$$u_3 = \frac{1}{42}[-2, 4, 1, -4, -4, 8, 39]. \tag{46.38}$$

Например, если бы последняя семерка членов ряда состояла из чисел 0, 1, 8, 27, 64, 125, 216, четвертый член тренда (следующий за средним членом, где $t = 0$) был бы равен

$$\frac{1}{42}[0 - 4 + 16 + 324 + 1216 + 2000 - 864] = 64.$$

Следующий член равен

$$\frac{1}{42}[0 - 7 - 32 + 162 + 1024 + 2375 + 1728] = 125.$$

Читатель может убедиться, что последний член равен 216. Эти результаты точны, конечно, потому, что мы выравнивали кубическим полиномом последовательность кубов.

Здесь следует отметить одно интересное явление. В формулах (46.35)—(46.38), так же как и в формуле для u_0, сумма коэффициентов равна единице. Но сами коэффициенты становятся все более и более неодинаковыми, так что сумма их квадратов растет по мере перехода от u_0 к u_3. Дисперсия любого случайного остаточного члена, таким образом, также растет, отражая тот факт, что при переходе от центра последовательности, которую мы выравниваем, к ее концу, полиномы становятся менее «надежными». Суммы квадратов коэффициентов в этом случае равны

u_0: 0,3333; u_1: 0,4524; u_2: 0,4524; u_3: 0,9286.

Суммы квадратов тем не менее растут не монотонно. Некоторые дальнейшие результаты приведены в упражнениях 46.1 и 46.5.

Коэффициенты и их суммы квадратов для $p \leqslant 5$ и $n \leqslant 25$ затабулированы Коуденом (1962).

46.12 Результаты, подобные тем, что сейчас обсуждались, могут быть также легко получены, если воспользоваться ортогональными полиномами (28.18 т. 2).

Пусть $n = 2m + 1$. Для первых четырех полиномов находим

$$\left.\begin{aligned}
\varphi_0(t) &= 1,\\
\varphi_1(t) &= \lambda_1 t,\\
\varphi_2(t) &= \lambda_2\left\{t^2 - \frac{1}{3}m(m+1)\right\},\\
\varphi_3(t) &= \lambda_3\left\{t^3 - \frac{1}{5}t(3m^2+3m-1)\right\},\\
\varphi_4(t) &= \lambda_4\left\{t^4 - \frac{1}{7}t^2(6m^2+6m-5) + \frac{3}{35}(m-1)m(m+1)(m+2)\right\}.
\end{aligned}\right\} \tag{46.39}$$

Если теперь, например,

$$u_t = b_0 + b_1\varphi_1 + b_2\varphi_2 + b_3\varphi_3, \tag{46.40}$$

мы сразу получаем, что

$$u_0 = b_0 - \frac{1}{3}\lambda_2 m(m+1)b_2. \tag{46.41}$$

Кроме того, из таблиц можно получить значения λ и $\sum\varphi_j^2$. В силу свойства ортогональности имеем

$$b_0 = \frac{\sum u_t\varphi_0}{\sum\varphi_0^2} = \frac{\sum u_t}{2m+1}, \tag{46.42}$$

$$b_2 = \frac{\sum u_t\varphi_2}{\sum\varphi_2^2}. \tag{46.43}$$

Например, если кубическим полиномом выравниваются семь значений ($m = 3$), из (46.42), (46.43) и таблиц получаем

$$b_0 = \frac{1}{7}\sum u_t, \qquad b_2 = \frac{1}{84}\sum(t^2-4)u_t.$$

Следовательно, в силу (46.40)

$$u_0 = \frac{1}{7}\sum u_t - \frac{1}{21}\left(\sum t^2u_t - 4\sum u_t\right) = \frac{1}{21}\left(7\sum u_t - \sum t^2u_t\right),$$

как и в примере 46.1. Используя полиномы таким же образом, можно получить и «краевые» значения, которые вычислялись в **46.11**.

46.13 До сих пор мы ничего не говорили о том, из каких соображений следует выбирать длину интервала, по которому берется скользящее среднее (т. е. $2m + 1$) и степень полинома p, с помощью которого это среднее вычисляется. В действительности никаких простых критериев здесь нет, в основном по той причине, что многое зависит от того, почему мы хотим выделить тренд, или, ставя вопрос несколько иначе, что представляет собой модель, определяющая разбиение ряда на компоненты. Если мы главным образом заняты общей тенденцией данных к тренду и не интересуемся более мелкими остаточными эффектами, может оказаться более пригодным один тип скользящих средних. Если же мы хотим выделить тренд для того, чтобы изучить оставшиеся компоненты, такой тип скользящих средних может оказаться совершенно неподходящим. Наконец, в ряде ситуаций может вообще возникнуть вопрос о целесообразности исключения тренда методом скользящих средних. Поэтому, для того, чтобы иметь основания обсуждать выбор подходящей методики для нахождения тренда, следует сначала изучить, как отражаются наши процедуры на оставшихся компонентах.

Влияние процедуры выделения тренда методом скользящих средних на остальные компоненты

46.14 Таблица 46.1 содержит результаты применения формулы Спенсера для 21 точки к одному искусственно построенному ряду. Последний получен добавлением случайной компоненты к кубическому полиному. (Мы использовали формулу Спенсера, а не одну из формул (46.7), потому что здесь можно увидеть и эффект последовательного простого усреднения.) Конкретно, пусть

$$u_t = (t-26) + \frac{1}{10}(t-26)^2 + \frac{1}{100}(t-26)^3 + \varepsilon_t. \tag{46.44}$$

Значения составляющей ε_t брались из таблицы случайных чисел и представляют собой значения случайной величины, принимающей с равными вероятностями все целые значения от 0 до 99. Различные столбцы таблицы иллюстрируют процесс выравнивания. Отметим, между прочим, что для таких коротких рядов удобно оставлять наиболее трудные вычисления на конец, так как там объем их существенно меньше.

Мы знаем, что формула Спенсера кубы «улавливает» точно, поэтому, вычитая тренд из первоначального ряда, мы (с точностью до округления систематической компоненты до ближайшей единицы) должны прийти к полному исключению систематической компоненты и остаться лишь со случайной компонентой. Однако это не так, в чем можно убедиться, сравнивая столбцы (2) и (9) таблицы 46.1 и помня при этом, что элементы последнего столбца включают в себя и число 49,5, равное среднему случайной компоненты. Причину далеко искать не надо: на самом деле мы вычисляли скользящие средние и случайных элементов, выделяя «линию тренда» и среди них.

Результат применения формулы Спенсера для 21-й точки только к случайным элементам ε_t приведен в столбце (11). Если

Таблица 46.1

Ряд, полученный по формуле (46.44). Линия тренда определяется формулой Спенсера для 21-й точки

(1) t	(2) Кубический член	(3) ε_t	(4) u_t	(5) [5] u_t	(6) [5] (5)	(7) [7] (6)	(8) [−1, 0, 1, 2, ...] (7)	(9) $\frac{(8)}{350}$	(10) Отклонение $u_t-(9)$	(11) «Выравнивание» ε_t
1	−119	23	−96							
2	−105	15	−90							
3	−92	75	−17	−246						
4	−80	48	−32	−209						
5	−70	59	−11	−87	−572					
6	−60	1	−59	−42	−241					
7	−51	83	32	12	162					
8	−44	72	28	85	413	2 233				
9	−37	59	22	194	670	3 801				
10	−31	93	62	164	844	5 120				
11	−26	76	50	215	957	5 984	14 352	41	9	67
12	−22	24	2	186	996	6 642	15 470	44	−42	66
13	−18	97	79	198	1 078	7 041	15 814	45	34	63
14	−15	8	−7	233	1 026	7 145	15 676	45	−52	60
15	−12	86	74	246	1 071	7 038	14 978	43	31	55
16	−10	95	85	163	1 069	6 934	14 166	40	45	51
17	−8	23	15	231	984	6 709	13 379	38	−23	47
18	−7	3	−4	196	850	6 535	12 703	36	−40	43
19	−6	67	61	112	892	6 408	12 169	35	26	40
20	−5	44	39	148	853	6 363	12 102	35	4	39
21	−4	5	1	205	852	6 446	12 279	35	−34	39
22	−3	54	51	192	944	6 611	12 676	36	15	39
23	−2	55	53	195	1 024	6 769	13 228	38	15	40
24	−2	50	48	204	1 031	7 052	13 857	40	8	41
25	−1	43	42	228	1 015	7 353	14 508	41	1	42
26	0	10	10	212	1 050	7 610	15 120	43	−33	43
27	1	74	75	176	1 136	7 923	15 634	45	30	44
28	2	35	37	230	1 153	8 249	16 251	46	−9	44
29	4	8	12	290	1 201	8 607	17 002	49	−37	45
30	6	90	96	245	1 337	9 019	17 717	51	45	44
31	9	61	70	260	1 357	9 424	18 499	53	17	44
32	12	18	30	312	1 373	9 870	19 307	55	−25	43
33	15	37	52	250	1 462	10 429	20 159	58	−6	42
34	20	44	64	306	1 541	10 989	21 133	60	4	41
35	24	10	34	334	1 599	11 679	22 417	64	−30	39
36	30	96	126	339	1 760	12 539	23 797	68	58	38
37	36	22	58	370	1 897	13 529	25 737	74	−16	37
38	44	13	57	411	2 047	14 699	27 955	80	−23	36
39	52	43	95	443	2 233	16 060	30 456	87	8	35
40	61	14	75	484	2 452	17 570	33 334	95	−20	34
41	71	87	158	525	2 711	19 353	36 716	105	53	34
42	83	16	99	589	2 960	21 394				
43	95	3	98	670	3 270	23 690				
44	109	50	159	692	3 680	26 255				

Продолжение табл. 46.1

(1) t	(2) Кубический член	(3) ε_t	(4) u_t	(5) [5] u_t	(6) [5] (5)	(7) [7] (6)	(8) [−1, 0, 1, 2, ...] (7)	(9) $\frac{(8)}{350}$	(10) Отклонение $u_t-(9)$	(11) «Выравнивание» ε_t
45	124	32	156	794	4088					
46	140	40	180	935	4529					
47	158	43	201	997	5017					
48	177	62	239	1111						
49	198	23	221	1180						
50	240	50	270							
51	244	5	249							

бы метод был совершенен, мы должны были бы в этом столбце получить числа 49,5, т. е. средние ε_t за выбросом нерегулярных выборочных эффектов, однако наблюденные значения не только отклоняются от среднего, но и делают это систематически, совершая некоторое колебательное движение, которое процедура выделила как часть тренда *).

46.15 Обнаруженный эффект существен, особенно если мы исключаем тренд для того, чтобы сконцентрировать внимание на колебаниях. Перейдем к более детальному рассмотрению.

Предположим, что нам дан ряд, состоящий из трех частей: тренда $x_1(t)$, «колебательного» члена $x_2(t)$ и случайного элемента $x_3(t)$, т. е.

$$u_t = x_1 + x_2 + x_3. \tag{46.45}$$

Выделим тренд методом скользящих средних, обозначая эту операцию символом T; имеем

$$Tu_t = Tx_1 + Tx_2 + Tx_3. \tag{46.46}$$

Предположим, что наш метод определения тренда совершенен в том смысле, что $Tx_1 = x_1$. Тогда, вычитая (46.46) из (46.45), для того чтобы исключить тренд, находим, что

$$u_t - Tu_t = x_2 - Tx_2 + x_3 - Tx_3. \tag{46.47}$$

Сейчас нас интересует то обстоятельство, что члены Tx_2 и Tx_3 в (46.47) могут исказить истинные колебания остаточного ряда и индуцировать ложное колебательное движение.

*) Эффект якобы систематического колебания случайной компоненты относительно среднего связан с законом арксинуса (см., например, В. Феллер, Введение в теорию вероятностей и ее приложения, ИЛ, 1955, т. 1, гл. III). (*Прим. перев.*)

46.16 Рассмотрим простой случай, когда член x_2 синусоидален, иначе, равен $\sin(\alpha+\lambda t)$, где t — целое. Так как

$$\sum_{t=1}^{k}\sin(\alpha+\lambda t)=\frac{\sin\frac{1}{2}k\lambda}{\sin\frac{1}{2}\lambda}\sin\left\{\alpha+\frac{1}{2}(k+1)\lambda\right\}, \qquad (46.48)$$

простое скользящее среднее k последовательных членов приведет к синусоидальному ряду с тем же периодом и фазой, что и у первоначального, но с амплитудой, равной первоначальной, умноженной на

$$\frac{1}{k}\,\frac{\sin\frac{1}{2}k\lambda}{\sin\frac{1}{2}\lambda}. \qquad (46.49)$$

q-кратная итерация приведет к умножению амплитуды на q-ю степень выражения (46.49).

Таким образом, член Tx_2 будет мал, если k велико и q велико или если $k\lambda/2$ кратно π, т. е. если длина отрезка усреднения будет равна периоду колебаний. Но если λ мало и $k\lambda$ мало, амплитуда изменяется очень мало и член x_2-Tx_2 становится очень мал, т. е. процедура вычисления скользящих средних практически уничтожает член x_2. В том случае, когда $k\lambda$ мало, длина отрезка усреднения мала по сравнению с периодом колебаний гармонического члена, т. е. здесь мы как бы имеем дело с медленными колебаниями. Этого результата можно было ожидать. Медленные колебания скользящим усреднением воспринимаются как тренд и соответственно исключаются. В общем случае скользящее усреднение будет выделять частые колебания и гасить медленные. К тому же, если длина отрезка усреднения немного больше, чем период, множитель (46.49) может быть отрицательным, и, следовательно, выделение тренда может даже несколько преувеличить истинные колебания.

Не столь легко выделить эффекты искажения точности при скользящем усреднении в том случае, когда веса неодинаковы и члены не являются гармоническими, но очевидно, что примерно такого же рода ситуации могут возникнуть и здесь.

46.17 Рассмотрим теперь влияние процедуры вычисления простых скользящих средних (т. е. усреднения с равными весами) на остаточный член x_3, который мы будем полагать случайной величиной ε_t с дисперсией v. Для члена Tx_3 имеем

$$Tx_3=\frac{1}{k}\sum_{-[k/2]}^{[k/2]}\varepsilon_{t+j}, \qquad (46.50)$$

где $[k/2]$ — наибольшее из целых, не превосходящих $k/2$. Последовательные величины ε_t независимы, но последовательные величины Tx_3 уже не являются таковыми, так как $Tx_3(a)$ и $Tx_3(b)$ зависят от $k-a+b$ общих величин ε и коррелированы, если $a-b<k$. Следовательно, ряд Tx_3 является значительно более «гладким», чем x_3, а если мы перейдем к дальнейшим усреднениям, то соответствующий ряд станет еще глаже. Мы уже видели в таблице 46.1 пример такого эффекта и встретим еще примеры в дальнейшем.

46.18 Применение скользящего усреднения к случайным рядам приводит к рядам, имеющим тенденцию к периодичности в колебаниях. Это обеспечивается тем, что здесь веса таковы, что порождают положительную коррелированность между членами нового ряда. Последнее всегда имеет место, если скользящее усреднение используется для выделения тренда. Мы назовем этот эффект эффектом Слуцкого — Юла, по имени двух статистиков, которые (независимо) детально изучали этот эффект.

Вновь образованные ряды нерегулярны в смысле цикличности, иначе говоря, их «пики» и «ямы» не появляются через равные временные интервалы, а амплитуда колебаний существенно меняется (хотя в главе 49 мы докажем теорему Слуцкого, показывающую, что определенные типы многократно примененных процедур усреднения приводит к синусоидальным кривым). Тем не менее колебательные движения подобного типа удивительно сходны с некоторыми типами движений, встречающихся на практике, особенно в экономике, и мы позже рассмотрим этот вопрос более детально. Сейчас же наша цель заключается в изучении того, в какой мере обсуждаемые здесь эффекты могут быть порождены процедурой выделения тренда. Нам необходимо это для того, чтобы быть уверенным, что колебательное движение не было привнесено в свободный от тренда ряд в результате наших собственных арифметических операций над членами ряда.

46.19 Для этой цели мы рассмотрим период и дисперсию у рядов, образовавшихся в результате эффекта Слуцкого — Юла.

Поскольку пики и ямы не появляются через равные интервалы, не существует и величины, которую было бы удобно назвать длиной колебания. Фактически можно изучать лишь распределение таких длин. В качестве средней длины можно определить средний период или от пика к пику, или от ямы к яме; однако здесь возникают некоторые трудности, связанные с тем, готовы ли мы учитывать периоды у небольших пульсаций, накладывающихся на основные колебания.

Сознавая, что ответ здесь до некоторой степени произволен, возьмем в качестве меры длины колебания среднее расстояние

между «пересечениями снизу вверх», точнее говоря, среднее расстояние между точками, где ряд меняет знак минус на плюс, или «пересекает ось времени». Предположим, что ряд получен скользящим усреднением с весами $a_1, \ldots, a_k$ нормальных величин с дисперсией v. Тогда вероятность того, что

$$u_k = \sum_{j=1}^{k} a_j \varepsilon_j < 0 \tag{46.51}$$

и

$$u_{k+1} = \sum_{j=1}^{k} a_j \varepsilon_{j+1} > 0, \tag{46.52}$$

т. е. вероятность того, что члены нового ряда изменят знак, равна интегралу по области, ограниченной гиперплоскостями $\sum a_j \varepsilon_j = 0$ и $\sum a_j \varepsilon_{j+1} = 0$, от функции

$$(2\pi v)^{-(k+1)/2} \exp\left\{-(2v)^{-1} \sum_{1}^{k+1} \varepsilon_j^2\right\}. \tag{46.53}$$

Этот интеграл равен деленному на 2π углу θ между двумя плоскостями, который дается соотношением

$$\cos\theta = \sum_{j=1}^{k} a_j a_{j+1} \Big/ \sum_{j=1}^{k} a_j^2. \tag{46.54}$$

Следовательно, $\theta/(2\pi)$ — математическое ожидание числа пересечений за единицу времени, а $2\pi/\theta$ приблизительно равно среднему расстоянию между пересечениями.

46.20 Аналогично вероятность того, что

$$u_{k+1} - u_k < 0, \tag{46.55}$$

$$u_k - u_{k-1} > 0, \tag{46.56}$$

т. е. что u_k есть пик ряда, равна углу между двумя гиперплоскостями

$$\sum_{j=1}^{k} a_j \varepsilon_{j+1} - \sum_{j=1}^{k} a_j \varepsilon_j = 0, \tag{46.57}$$

$$\sum_{j=1}^{k} a_j \varepsilon_j - \sum_{j=1}^{k} a_j \varepsilon_{j-1} = 0, \tag{46.58}$$

который дается формулой

$$\cos\theta_1 = \frac{(a_2 - a_1)\, a_1 + (a_3 - a_2)(a_2 - a_1) + \ldots + a_k (a_k - a_{k-1})}{a_1^2 + (a_2 - a_1)^2 + \ldots + a_k^2}. \tag{46.59}$$

Итак, среднее расстояние между пиками равно $2\pi/\theta_1$. Очевидно, что та же формула справедлива и для среднего расстояния между ямами.

46.21 Если мы хотим исключить из рассмотрений пульсации некоторой определенной длины d, мы можем искать вероятность выполнения (46.55) и (46.56) совместно с условием

$$u_k > u_{k+d}. \tag{46.60}$$

Последняя вероятность равна деленной на 4π площади части единичной сферы, вырезаемой гиперплоскостями (46.57), (46.58) и

$$\sum a_j \varepsilon_j - \sum a_j \varepsilon_{j+d} = 0. \tag{46.61}$$

Если углы между плоскостями равны A, B и C, искомая площадь равна $\theta_2 = A + B + C - 2\pi$.

Средняя длина между пиками, при исключенных пульсациях равна тогда $4\pi/\theta_2$.

Пример 46.6

В таблице 46.2 мы приводим 480 членов ряда случайных чисел, принимающих целые значения от 0 до 19. Там же приведены члены ряда, полученного из первоначального путем скользящего усреднения по три члена с последующим скользящим усреднением по пять членов. На рис. 46.1 графически показана часть полученного ряда. Сглаженный ряд состоит из 474 членов.

Среднее значение членов нового ряда равно $15 \times 9{,}5 = 142{,}5$. Число пересечений находится из таблицы и равняется 23. Первое пересечение происходит между 19-м и 20-м членами сглаженного ряда, последнее между 459-м и 460-м. Среднее расстояние между пересечениями равно $440/22 = 20$ единицам. Как это выглядит в сравнении со средним расстоянием, которое дает «нормальная» теория?

Веса, которыми мы пользуемся, задаются числами [1, 2, 3, 3, 3, 2, 1] и из (46.54) мы получаем, что $\cos\theta = 34/37 = 0{,}9189$ и $\theta = 23°14'$. Следовательно, среднее расстояние равно $360/23{,}233 = 15{,}5$ единицы. Наблюденное среднее значение равно 20, но оно основано на равномерном распределении, и мы имеем право ожидать некоторого отличия от теории для нормального распределения. У равномерно распределенных величин отклонение от среднего выражается более резко, и не удивительно обнаружить колебания, происходящие целиком выше или ниже среднего уровня.

Далее находим, что число пиков в ряде равно 62, первый приходится на 7-й член, последний на 466-й. Следовательно, среднее расстояние между пиками равно $459/61 = 7{,}5$ единицы. Из (46.59) находим, что $\cos\theta_1 = 2/3$ и $\theta_1 = 48°\,11'$. Теоретическое среднее расстояние равно $360/48{,}187 = 7{,}5$ единицы — прекрасное согласие с экспериментом. Можно увидеть, что несколь-

Таблица 46.2

Ряд из 480 равномерных случайных чисел ε и результат сглаживания S по формуле [5] [3]

t	ε	S	t	ε	S	t	ε	S	t	ε	S	t	ε	S	t	ε	S
1	3		40	3	140	79	16	175	118	10	111	157	7	174	196	5	128
2	15		41	14	135	80	15	188	119	8	116	158	15	160	197	7	122
3	15		42	7	146	81	17	197	120	10	131	159	5	157	198	8	126
4	8	164	43	16	141	82	11	200	121	3	145	160	11	141	199	18	120
5	19	147	44	3	139	83	6	206	122	16	156	161	9	140	200	0	121
6	1	143	45	10	117	84	17	215	123	12	173	162	14	122	201	7	105
7	3	145	46	12	96	85	18	228	124	8	175	163	1	117	202	9	99
8	12	165	47	0	75	86	19	230	125	19	160	164	11	94	203	5	93
9	19	175	48	3	65	87	15	220	126	11	145	165	1	98	204	3	95
10	13	196	49	2	61	88	13	198	127	1	129	166	8	93	205	12	91
11	16	191	50	3	71	89	8	175	128	4	123	167	2	106	206	4	93
12	4	178	51	10	84	90	10	159	129	16	108	168	18	103	207	2	97
13	17	159	52	5	91	91	14	158	130	3	115	169	1	121	208	11	97
14	8	150	53	10	92	92	5	158	131	13	108	170	7	117	209	6	107
15	6	134	54	3	101	93	12	159	132	0	118	171	9	127	210	7	115
16	15	118	55	2	119	94	18	153	133	10	112	172	13	120	211	6	128
17	3	101	56	11	141	95	1	145	134	4	122	173	2	137	212	15	125
18	3	88	57	14	166	96	14	124	135	19	113	174	16	139	213	4	130
19	7	87	58	18	190	97	8	112	136	3	110	175	1	145	214	13	126
20	4	100	59	8	212	98	1	108	137	4	100	176	17	142	215	4	125
21	5	126	60	14	211	99	5	123	138	7	103	177	13	145	216	7	123
22	14	140	61	15	204	100	13	131	139	0	103	178	0	149	217	13	119
23	15	147	62	17	191	101	11	150	140	16	107	179	15	157	218	4	111
24	10	150	63	7	185	102	14	151	141	13	102	180	7	166	219	13	101
25	3	153	64	9	166	103	6	140	142	0	103	181	16	164	220	0	91
26	10	156	65	11	160	104	13	120	143	2	114	182	16	171	221	3	82
27	13	165	66	14	167	105	1	119	144	4	127	183	7	169	222	11	76
28	14	175	67	5	188	106	4	120	145	18	136	184	6	174	223	0	75
29	15	168	68	17	203	107	13	133	146	18	140	185	13	168	224	10	72
30	8	160	69	18	204	108	13	147	147	6	131	186	17	170	225	1	86
31	10	154	70	13	205	109	8	172	148	0	121	187	14	170	226	4	92
32	1	156	71	18	185	110	12	186	149	4	120	188	2	159	227	6	109
33	18	154	72	0	171	111	12	195	150	11	137	189	15	140	228	18	116
34	17	165	73	19	149	112	19	204	151	15	162	190	9	139	229	3	139
35	4	164	74	1	146	113	13	203	152	15	179	191	1	145	230	7	149
36	10	159	75	14	130	114	11	184	153	11	188	192	15	151	231	12	149
37	16	138	76	12	135	115	18	156	154	9	184	193	16	147	232	15	141
38	2	137	77	2	139	116	2	135	155	12	183	194	6	144	233	11	137
39	13	131	78	7	160	117	4	121	156	18	175	195	10	132	234	1	134

Продолжение табл. 46.2

t	ε	S	t	ε	S	t	ε	S	t	ε	S	t	ε	S	t	ε	S
235	8	128	276	2	134	317	9	151	358	4	150	399	13	178	440	1	72
236	9	130	277	18	141	318	2	152	359	5	133	400	18	183	441	5	91
237	14	132	278	8	172	319	16	160	360	9	120	401	17	191	442	16	96
238	9	128	279	9	184	320	10	185	361	12	117	402	3	205	443	8	91
239	6	122	280	19	185	321	17	196	362	2	127	403	18	198	444	2	78
240	7	108	281	17	167	322	15	209	363	11	118	404	14	192	445	0	75
241	1	99	282	14	150	323	13	194	364	12	112	405	14	191	446	2	85
242	17	80	283	8	123	324	10	179	365	5	105	406	13	197	447	7	109
243	0	75	284	5	115	325	18	151	366	0	100	407	5	205	448	17	124
244	3	73	285	6	131	326	0	133	367	12	84	408	19	204	449	12	124
245	0	94	286	7	168	327	9	112	368	6	88	409	17	202	450	5	117
246	6	124	287	19	186	328	8	108	369	2	96	410	18	192	451	2	106
247	17	169	288	19	196	329	3	105	370	4	104	411	5	174	452	2	97
248	16	195	289	16	188	330	9	111	371	15	109	412	7	140	453	15	92
249	17	204	290	2	180	331	12	107	372	6	130	413	15	107	454	8	100
250	15	191	291	10	158	332	3	101	373	5	148	414	1	86	455	2	111
251	3	175	292	12	147	333	8	85	374	14	156	415	4	66	456	4	120
252	14	150	293	15	145	334	5	77	375	14	164	416	2	58	457	11	121
253	9	144	294	5	148	335	1	75	376	11	180	417	3	50	458	15	119
254	11	131	295	6	145	336	2	93	377	8	187	418	2	62	459	8	110
255	3	135	296	15	134	337	13	107	378	15	174	419	10	78	460	3	98
256	15	125	297	6	137	338	6	134	379	18	151	420	0	105	461	1	98
257	1	138	298	13	136	339	15	151	380	7	127	421	9	126	462	4	121
258	14	142	299	2	136	340	7	161	381	1	99	422	16	146	463	13	150
259	9	162	300	14	129	341	13	160	382	2	88	423	9	152	464	17	170
260	13	166	301	9	128	342	13	162	383	7	89	424	11	144	465	19	176
261	11	182	302	4	115	343	5	155	384	4	119	425	12	124	466	5	169
262	15	190	303	14	100	344	15	153	385	18	143	426	2	106	467	4	149
263	8	203	304	2	93	345	10	162	386	5	166	427	3	106	468	15	136
264	17	210	305	0	96	346	3	174	387	17	170	428	9	119	469	8	137
265	19	214	306	8	109	347	18	176	388	12	179	429	6	139	470	6	136
266	10	211	307	12	131	348	19	177	389	7	179	430	17	159	471	14	133
267	17	188	308	10	159	349	8	174	390	14	184	431	15	174	472	9	126
268	11	163	309	11	178	350	5	173	391	15	190	432	5	179	473	0	125
269	9	146	310	17	187	351	16	159	392	11	194	433	14	172	474	15	109
270	1	153	311	10	200	352	7	157	393	13	201	434	14	155	475	7	103
271	11	154	312	12	212	353	16	157	394	18	199	435	9	133	476	5	96
272	17	162	313	17	216	354	8	169	395	6	193	436	8	107	477	1	95
273	17	155	314	17	211	355	6	168	396	19	178	437	3	75	478	11	
274	4	154	315	16	192	356	15	165	397	13	173	438	1	53	479	5	
275	8	137	316	9	173	357	19	153	398	1	178	439	3	55	480	5	

ко расстояний между пиками соответствует очень малым пульсациям.

46.22 Исследуем теперь, как дисперсия индуцированных колебаний соотносится с дисперсией первоначального случайного ряда.

Дисперсия суммы k независимых случайных величин, каждая из которых имеет дисперсию v, равна $v \cdot k$, а среднее этих величин имеет дисперсию v/k. Отсюда не следует, что простое

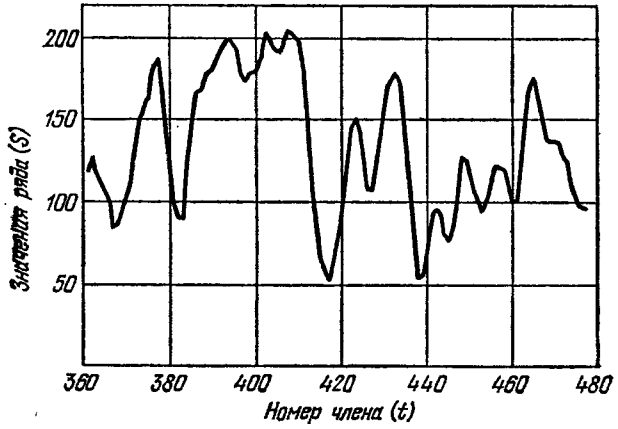

Рис. 46.1. График для последних 117 членов ряда S таблицы 46.2.

скользящее среднее имеет дисперсию в k раз меньшую, чем случайные величины. Это объясняется наличием корреляции между последовательными членами преобразованного ряда.

Если первоначальный ряд состоял из величин $\varepsilon_1, \ldots, \varepsilon_n$, то преобразование с весами $a_1, \ldots, a_k$ приводит к ряду

$$\begin{aligned} \eta_1 &= \sum a_j \varepsilon_j, \\ \eta_2 &= \sum a_j \varepsilon_{j+1}, \\ &\ldots\ldots\ldots\ldots \\ \eta_{n-k+1} &= \sum a_j \varepsilon_{j+n-k}. \end{aligned} \tag{46.62}$$

Если в качестве математического ожидания ε взять нуль, то математическое ожидание суммы величин η также будет равняться нулю. Так как имеется $n-k+1$ членов, выборочная дисперсия равна

$$\frac{1}{n-k+1} \sum \eta_m^2. \tag{46.63}$$

Поскольку величины ε независимы, математическое ожидание (46.63) равно

$$\frac{1}{n-k+1} \mathsf{M} \sum \eta_m^2 = \mathsf{M}\eta^2 = v \sum_{i=1}^{k} a_i^2. \tag{46.64}$$

В частности, если все величины a равны $1/k$, математическое ожидание выборочной дисперсии равно v/k. Мы видим, как меняется дисперсия в результате усреднения.

Если простое усреднение по k элементам повторяется q раз, соответствующие веса являются последовательными коэффициентами в

$$\frac{1}{k^q}(1 + x + \ldots + x^{k-1})^q = \frac{1}{k^q}\left(\frac{1-x^k}{1-x}\right)^q.$$

Сумма квадратов этих коэффициентов равна коэффициенту при $x^{q(k-1)}$ в

$$\frac{1}{k^{2q}} \frac{(1-x^k)^{2q}}{(1-x)^{2q}}. \tag{46.65}$$

Последняя сумма и дает новую дисперсию после q-кратного простого усреднения по k элементам. Ниже приводятся значения поправочного множителя для некоторых значений k и q.

k \ q	1	2	3	4	5
3	0,33	0,23	0,19	0,17	0,15
4	0,25	0,17	0,14	0,12	0,11
5	0,20	0,14	0,11	0,10	0,09
6	0,17	0,11	0,09	0,08	0,07
7	0,14	0,10	0,08	0,07	0,06

Ясно, что первое скользящее усреднение приводит к ряду с гораздо меньшей дисперсией, но вторая и последующие итерации не уменьшают дальше дисперсию столь же быстро. При $k = 7$ первое усреднение уменьшает дисперсию в семь раз, а три последующие — лишь еще наполовину.

46.23 Полезно также изучить влияние скользящего усреднения на сериальные корреляции в остаточной компоненте. Для ряда (46.51), полученного скользящим усреднением случайного ряда, как и в (46.54), получаем, что

$$\operatorname{cov}(u_t, u_{t+s}) = \mathsf{M}\left\{\sum a_j \varepsilon_{j+t} \sum a_j \varepsilon_{j+s+t}\right\} = v \sum_{j=1}^{k-s} a_j a_{j+s}. \tag{46.66}$$

Таким образом, s-я сериальная корреляция нового ряда равна

$$r_s = \begin{cases} \dfrac{\sum_{j=1}^{k-s} a_j a_{j+s}}{\sum_{j=1}^{k} a_j^2}, & |s| < k, \\ 0, & |s| \geqslant k. \end{cases} \qquad (46.67)$$

Итак, мы видим, что у полученного таким способом бесконечного ряда, несмотря на равенство нулю сериальных корреляций первоначального (случайного) ряда, новые сериальные корреляции не равны нулю вплоть до порядка k, т. е. до тех пор, пока члены нового ряда не будут зависеть от одних и тех же членов первоначального ряда.

Например, взяв простое скользящее усреднение по k элементам (все a равны $1/k$), из (46.67) легко получаем, что

$$r_s = 1 - \frac{|s|}{k}. \qquad (46.68)$$

Следовательно, корреляция может быть довольно большой при $s = 1$ и линейно уменьшается с ростом s до нуля при $s = k$. Высокая коррелированность такого типа между соседними величинами есть проявление эффекта Слуцкого — Юла.

Пример 46.7

Веса в формуле Спенсера для 21 точки таковы:

$$1/350\,[-1, -3, -5, -5, -2, 6, 18, 33, 47, 57, 60].$$

Отбрасывая делитель 350, который для наших целей несущественен, получаем, что сумма квадратов весов равна 17 542. Произведения (46.66) и соответствующие сериальные корреляции таковы:

k	$\Sigma a_j a_{j+k}$	r_k	k	$\Sigma a_j a_{j+k}$	r_k
0	17 542	1,000	11	−930	−0,053
1	16 786	0,957	12	−528	−0,030
2	14 667	0,836	13	−214	−0,012
3	11 584	0,660	14	−27	−0,002
4	8 085	0,461	15	50	+0,003
5	4 726	0,269	16	59	+0,003
6	1 951	0,111	17	40	0,002
7	6	0,000	18	19	0,001
8	−1 074	−0,061	19	6	0,000
9	−1 430	−0,082	20	1	0,000
10	−1 298	−0,074	21	0	0,000

Коррелограмма дана на рис. 46.2. Начиная с $k = 13$ и далее корреляции очень малы, и начиная с $k = 21$ и далее равны нулю.

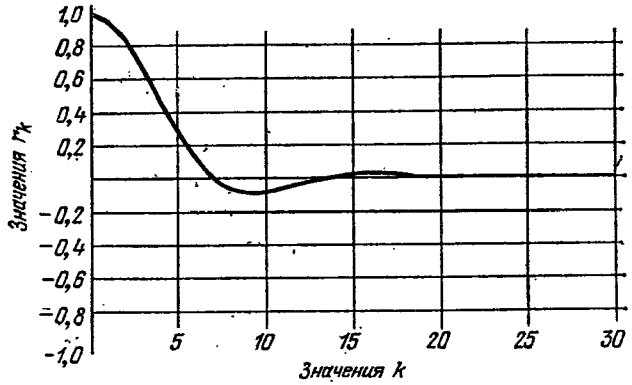

Рис. 46.2. Коррелограмма ряда, полученного с помощью формулы Спенсера для 21-й точки (пример 46.7).

Метод переменных разностей

46.24 Само понятие ряда, состоящего из полиномиальной компоненты и в той или иной мере случайного остатка, подсказывает метод, цель которого — исключить первую компоненту путем вычисления конечных разностей. Ясно, что вычисление последовательных разностей в конце концов приведет к полному исключению любой компоненты, которая действительно является полиномом. Можно быть практически уверенным, что такая операция исключит вообще любую систематическую компоненту, если только она не является экспоненциальной или циклической. Рассмотрим, как сказывается вычисление разностей на случайном ряде. Имеем

$$\Delta^r \varepsilon_t = \varepsilon_{t+r} - \binom{r}{1} \varepsilon_{t+r-1} + \binom{r}{2} \varepsilon_{t+r-2} - \ldots$$
$$\ldots + (-1)^r \varepsilon_t = (U-1)^r \varepsilon_t. \qquad (46.69)$$

Полагая, без ограничения общности, что ε_t имеет нулевое среднее, получаем, что

$$\mathbf{M}(\Delta^r \varepsilon_t) = 0, \qquad (46.70)$$

и если ε_t имеют одну и ту же для всех t дисперсию v,

$$\mathbf{D}(\Delta^r \varepsilon_t) = v \sum_{j=0}^{r} \binom{r}{j}^2 = v \times \{\text{коэфф. } x^r \text{ в } (1+x)^r (x+1)^r\} = \binom{2r}{r} v. \qquad (46.71)$$

Теперь можно получить оценку для v в виде

$$\hat{v} = \frac{\mu_2'(\Delta^r \varepsilon_t)}{\binom{2r}{r}}. \qquad (46.72)$$

Следует заметить, что в качестве $\mu_2'(\Delta^r \varepsilon_t)$ мы здесь используем второй момент относительно нуля, а не наблюденную (выборочную) дисперсию величины $\Delta^r \varepsilon_t$, поскольку известно, что среднее равно нулю. Это до некоторой степени сокращает вычисления. Множитель $\binom{2r}{r}$ при изменении r от 1 до 10 принимает следующие значения:

r	$\binom{2r}{r}$	$1/\binom{2r}{r}$	r	$\binom{2r}{r}$	$1/\binom{2r}{r}$
1	2	0,5	6	924	$0{,}0^2 108225$
2	0	0,166667	7	3 432	$0{,}0^3 291375$
3	20	0,05	8	12 872	$0{,}0^4 777001$
4	70	0,0142857	9	48 620	$0{,}0^4 205677$
5	252	$0{,}0^2 3968825$ *)	10	184 756	$0{,}0^5 5411254$

*) Запись $0{,}0^k XY\ldots$ означает наличие k нулей до первой значащей цифры X. (*Прим. ред.*)

46.25 Основанный на равенстве (46.72) метод переменных разностей состоит в следующем. Вычисляем разности ряда один раз, находим для них второй момент относительно нуля, делим его на 2, затем снова вычисляем разности и снова находим второй момент относительно нуля, деля затем на 6 и т. д. Если последовательные оценки v уменьшаются, продолжаем вычисление разностей. Вообще говоря, наступит момент, когда оценки перестанут уменьшаться и будут оставаться постоянными, конечно, с точностью до выборочных ошибок (которые могут быть довольно велики). На указанной стадии мы можем предположить, что систематическая компонента из первоначального ряда исключена. Последняя из полученных оценок дает нам оценку дисперсии случайного элемента в первоначальном ряде; а порядок разности, на котором мы остановились, указывает на степень полинома, представляющего систематическую компоненту.

Пример 46.8

Применим технику переменных разностей к ряду таблицы 46.1. Отталкиваясь от способа построения этого ряда, заключаем, что, во-первых, систематическая составляющая должна быть полностью исключена после вычисления третьих раз-

Таблица 46.3

Разности ряда u_t из таблицы 46.1 *)

t	u_t	Δ	Δ^2	Δ^3	Δ^4	Δ^5	Δ^6
1	−96	−6	67	155	279	508	1050
2	−90	−73	−88	−124	−229	−542	−1297
3	−17	15	36	105	313	755	1524
4	−32	−21	−69	−208	−442	−769	−1141
5	−11	48	139	234	327	372	271
6	−59	−91	−95	−93	−45	101	361
7	32	4	−2	−48	−146	−260	−229
8	28	6	46	98	114	−31	−625
9	22	−40	−52	−16	145	594	1661
10	62	12	−36	−161	−449	−1067	−2252
11	50	48	125	288	618	1185	1978
12	2	−77	−163	−330	−567	−793	−876
13	79	86	167	237	226	83	−159
14	−7	−81	−70	11	143	242	137
15	74	−11	−81	−132	−99	105	551
16	85	70	51	−33	−204	−446	−655
17	15	19	84	171	242	209	−64
18	−4	−65	−87	−71	33	273	690
19	61	22	−16	−104	−240	−417	−629
20	39	38	88	136	177	212	216
21	1	−50	−48	−41	−35	−4	175
22	51	−2	−7	−6	−31	−179	−650
23	53	5	−1	25	148	471	1110
24	48	6	−26	−123	−323	−639	−975
25	42	32	97	200	316	336	41
26	10	−65	−103	−116	−20	295	925
27	75	38	13	−96	−315	−630	−965
28	37	25	109	219	315	335	207
29	12	−84	−110	−96	−20	128	316
30	96	26	−14	−76	−148	−188	−32
31	70	40	62	72	40	−156	−798
32	30	−22	−10	32	193	642	1597
33	52	−12	−42	−164	−446	−955	−1719
34	64	30	122	282	509	764	950
35	34	−92	−160	−227	−225	−186	141
36	126	68	67	28	−69	−327	−991
37	58	1	39	97	258	664	1515
38	57	−38	−58	−161	−406	−851	−1492
39	95	20	103	245	445	641	707
40	75	−83	−142	−200	−196	−66	281
41	158	59	58	−4	−130	−347	−685
42	99	1	62	126	217	338	509
43	98	−61	−64	−91	−121	−171	−314
44	159	3	27	30	50	143	432

*) В данной таблице разности Δu_t определены, в отличие от (46.9), соотношением $\Delta u_t = u_t - u_{t+1}$. (*Прим. ред.*)

Продолжение табл. 46.3

t	u_t	Δ	Δ^2	Δ^3	Δ^4	Δ^5	Δ^6
45	156	−24	−3	−20	−93	−289	−745
46	180	−21	17	73	196	456	
47	201	−38	−56	−123	−260		
48	239	18	67	137			
49	221	−49	−70				
50	270	21					
51	249						

ностей, во-вторых, что случайная часть состоит из элемента с дисперсией, приблизительно равной 833. В действительности дисперсия случайной величины, принимающей целые значения от 1 до N, равна $(N^2-1)/12$, а N в этом случае равно 100. Фактическая дисперсия случайной величины в таблице 46.1 равна 843.

В таблице 46.3 приведены ряд и разности вплоть до Δ^6. Величины S_j — суммы квадратов для столбцов со значениями Δ^j — таковы:

$$S_1 = 107\,541$$
$$S_2 = 318\,115$$
$$S_3 = 1\,033\,513$$
$$S_4 = 3\,445\,308$$
$$S_5 = 11\,720\,069$$
$$S_6 = 40\,548\,844$$

Чтобы получить второй момент, делим S_j на $51-j$, а затем, чтобы получить оценку, еще на $\binom{2j}{j}$. Окончательно получаем следующее:

j	Оценка	j	Оценка
1	1075,41	4	1047,21
2	1082,02	5	1011,05
3	1076,58	6	975,20

Довольно любопытно, что при $j=2$ оценка больше, чем при $j=1$, и все оценки мало отличаются друг от друга. Рассуждая стандартно, мы должны были бы прийти к выводу, что систематической компонентой является полином первой степени, т. е. прямая, а дисперсия остаточного члена равна примерно 1000.

Читатель не должен удивляться, встречая при анализе такого рода небольших рядов расхождения между теорией и экспериментом. На самом деле расхождение не столь велико, как это кажется. Дисперсия первоначального ряда равна 6272,61. Деленный на 2 средний квадрат первой разности равен 1075,41, так что при вычислении первых разностей было исключено примерно 5/6 дисперсии, и метод показывает, и совершенно правильно, что большая часть систематической компоненты линейна. Случайная же компонента довольно велика по сравнению с нелинейной частью систематической компоненты, так что последняя как бы теряется в первой. Наш ряд слишком мал, чтобы метод переменных разностей мог разделить эти две составляющие. Рассмотрим, например, кубический член $(t-26)^3/100$. В первоначальном ряде его значения меняются от −156,25 до +156,25. Первая разность преобразует этот член в $3(t-26)^2/100$. Последнее меняется от 18,75 до нуля и снова до 18,75, тогда как случайная компонента меняется в диапазоне от 0 до 198. Уже здесь систематическая компонента поглощена случайной, и небольшая «случайная корреляция» между ними легко может вызвать увеличение среднего квадрата вторых разностей.

Возможен также и несколько иной подход. Допустим, что, положившись на метод переменных разностей, мы стали рассматривать наш ряд как представимый в виде суммы линейной и случайной компонент. Если бы мы методом наименьших квадратов выделили прямую и стали изучать остаточный член, то, вероятно, обнаружили бы весьма слабое отклонение от случайности.

Выбранное представление отличается от истинной конструкции ряда, но оно могло бы послужить основой для построения почти такого же ряда.

Следовательно, обнаружить слабость нашего представления можно, лишь установив его несоответствие последующим членам ряда.

46.26 Метод переменных разностей, таким образом, дает довольно низкую границу для степени полинома, участвующего в общем или локальном представлении ряда. Остается рассмотреть вопрос о том, когда различия между последовательными оценками v начинают обусловливаться лишь эффектом случайности. Иначе, нам надо решать, когда величины достигают некоторого стационарного уровня. Сумма квадратов S_j равна постоянной, умноженной на второй момент, но, так как ее члены коррелированы, мы не можем использовать дисперсию второго момента, чтобы проверить значимость отклонений от среднего. Кроме того, S_j и S_{j+1} коррелированы. Мы перейдем к вычислению выборочной дисперсии их разности, несколько сложная формула для которой принадлежит О. Андерсону (1914).

46.27 Пусть

$$b_j = \binom{r}{j}. \tag{46.73}$$

Тогда имеем

$$\frac{\mathbf{M}(\Delta^r u)^2}{\binom{2r}{r}} = \frac{\left(\sum b_j^2\right)\mu_2}{\sum b_j^2} = \mu_2, \tag{46.74}$$

где μ_2 — дисперсия u. Далее

$$\mathbf{M}(\Delta^r u)^4 = \mathbf{M}\{[b_0 u_{r+1} - b_1 u_r + \ldots + (-1)^r b_r u_1]^2 + \\ + [b_0 u_{r+2} - b_1 u_{r+1} + \ldots + (-1)^r b_r u_2]^2 + \ldots \\ \ldots + [b_0 u_n - b_1 u_{n-1} + \ldots + (-1)^r b_r u_{n-r}]^2\}^2. \tag{46.75}$$

Рассмотрим прежде всего члены, содержащие четвертую степень u. Мы получим их из выражения

$$\mathbf{M}\{b_0^2 u_{r+1}^2 + \ldots + b_r^2 u_1^2 + b_0^2 u_{r+2}^2 + \ldots + b_r^2 u_2^2 + \ldots \\ \ldots + b_0^2 u_n^2 + \ldots + b_r^2 u_{n-r}^2\}^2. \tag{46.76}$$

Пусть теперь

$$B_0^2 = (b_0^2)^2 + (b_0^2 + b_1^2)^2 + \ldots + (b_0^2 + \ldots + b_{r-1}^2)^2, \tag{46.77}$$

$$A_0^2 = (b_0^2 + b_1^2 + \ldots + b_r^2)^2 = \binom{2r}{r}^2. \tag{46.78}$$

Тогда член, содержащий $\mathbf{M}u^4$, равен

$$\{A_0^2(n-r) + 2B_0^2\}\,\mathbf{M}u^4. \tag{46.79}$$

Остальные члены, содержащиеся в (46.75), есть «члены типа $\mathbf{M}(u_l^2 u_m^2)$, $l \neq m$». Раскрыв скобки в (46.75), читатель легко найдет, что коэффициенты при $\mathbf{M}(u_l^2 u_m^2)$ выражаются в виде функций от величин

$$A_j^2 = (b_0 b_j + b_1 b_{j+1} + \ldots + b_{r-j} b_r)^2 = \binom{2r}{r-j}^2 \tag{46.80}$$

и

$$B^2 = (b_0 b_j)^2 + (b_0 b_j + b_1 b_{j+1})^2 + (b_0 b_j + \\ + b_1 b_{j+1} + \ldots + b_{r-j-1} b_{r-1})^2. \tag{46.81}$$

Выражение для $\mathbf{M}(\Delta^r u)^4$ сводится к выражению

$$(n-2r) A_0^2 \mathbf{M}u^4 + 4\{(n-2r+1) A_1^2 + \\ + (n-2r+2) A_2^2 + \ldots + (n-2r+r) A_r^2\}\,\mathbf{M}(u_l^2 u_m^2) + \\ + 2B_0^2 \mathbf{M}u^4 + 8\{B_1^2 + B_2^2 + \ldots + B_r^2\}\,\mathbf{M}(u_l^2 u_m^2). \tag{46.82}$$

Заменяя $\mathbf{M}u^4$ на μ_4, $\mathbf{M}(u_l^2 u_m^2)$ на μ_2^2, деля на $(n-r)^2\binom{2r}{r}^2$ и вычитая μ_2^2, находим дисперсию оценки $\hat{\sigma}$. Выражение можно, однако, несколько упростить. Положив

$$T_r = \sum_{j=0}^{r-1}\binom{r}{j}^2\binom{r}{j+1}^2 + 2\sum_{j=0}^{r-2}\binom{r}{j}^2\binom{r}{j+2}^2 + \\ + 3\sum_{j=0}^{r-3}\binom{r}{j}^2\binom{r}{j+3}^2 + \ldots + r\binom{r}{0}\binom{r}{r}^2, \tag{46.83}$$

после долгих алгебраических преобразований находим, что

$$\mathbf{D}\left[\frac{S_r}{(n-r)\binom{2r}{r}}\right] = \frac{\mu_4 - 3\mu_2^2}{n-r}\left\{1 - \frac{2T_r}{(n-r)\binom{2r}{r}^2}\right\} + \\ + \frac{2\mu_2^2}{n-r}\left\{\binom{4r}{2r}\Big/\binom{2r}{r}^2 - \frac{r}{2(n-r)}\right\}, \qquad r \leqslant n/2. \tag{46.84}$$

Пренебрегая членами порядка $(n-r)^{-2}$, приходим к выражению

$$\frac{\mu_4 - 3\mu_2^2}{n-r} + \binom{4r}{2r}\frac{2\mu_2^2}{n-r}\Big/\binom{2r}{r}^2, \tag{46.85}$$

или, используя приближение Стирлинга для факториалов, к выражению

$$\frac{1}{n-r}\{\mu_4 - 3\mu_2^2 + \mu_2^2\sqrt{2r\pi}\}. \tag{46.86}$$

Последнее служит хорошим приближением для (46.85), ошибка составляет не более 3 процентов при $r \geqslant 6$.

Если величины u соответствуют нормальной совокупности, $\mu_4 - 3\mu_2^2$ равно нулю и формула соответственно упрощается.

46.28 Сходным образом можно показать, что

$$\operatorname{cov}\left\{\frac{S_r}{(n-r)\binom{2r}{r}}, \frac{S_{r+1}}{(n-r-1)\binom{2r+2}{r+1}}\right\} = \\ = \frac{\mu_4 - 3\mu_2^2}{n-r}\left\{1 - \frac{2T_r'}{\binom{2r}{r}\binom{2r+2}{r+1}(n-r-1)}\right\} + \\ + \frac{2\mu_2^2}{n-r}\left\{\frac{\binom{4r+1}{2r}}{\binom{2r}{r}\binom{2r+2}{r+1}}\,\frac{2n-2r-1}{n-r-1} - \frac{r+1}{2(n-r-1)}\right\}, \tag{46.87}$$

где

$$T'_r = \sum_{j=0}^{r-1} \binom{r}{j}^2 \binom{r+1}{j+2}^2 + 2\sum_{j=0}^{r-2} \binom{r}{j}^2 \binom{r+1}{j+3}^2 + \ldots$$
$$\ldots + r\binom{r}{0}^2 \binom{r+1}{r+1}^2.$$

Мы можем теперь определить дисперсию разности

$$\frac{S_r}{(n-r)\binom{2r}{r}} \quad \text{и} \quad \frac{S_{r+1}}{(n-r-1)\binom{2r+2}{r+1}}.$$

Общая формула сложна, но для нормального распределения, больших n и $r \geqslant 6$ аналогично (46.86) получаем, что *)

$$\mathbf{D}\,(\text{разность}) = \frac{(3r+1)\sqrt{2\pi r}}{2(2r+1)^3(n-r-1)} \left\{ \frac{S_r}{(n-r)\binom{2r}{r}} \right\}^2. \qquad (46.88)$$

Использование этой формулы для вычислений облегчается, если заранее заготовить таблицы для входящих в нее коэффициентов. Здесь можно сослаться на Тинтнера (1940), где приведены таблицы, принадлежащие самому автору, О. Андерсону и Зейкову. Ниже мы рассмотрим несколько дальнейших модификаций, которые до некоторой степени упрощают формулу.

Пример 46.9

Применение метода переменных разностей вплоть до десятого порядка к данным из таблицы 45.4 (поголовье овец) приводит к следующим результатам:

r	$S_r/\binom{2r}{r}(n-r)$	r	$S_r/\binom{2r}{r}(n-r)$
1	3468	6	448
2	1442	7	401
3	854	8	371
4	629	9	357
5	518	10	347

Значения систематически уменьшаются с увеличением r от 1 до 10, но в конце такое уменьшение очень незначительно. При $r = 6$ из (46.88) мы получаем, что дисперсия разности равна

*) Очевидно, равенство (46.88) следует понимать как приближенное, интерпретируя его правую часть как статистическую оценку теоретической характеристики, стоящей в левой части. (*Прим. ред.*)

приблизительно 80,7, а при $r = 10$ — приблизительно 25,8. Выясняется, что уменьшение дисперсии теряет значимость при $r = 8$. Отсюда не следует, конечно, что линии тренда нужно приписать в точности такой порядок, поскольку совсем не следует исключать возможности колебаний в линии тренда. В сущности, мы бы немного потеряли, если бы ограничились выделением «кубического тренда».

46.29 Ясно, что метод переменных разностей не исключает таких систематических эффектов, как периодические колебания с малым периодом. Рассмотрим, например, ряд 1, —1, 1, —1, ... Вычисление первых разностей даст нам ряд 2, —2, 2, —2, ..., вычисление вторых разностей — ряд 4, —4, 4, —4, ... и т. д. Пренебрегая эффектами, связанными с небольшими длинами рядов, получаем, что дисперсия ряда r-х разностей в 2^{2r} раз больше, чем у первоначального ряда и частное при делении дисперсии на $\binom{2r}{r}$ равно

$$\frac{2^{2r}(r!)^2}{(2r)!} \sim \sqrt{\pi r},$$

и, таким образом, неограниченно возрастает. В такого рода ситуациях мы не можем получить оценку дисперсии случайного элемента, который может в принципе существовать.

Задача исследования разностей между S_r и S_{r+1}, или, что эквивалентно, задача проверки близости к единице отношения S_r/S_{r+1} усложняется тем, что разности, которые в этих величинах содержатся, коррелированы. Тинтнер (1940) и Н. Джонсон (1948) предложили методы преодоления указанной трудности. Однако эти методы предусматривают отказ от большой части данных.

46.30 Существует тесная связь между дисперсией разностей ряда и его сериальными корреляциями. Для ряда из n членов имеем

$$\sum_{t=1}^{n-1} (\Delta u_t)^2 = \sum_1^{n-1} (u_{t+1} - u_t)^2 = \sum_1^{n-1} \{(u_{t+1} - \bar{u}) - (u_t - \bar{u})\}^2 =$$
$$= \sum_1^{n-1} (u_{t+1} - \bar{u})^2 - 2\sum_1^{n-1} (u_{t+1} - \bar{u})(u_t - \bar{u}) + \sum_1^{n-1} (u_t - \bar{u})^2.$$

Тогда, после деления на $n - 1$, приближенно

$$\mathbf{D}\,(\Delta u_t) = \mathbf{D}u - 2\,\mathrm{cov}\,(u_{t+1}, u_t) + \mathbf{D}u = 2\mathbf{D}u\,(1 - r_1). \qquad (46.89)$$

С той же точностью

$$\mathbf{D}\,(\Delta^2 u_t) = \sum (u_{t+2} - 2u_{t+1} + u_t)^2 = \mathbf{D}u\,(6 - 8r_1 + 2r_2). \qquad (46.90)$$

Точно так же в общем случае

$$\mathrm{D}\Delta^p u_t = \mathrm{D}\left\{\sum_{j=0}^{p}\binom{p}{j}(-1)^j u_{t+j}\right\} =$$

$$= \mathrm{D}u\left\{\sum_{j=0}^{p}\binom{p}{j}^2 - 2r_1\sum_{0}^{p-1}\binom{p}{j}\binom{p}{j+1} + 2r_2\sum_{0}^{p-2}\binom{p}{j}\binom{p}{j+2} - \ldots\right\} =$$

$$= \mathrm{D}u\left\{\binom{2p}{p} - 2r_1\binom{2p}{p-1} + 2r_2\binom{2p}{p-2} - \ldots\right\} =$$

$$= \binom{2p}{p}\mathrm{D}u\left\{1 - \frac{2p}{p+1}r_1 + \frac{2p(p-1)}{(p+1)(p+2)}r_2 - \ldots\right\}. \quad (46.91)$$

Мы можем аналогично выразить сериальные корреляции в виде функций от дисперсий разностей.

Положим

$$V_j = \frac{\mathrm{D}\Delta^j u}{\binom{2j}{j}}, \quad V_0 = \mathrm{D}u. \quad (46.92)$$

Тогда можно показать (см. упражнение 46.14), что

$$V_0 r_p = V_0 - p^2 V_1 + \frac{p^2(p^2-1)}{(2!)^2}V_2 - \frac{p^2(p^2-1)(p^2-4)}{(3!)^2}V_3 + \ldots \quad (46.93)$$

46.31 Относительная простота этих формул обусловлена тем, что мы имеем дело с длинными рядами и можем пренебречь краевыми эффектами, фактически считая, что

$$\sum_{i=1}^{n-1} u_i^2 \doteqdot \sum_{i=1}^{n-1} u_{i+1}^2.$$

Обсуждаемые формулы можно сделать и точными, если видоизменить определения так, чтобы исключить краевые эффекты с самого начала. Определим последовательность специальных сумм следующим образом:

$$\sum\nolimits_{(0)}(x_i y_i) = x_1y_1 + x_2y_2 + \ldots + x_{n-1}y_{n-1} + x_n y_n, \quad (46.94)$$

$$\sum\nolimits_{(1)}(x_i y_i) = \frac{1}{2}x_1y_1 + x_2y_2 + \ldots + x_{n-1}y_{n-1} + \frac{1}{2}x_n y_n, \quad (46.95)$$

$$\sum\nolimits_{(2)}(x_i y_i) = \frac{1}{4}x_1y_1 + \frac{3}{4}x_2y_2 + x_3y_3 + \ldots$$

$$\ldots + x_{n-2}y_{n-2} + \frac{3}{4}x_{n-1}y_{n-1} + \frac{1}{4}x_n y_n. \quad (46.96)$$

Общий закон формирования таких сумм определяется следующей рекуррентной формулой:

$$\sum_{i=1}^{n}{}_{(m)}(x_i y_i) = \frac{1}{2}\sum_{i=1}^{n-1}{}_{(m-1)}(x_i y_i + x_{i+1}y_{i+1}), \quad (46.97)$$

например, у первых трех членов в $\sum_{(3)}$ коэффициенты равны $\frac{1}{8}, \frac{1}{2}, \frac{7}{8}$.

Теперь определим новые величины

$${}_mV_0 = \sum\nolimits_{(m)} u^2/n, \quad (46.98)$$

$${}_mV_1 = \sum\nolimits_{(m-1)}(\Delta u)^2/2n, \quad (46.99)$$

. . . . . .

$${}_mV_p = \sum\nolimits_{(m-p)}(\Delta^p u)^2 \Big/ \binom{2p}{p} n. \quad (46.100)$$

Аналогично определим

$${}_mr_p = \sum\nolimits_{(m-p)} u_i u_{i+p} \Big/ \sum\nolimits_{(m)} u_i^2. \quad (46.101)$$

Определенные величины в точности удовлетворяют соотношениям (46.91) и (46.93). По-видимому, проще всего в этом убедиться, рассмотрев ряд

$$0,\ 0,\ 0,\ u_1,\ u_2,\ \ldots,\ u_n,\ 0,\ 0,\ 0,\ \ldots \quad (46.102)$$

Первые разности равны

$$0,\ 0,\ u_1,\ u_2 - u_1,\ \ldots,\ u_n - u_{n-1},\ -u_n,\ 0,\ 0,\ \ldots, \quad (46.103)$$

а сумма их квадратов равна

$$u_1^2 + \sum_{j=1}^{n-1}(u_{j+1} - u_j)^2 + u_n^2 = 2\sum_{j=1}^{n} u_j^2 - 2\sum_{j=1}^{n-1} u_j u_{j+1} = 2_1V_0(1_1 - r_1). \quad (46.104)$$

Для такого ряда (46.97) эквивалентно равенству

$$\sum_{i=1}^{n}{}_{(m)} u_i u_{i+k} = \frac{1}{2}\sum_{i=1}^{n}{}_{(m-1)}(u_i u_{i+k} + u_{i+1}u_{i+1+k}) = \left(\sum\nolimits_{(m)} u_i^2\right)({}_mr_k).$$

Соображения, которые привели нас к (46.91) и (46.93), остаются в силе и для бесконечных рядов, а значит, и для ряда (46.102). Следовательно, (46.91) и (46.93) справедливы и для наших модификаций величин V и r.

46.32 До сих пор мы говорили прежде всего об использовании метода переменных разностей для определения порядка полинома, осуществляющего наилучшее выравнивание. Процедура поиска кончалась, когда величины V переставали существенно меняться при вычислении дальнейших разностей. Но предпо-

лагая, что первоначальный ряд представим в виде суммы полинома и случайной «ошибки», мы можем при заданных величинах V интересоваться и тем, какова наилучшая оценка дисперсии ошибки. Этот вопрос изучался Кенуем (1953b), который искал линейную функцию величин V, имеющую минимальную дисперсию. В большинстве практических ситуаций было бы более реалистично исследовать наличие сериальных корреляций среди ошибок. Если же наличие сериально коррелированных ошибок задано, задача состоит в распространении метода переменных разностей на случай оценивания сериальных корреляций. Этот вопрос также рассматривался Кенуем. См. упражнения 46.9—11.

46.33 Как бы мы ни решали нашу задачу, подбирая ли полином непосредственно, используя ли скользящие средние или какой-либо другой процесс сглаживания, мы все равно столкнемся с трудностями, о которых упоминалось в **46.18.** Исключение тренда исказит остаточный член, и, по-видимому, избежать этого нельзя. Можно лишь стремиться к наилучшему поведению в такой ситуации и поступать здесь можно двояко: можно выбирать методы, которые при прочих равных обстоятельствах минимизируют искажения, и можно организовывать процедуру так, чтобы при возникновении опасений на любой последующей стадии анализа можно было бы выделить обусловленные сглаживанием искажения из элементов остаточного ряда. Перейдем к исследованию возможностей второго направления.

46.34 Предположим, что мы разбили наш ряд из n членов на последовательные группы и в каждой группе выравниваем члены полиномом одного типа. Внутри каждой группы мы можем получить удовлетворительную линию тренда. Но ясно, что линия тренда каждой группы должна соединяться с линией последующей группы, причем, в каком-то приемлемом смысле, гладко.

Это условие, которое мы вскоре рассмотрим более детально, является ограничением, налагаемым на методику, в целом же описываемый подход имеет то преимущество, что позволяет исследовать ряд как последовательность независимых блоков и применять к последним дисперсионный анализ. Метод можно рассматривать как компромиссный вариант между методом скользящих средних и полиномиальным выравниванием членов всего ряда. Он подробно рассматривался Роудсом (1921). Обобщения принадлежат Кеную (1949a).

В качестве простого примера рассмотрим выравнивание ряда прямыми линиями по группам, состоящим из трех элементов. Если сглаженное значение u_1 есть $2b_1$, а сглаженное значение u_3 есть $2b_2$, то в качестве сглаженного значения u_2 следует взять b_1+b_2. Далее, пусть сглаженное значение u_5 есть $2b_3$ (u_3 мы

уже сгладили с помощью $2b_2$), сглаженное значение u_4 есть b_2+b_3 и т. д. В результате имеем следующие значения:

$$\begin{array}{lccccccc} \text{действительные:} & u_1 & u_2 & u_3 & u_4 & u_5 & u_6 & u_7 \ldots \\ \text{выровненные;} & 2b_1 & b_1+b_2 & 2b_2 & b_2+b_3 & 2b_3 & b_3+b_4 & 2b_4 \ldots \end{array} \tag{46.105}$$

Так определенная линия тренда непрерывна, но ее первая производная терпит разрыв в моменты «времени» 3, 5, 7 и т. д., иначе, линия состоит из прямых отрезков, каждый из которых содержит три выравненных элемента.

46.35 Значения постоянных b можно определить, как обычно, методом наименьших квадратов, т. е. минимизируя

$$(u_1-2b_1)^2+\{u_2-(b_1+b_2)\}^2+(u_3-2b_2)^2+\ldots, \tag{46.106}$$

что приводит к системе уравнений

$$\left.\begin{array}{r} 2u_1+u_2-5b_1-b_2=0, \\ u_2+2u_3+u_4-b_1-6b_2-b_3=0, \\ \cdots\cdots\cdots\cdots\cdots \end{array}\right\} \tag{46.107}$$

Решить эти уравнения нетрудно, но их можно сильно упростить, если вновь модифицировать схему так, чтобы исключить краевые эффекты. Пусть исследуемый временной ряд представлен нечетным числом членов $u_1, \ldots, u_{2m+1}$. Заменим наш ряд на ряд

$$\tfrac{1}{2}(u_1+u_{2m+1}),\ u_2,\ u_3,\ \ldots,\ u_{2m},\ \tfrac{1}{2}(u_1+u_{2m+1}). \tag{46.108}$$

«Выравненные величины» b выберем равными

$$b_1+b_m,\ 2b_1,\ b_1+b_2,\ \ldots,\ 2b_m,\ b_1+b_m. \tag{46.109}$$

Такой выбор аналогичен процедуре **45.34**, где мы последний член приравнивали первому и, следовательно, превращали ряд в циклический. Положив $u_1'=\frac{1}{2}(u_1+u_{2m+1})$, минимизируем сумму

$$\{u_1'-(b_1+b_m)\}^2+(u_2-2b_1)^2+\ldots+(u_{2m}-2b_m)^2.$$

Приходим к уравнениям

$$\left.\begin{array}{l} 6b_1+b_2+\ldots\ldots\ldots\ldots+b_m=u_1'+2u_2+u_3=U_2, \\ b_1+6b_2+b_3 \qquad\qquad\qquad\qquad\quad =u_3+2u_4+u_5=U_4, \\ \qquad b_2+6b_3+b_4 \qquad\qquad\qquad\quad =u_5+2u_6+u_7=U_6, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ b_1+\ldots\ldots\ldots+b_{m-1}+6b_m=u_{2m-1}+2u_m+u_1'=U_{2m}. \end{array}\right\} \tag{46.110}$$

Преимущество такой формы в том, что коэффициенты при величинах b образуют симметрическую циркулянтную матрицу и система уравнений может быть решена одним из стандартных способов. (Методику смотри у Кенуя, 1949а, и Гуда, 1950 — см. упражнение 46.13). В результате величины b должны быть выражены в виде линейных функций от величин U из (46.110).

В упражнении 46.12 полученные результаты обобщаются.

Пример 46.10

Вернемся к данным таблицы 46.1. Значения u_t у нас будут такие же, как и в столбце (4) с тем исключением, что u_1 и u_{51} мы заменим на $(u_1 + u_{51})/2$. Все u_t вновь приведены в таблице 46.4. Значения U_t приведены в третьем столбце, а соответствующие значения b_t — в четвертом. Так, например,

$$U_2 = 76{,}5 + 2(-90) + (-17) = -120{,}5,$$

$$U_{50} = 221 + 2(270) + 76{,}5 = 837{,}5.$$

Таблица 46.4

Ряд таблицы 46.1, «выровненный прямыми по трем точкам»

t	u_t	U_t	b_t	t	u_t	U_t	b_t
1	76,5			28	37	161	17,6740
2	—90	—120,5	—40,3681	29	12		
3	—17			30	96	274	39,4348
4	—32	—92	—5,6002	31	70		
5	—11			32	30	182	19,7171
6	—59	—97	—18,0309	33	52		
7	32			34	64	214	24,2620
8	28	110	16,7855	35	34		
9	22			36	126	344	48,7106
10	62	196	27,3178	37	58		
11	50			38	57	267	27,4740
12	2	133	15,3071	39	95		
13	79			40	75	403	53,4452
14	—7	139	13,8390	41	158		
15	74			42	99	454	54,8549
16	85	259	40,6586	43	98		
17	15			44	159	572	71,4253
18	—4	68	1,2096	45	156		
19	61			46	180	717	88,5930
20	39	140	20,0840	47	201		
21	1			48	239	900	114,0164
22	51	156	18,2862	49	221		
23	53			50	270	837,5	127,3086
24	48	191	26,1987	51	76,5		
25	42						
26	10	137	15,5212	Всего	—	6545,0	818,1244
27	75						

Также

$$6b_1 + b_2 + b_m = -242{,}2086 - 5{,}6002 + 127{,}3086 = -120{,}5002 = U_2$$

и т. д. Выровненные значения получаются немедленно, например при $t = 2$ $u_2 = -90$; $2b_1 = -80{,}74$; при $t = 3$ $u_3 = -17$; $b_1 + b_2 = -45{,}97$. При $t = 1$ $u_1 = 76{,}5$; $b_1 + b_{25} = 86{,}94$.

Как обычно, при «выравнивании наименьшими квадратами», нам не нужно вычислять каждый остаточный член, чтобы получить сумму квадратов, так как (см. (35.1)) остаточная СК равна

$$\sum u_t^2 - \sum b_i U_i.$$

В нашем примере

$$\sum u_t^2 = 474\,458{,}25,$$

$$\sum b_i U_i = 448\,274{,}26,$$

$$\text{Разность} = 26{,}184.$$

Мы получили 25 величин b для 51 наблюдения. На наблюдения было наложено одно условие, так что число ст. св. при оценивании остаточной дисперсии равно 25. Значение оценки равно 26,184/25 = 1047, в то время как в примере 46.8 (при вычислении лишь первых разностей) мы получили значение 1075.

Можно поступать и лучшим образом. Метод выравнивания по трем точкам предполагает, что может существовать корреляция между наблюденными остатками в соседних точках внутри троек, и корреляции нет, если точки принадлежат разным группам.

В действительности мы можем рассматривать ряд в виде $(n-1)/2$ блоков по два элемента. Две разности в выровненной триаде имеют вид $b_2 - b_1$, $b_2 - b_1$. Сумма квадратов внутри блока оценивается выражением $\frac{1}{50}\left(\sum u_i - \sum u_{i+1}\right)^2$, которое оказывается равным 406,12. Мы можем теперь рассмотреть полную сумму квадратов. Имеем

	Ст. св.	СК	«Средний квадрат»
Выравненные величины	25	448274,26	
Внутри блоков	1	406,12	
Остатки	24	25777,87	1074
	50	474458,25	

«Остаточный средний квадрат» теперь почти полностью согласуется со значением, полученным в примере 46.8 с помощью

первых разностей. Фактически согласие более полное, чем мы имели право ожидать.

Более пространное изложение этой темы можно найти у Кенуя (1949а). Впоследствии Кенуй решал задачу выравнивания так, чтобы отдельные элементы линии тренда соединялись гладко.

Сезонные колебания

46.36 При анализе сезонных колебаний одно важное обстоятельство облегчает решение задачи и одно второстепенное мешает решению. Облегчает положение нам тот факт, что мы знаем, что период колебаний равен одному году. Мешает решению задачи то обстоятельство, что наблюдения обычно проводятся лишь через определенные интервалы: ежеквартально, ежемесячно, еженедельно. Большинство выводов, относящихся к сезонным колебаниям в течение года, может быть адекватно распространено и на циклические процессы с иным периодом. Примерами могут служить изменение температуры в течение дня или цен в течение недели. Ради простоты изложения мы ограничимся сезонностью в строгом смысле этого слова.

46.37 Часто сезонность выражена настолько ярко, что нет необходимости доказывать ее существование. Однако имеются такие ситуации, когда мы не уверены, не обусловлена ли эволюция членов ряда случайными колебаниями наложенными на тренд или флуктуациями нециклического характера. Поэтому прежде всего нам нужен критерий для проверки наличия сезонности, и в любом случае мы должны уметь измерять сезонные эффекты. Проиллюстрируем наши возможности на поквартальных данных таблицы 46.5 (поквартальные индексы розничной цены на овощи). В таблице 46.6 приведены данные после выбора нового начала координат и масштаба.

Таблица 46.5

Поквартальные индексы розничной цены на овощи в Соединенном Королевстве, 1951—8. (Данные из Journal of the Royal Statistical Society.)

	1951	1952	1953	1954	1955	1956	1957	1958
Первый квартал	295,0	324,7	372,9	354,0	333,7	323,2	304,3	312,5
2-й	317,5	323,7	380,9	345,7	323,9	342,9	285,9	336,1
3-й	314,9	322,5	353,0	319,5	312,8	300,3	292,3	295,5
4-й	321,4	332,9	348,9	317,6	310,2	309,8	298,7	318,4

Таблица 46.6

Данные таблицы 46.5 при начале отсчета 300. Все значения умножены на 10

	1951	1952	1953	1954	1955	1956	1957	1958
Первый квартал	—50	247	729	540	337	232	43	125
2-й	175	237	809	457	239	429	—141	361
3-й	149	225	530	195	128	3	—77	—45
4-й	214	329	489	176	102	98	—13	184

46.38 Исследуем прежде всего возможности критериев, свободных от распределения. Было бы соблазнительно ранжировать кварталы внутри каждого года (от 1 до 4) и рассмотреть, как меняется ранжировка от года к году. Однако недолгое размышление показывает, что такая процедура не может отличить сезонность от тренда. Если бы данные равномерно росли по времени, а в этом случае сезонности нет, у первого квартала всегда был бы наименьший ранг. Фактически, чтобы добиться успеха, нам необходимо в качестве первого шага исключить тренд.

Модель, которую мы использовали до сих пор, заключалась в аддитивном представлении временного ряда, но по очевидным причинам разумно считать, что влияние сезонности, носит мультипликативный характер. Таким образом, введем «годовые» величины y, «сезонную» величину s (постоянная пропорциональности, не меняющаяся от года к году), случайную ошибку ε и рассмотрим следующую модель:

$$u_{tq} = y_t s_q + \varepsilon, \qquad t = 1, \ldots, n; \; q = 1, 2, 3, 4. \tag{46.111}$$

Если тренд мал, сезонность можно характеризовать аддитивной постоянной (а не постоянной пропорциональности), и тогда можно записать, что

$$u_{tq} = y_t + s_q + \varepsilon. \tag{46.112}$$

Последнее представляет обычную модель дисперсионного анализа с двухфакторной перекрестной классификацией.

Если тренд ярко выражен, (46.112) представляется посредственной аппроксимацией, и безопаснее работать с моделью

$$\log u_{tq} = \log y_t + \log s_q + \eta. \tag{46.113}$$

Простейший путь использования (46.111) состоит в том, чтобы делить u_t на среднее y_t по рассматриваемому году и затем рассматривать полученные отношения как оценки мультипликативного сезонного фактора.

Применительно к таблице 46.5 это выглядит так. На 1951 г. квартальное среднее равно 312,2. Разделив поквартальные величины на это среднее, получим ряд (в %): 94,49; 101,70; 100,86; 102,95. Аналогичные вычисления по всем годам дадут следующее:

	1951	1952	1953	1954	1955	1956	1957	1958	Среднее
Первый квартал	94,49	99,62	102,47	105,92	104,23	101,30	103,05	99,0	101,3
2-й	101,70	99,31	104,66	103,44	101,17	107,48	96,82	106,49	102,6
3-й	100,86	98,94	97,00	95,60	97,70	94,12	98,99	93,62	97,1
4-й	102,95	102,13	95,87	95,03	96,89	97,10	101,15	100,88	99,0

Представляется разумным выписать среднее поквартальных данных по всем годам для того, чтобы оценить сезонную компоненту (определенная выше постоянная). Это сделано в последнем столбце. Если бы мы хотели «подправить» первоначальные данные с целью исключения сезонности, мы должны были бы разделить данные за первые кварталы на 101,3, данные за вторые кварталы на 102,6 и т. д.

46.39 Иной подход заключается в предварительном исключении тренда методом скользящих средних и последующем изучении остаточных характеристик сезонности. Мы здесь конечно, рискуем исказить остаток. Однако если организовывать скользящее усреднение с осторожностью, можно минимизировать искажение настолько, насколько это касается сезонности. В сущности мы отмечаем в **46.16**, что если простое скользящее среднее (с равными весами) берется по отрезку длиной, равной периоду циклической компоненты, то значение тренда от этой компоненты равно нулю, так что остаток остается незатронутым.

Для данных таблицы 46.5 это означает использование простого скользящего среднего четырех элементов, что эквивалентно исключению линейного тренда. Наши первоначальные данные, однако, являются средними ценами за квартал и относятся, следовательно, к временным интервалам длиной в три месяца и с центрами: 15 февраля, 15 мая, 15 августа, 15 ноября (или что-то возле этого). Усреднение по четырем элементам приведет нас к значениям тренда в точках, находящихся где-то посередине между этими датами. Чтобы иметь возможность сравнивать полученные средние с первоначальными данными, соответствующие временные точки надо сдвинуть. Для этого мы «центрируем» среднее. Наиболее просто это сделать, взяв средние последовательных пар средних, вычисленных по четырем элементам. Так, в таблице 46.6 среднее первых четырех значений равно 122, а от второго до пятого равно 196,25. Среднее этих двух чисел — 159,125 рассматривается как значение тренда, соответствующее третьему кварталу 1951 г., где первоначальное значение ряда равно 149. Ясно, что процесс эквивалентен вычислению среднего пяти элементов с весами

$$\frac{1}{8}[1,\ 2,\ 2,\ 2,\ 1]. \qquad 46.114)$$

Остатки, полученные при обработке в том же духе всех данных таблицы 46.6 приведены в таблице 46.7. Отклонения средних (вычисленных по семи годам для каждого из четырех кварталов) от общего среднего, равного 24,01/4, равны 62,45, 86,17, —88,39, —60,26. Если бы использовались первоначальные величины, соответствующие значения равнялись бы 306,25, 308,62, 291,16, 293,97 или (в процентах от среднего)— 102,1, 102,9, 97,1, 98,0. Наблюдается существенное отличие от результатов, полученных методом **46.38**.

46.40 Интересно также посмотреть, что будет, если мы исключим тренд, осуществляя более тщательное скользящее усреднение. Рассмотрим выравнивание кубическим полиномом по семи точкам с весами [—2, 3, 6, 7]/21. Остатки указаны в таблице 46.8. Индексы сезонности оказываются равными

102,3, 102,3, 97,3, 98,1,

в то время как ранее мы получили

101,3, 102,6, 97,1, 99,0 (в **46.38**)

102,1, 102,9, 97,1, 98,0 (в **46.39**)

Хотя общая картина во всех случаях одна и та же («сезонный пик» во втором квартале, «сезонная яма» в третьем), между результатами имеются достаточно большие различия, чтобы толкнуть нас на вычисления, требующие большей аккуратности. Мы склонны использовать метод **46.39**. Есть некоторая опасность, что выравнивание методом **46.40** слишком хорошо в том смысле, что линия тренда включает часть сезонных эффектов. Представляется невозможным, однако, установить полностью объективное правило, разделяющее тренд и сезонность. Мы бы рекомендовали попробовать несколько методов и выбрать тот, что дает наиболее разумные результаты. В публикациях же следует точно устанавливать, что именно было сделано.

Иногда к хорошим результатам приводит итеративный процесс: исключается грубый тренд, оценивается сезонный фактор, первоначальные данные подправляются с целью исключения сезонности, вновь оценивается тренд и т. д. Или снова вместо простого среднего (как в **46.38**) можно взять скользящее среднее, чтобы получить скользящий индекс сезонности.

Таблица 46.7

Остатки для данных таблицы 46.6 после исключения тренда скользящими средними по четырем точкам

	1951	1952	1953	1954	1955	1956	1957	1958	Всего	Средние
Первый квартал		25,750	167,875	77,875	108,625	24,875	52,250	22,000	479,250	68,46
2-й		−8,125	189,750	75,875	28,250	238,000	107,875	229,375	645,250	92,18
3-й	−10,125	−94,750	−85,625	−121,625	−60,375	−163,875	−40,250		−576,625	−82,38
4-й	10,000	−122,500	−59,000	−88,000	−97,000	26,000	−49,250		−379,750	−54,25

Таблица 46.8

Остатки для данных таблицы 46.6 после исключения тренда выравниванием кубическим полиномом по семи точкам

	1951	1952	1953	1954	1955	1956	1957	1958	Всего	Средние
Первый квартал		+30,38	122,14	79,95	113,19	16,33	91,14	5,43	458,56	65,51
2-й		+29,19	135,62	82,24	27,43	206,57	−84,24		396,81	66,13
3-й		−53,71	−123,95	−106,81	−35,76	−191,57	12,38		−499,42	−83,24
4-й	−12,67	−128,67	−72,57	−70,48	−97,90	25,81	−54,00		−410,48	−58,64

46.41 Как мы увидим в главе 49, спектральный анализ поможет нам глубже разобраться в вопросах исключения тренда при нахождении как случайного остатка, так и сезонной компоненты. Мы увидим, что по крайней мере в определенных случаях можно подправлять спектральную функцию с целью устранения искажений, обусловленных выравниванием.

46.42 Существует несколько сложных программ для электронно-вычислительных машин, позволяющих разбить ряд на тренд, сезонную и остаточную компоненты. Эти программы могут включать в себя не один процесс изоляции каждой компоненты; например, могут использоваться предварительное сглаживание, первое приближение к сезонной компоненте, более точное выделение тренда, дальнейшее приближение к сезонной компоненте и т. д. «Чтобы узнать вкус пудинга, надо его отведать», и следует отметить, что такие программы работают вполне хорошо при решении широкого класса практических задач из экономики и социологии. Обсуждаемый вопрос рассматривался в работах Шискина (1955), Эйзенпресса (1956), Шискина и Эйзенпресса (1957) и Бёрмэна (1965).

УПРАЖНЕНИЯ

46.1 Значения ряда, заданные во временных точках $-m, \ldots, m$ (всего $2m+1$ значений), выравниваются прямой. Показать, что для этой прямой

$$a_0 = \frac{1}{2m+1}\sum u_t, \qquad a_1 = \frac{3}{m(m+1)(2m+1)}\sum t u_t.$$

Получить отсюда, что сумма квадратов коэффициентов основанного на такой прямой скользящего среднего в точке $t = j$ равна

$$\frac{1}{2m+1}\left[1 + \frac{3j^2}{m(m+1)}\right],$$

46.2 Выравнять кубическим полиномом последние семь значений ряда из таблицы 45.4 (поголовье овец) и показать, что тренд для последних четырех значений равен 1639, 1687, 1750 и 1807.

46.3 Показать, что веса в формуле Спенсера для 21-й точки равны

$$\frac{1}{350}[-1, -3, -5, -5{,}6, 18, 13{,}47, 57, 60]$$

и что если формулу применить к случайному ряду, новая дисперсия будет составлять приблизительно одну седьмую от первоначальной — приблизительно то же дало бы простое скользящее усреднение по семи элементам.

46.4 Показать, что веса в формуле Меколи для 43 точек —

$$\frac{1}{960}[12][8][5]^2\left[\frac{7}{10}, -1, 0, 0, 0, 0, 0, 0, 1\right]$$

равны $\frac{1}{9600}$ [7, 18, 30, 40, 45, 28, −8, −60, −122, −178, −205, −190, −127, −6, 163, 360, 562, 760, 928, 1050, 1127, 1156] и что уменьшение дисперсии случайного ряда здесь приблизительно такое же, как и при простом усреднении по девяти элементам.

46.5 Пусть ε_t — случайный ряд. Показать, что корреляция между последовательными членами ряда $\Delta^k\varepsilon_t$ для больших рядов равна $-k/(k+1)$ и, следовательно, стремится к -1 при увеличении k. Показать, что если u_t — сумма случайного и систематического полиномиального элементов, знаки у соседних членов ряда $\Delta^k u_t$ с ростом k все чаще оказываются противоположными. Проверить это, обратившись к таблице 46.3.

46.6 Исключая δ^2 из (46.28), показать, что для кривой третьего порядка точная линия тренда получается с помощью процедуры

$$\frac{1}{h^2-k^2}\left\{\frac{h^2-1}{k}[k]-\frac{k^2-1}{h}[h]\right\}.$$

Обобщить этот результат.

(См. Хигхем, 1882—5.)

46.7 Показать, что для больших случайных рядов нормальных величин с дисперсией σ^2 выборочные сериальные корреляции не коррелированы и что

$$\mathrm{D}r_k = 1/n.$$

Отсюда и из (46.92) получить для больших выборок формулу

$$\mathrm{D}V_p = \frac{2\sigma^4}{n}\left\{1+\frac{2p^2}{(p+1)^2}+\frac{2p^2(p-1)^2}{(p+1)^2(p+2)^2}\cdots\right\}.$$

(Кенуй, 1953b.)

46.8 Показать, что в обозначениях **46.30**, применительно к случайным рядам с дисперсией σ^2 и четвертым семиинвариантом $\varkappa_4$

$$\operatorname{cov}(V_i, V_j)-\frac{\varkappa_4}{n}=\frac{\sigma^4}{n}a_{ij}=\frac{\sigma^4}{2n}\binom{2i+2j}{i+j}\Big/\binom{2i}{i}\binom{2j}{j}$$

(Кенуй, 1953b.)

46.9 Пусть в предыдущем примере дана оценка дисперсии ошибки

$$t=\sum_{i=m+1}^{m+p} c_i V_i$$

с дисперсией $(\varkappa_4+\lambda\sigma^4)/n$. Показать, что дисперсия минимальна, если

$$\sum_{i=m+1}^{m+p} a_{ij}c_i-\lambda=0, \qquad j=m+1,\ldots,m+p,$$

$$-\sum_{i=m+1}^{m+p} c_i=-1.$$

(Кенуй, 1953b.)

46.10 Ряд состоит из полинома от t степени m и компоненты, у которой дисперсия для всех t одна и та же, но последовательные члены могут быть сериально коррелированы.

Пусть $R_{1p}=(p+1)(p+2)\Delta V_p/2!$, $p=m, m+1, \ldots$ Показать, что для больших рядов

$$R_{1p}=V_0\left\{r_1-\frac{4p}{p+3}r_2+\frac{9p(p-1)}{(p+3)(p+4)}r_3-\frac{16p(p-1)(p-2)}{(p+3)(p+4)(p+5)}r_4+\ldots\right\}$$

и, следовательно, что $\mathbf{M}(R_{1p})=\sigma^2 r_1$, если сериальные корреляции порядка больше первого равны нулю.

46.11 Вновь обращаясь к предыдущему примеру, показать, что если сериальные корреляции порядка, большего m, равны нулю, $\mathbf{M}(R_{mp})=\sigma^2 r_m$, где

$$R_{mp}=\frac{(p+2m-1)(p+2m)}{(2m-1)(2m)}\Delta R_{m-1,p}.$$

(Кенуй, 1953b.)

46.12 Обобщая сказанное в **46.34**, показать, что если постоянные a определить соотношением

$$(1+t+t^2+\ldots+t^{s-1})^{d+1}=a_0+a_1t+a_2t^2+\ldots,$$

то группы весов

$$\sum a_{is}b_{i+1}, \quad \sum a_{is-1}b_{i+1}, \quad \sum a_{is-2}b_{i+1}$$

определяют выравнивание кривой степени d по группам из s точек.

(Кенуй, 1949a.)

46.13 Пусть циркулянтная матрица

$$A=\{a_{i-j}\}, \qquad i, j=0, 1, \ldots, n-1,$$

где индексы складываются по модулю n. Показать, что

$$|A|=\prod_{s=0}^{n-1}\sum_{r=0}^{n-1} a_r\omega^{rs},$$

где $\omega=\exp\{2\pi i/n\}$ (n-й корень -1). Показать также, что собственные числа λ матрицы A даются соотношениями

$$\lambda_s=\sum_r a_r\omega^{rs}, \qquad s=0, 1, \ldots, n-1,$$

а

$$a_s=\frac{1}{n}\sum_{r=0}^{n-1}\lambda_r\bar{\omega}^{rs}.$$

Учитывая, что собственные числа A^{-1} равны λ^{-1}, показать, как обратить A.

(Гуд, 1950.)

46.14 Воспользовавшись соотношением

$$(-1)^p\cos 2p\theta=1-\frac{p^2}{2!}(2\cos\theta)^2+\frac{p^2(p^2-1)}{4!}(2\cos\theta)^4-\ldots$$

и положив $z=e^{i\theta}$, вывести равенство (46.93).

46.15 Последовательность значений ряда u_t, заданных при $t=-m, \ldots, 0, \ldots, m$, выравнивается методом наименьших квадратов кубическим полиномом $\sum_{j=0}^{3} a_j t^j$. Показать, что последний равен

$$\frac{15}{n(n^2-1)(n^2-4)}\left\{\frac{1}{20}(n^2-1)(3n^2-7)\sum u-(n^2-1)\sum t^2u\right\}+$$

$$+\frac{140t}{n(n^2-1)(n^2-4)(n^2-9)}\left\{\frac{5}{28}(3n^4-18n^2+31)\sum tu_t-(3n^2-7)\sum t^3u\right\}+$$

$$+\frac{15t^2}{n(n^2-1)(n^2-4)}\left\{-(n^2-1)\sum u+12\sum t^2u\right\}+$$

$$+\frac{140t^3}{n(n^2-1)(n^2-4)(n^2-9)}\left\{-(3n^2-7)\sum tu_t+20\sum t^3u\right\} \quad (n=2m+1).$$

46.16 У случайного ряда «исключен тренд» методом скользящих средних с весами $[a_{-m}, a_{-(m-1)}, \ldots, a_0]$. Показать, что сериальные корреляции остатков приблизительно равны величинам

$$r_k = \frac{\sum_{i=-m}^{m-k} a_i a_{i+k} - 2a_k}{\sum_{i=-m}^{m} a_i^2 - 2a_0 + 1}.$$

Сравнить значения, полученные после применения этих формул к данным таблицы 46.1 (ряд (10)), с истинными значениями

$$r_1 = -0{,}411, \quad r_2 = -0{,}244, \quad r_3 = 0{,}231,$$
$$r_4 = 0{,}143, \quad r_5 = 0{,}007.$$

ГЛАВА 47

СТАЦИОНАРНЫЕ ВРЕМЕННЫЕ РЯДЫ

47.1 После исключения из временного ряда компонент, соответствующих тренду и сезонным изменениям, остается ряд, представляющий осцилляции около некоторой постоянной величины. Последние могут оказаться пренебрежимо малыми, и тогда следует признать, что на самом деле первоначальный ряд состоял лишь из тренда и сезонной компоненты. Напротив, в ряде может отсутствовать тренд и сезонность, и тогда мы имеем дело с чистыми осцилляциями. В настоящей главе мы будем изучать такие осцилляции, предполагая, что тренд и сезонность или отсутствуют или исключены. Строго говоря, сезонные эффекты следует также рассматривать как осцилляции и не исключать их заранее. Но, как мы позже увидим, осцилляции, носящие сезонный характер, составляют скорее исключение, чем правило, и лучше всего, настолько, насколько это возможно, рассматривать их отдельно.

47.2 Начнем с некоторых интуитивных предположений. Таблица 45.1 (урожайность ячменя) дает нам пример флуктуаций относительно среднего значения, которые происходят в некоторых ограниченных пределах. С другой стороны, можно встретить и ряды, где размах осцилляций систематически растет (как, например, на рис. 47.1, а) или сам претерпевает колебания (как на рис. 47.1, б). Мы исключим такие случаи из рассмотрения и сконцентрируем наше внимание на рядах, у которых амплитуда остается более или менее постоянной во времени. Это не означает, что амплитуда колебаний должна быть всегда одной и той же, мы исключаем лишь случай систематических изменений.

47.3 Чтобы сделать наши рассуждения строгими, рассмотрим упорядоченную последовательность случайных величин $u_1, u_2, \ldots, u_n$. Пусть функция распределения любых n последовательных величин u, скажем $u_{t+1}, \ldots, u_{t+n}$, равна

$$F(u_{t+1}, \ldots, u_{t+n}). \tag{47.1}$$

Если F не зависит от t при всех целых $n > 0$, будем говорить, что F определяет *стационарный ряд*. В этом случае любые n последовательных величин имеют одно и то же распределение независимо от места, которое они занимают в ряде.

46.16 У случайного ряда «исключен тренд» методом скользящих средних с весами $[a_{-m}, a_{-(m-1)}, \ldots, a_0]$. Показать, что сериальные корреляции остатков приблизительно равны величинам

$$r_k = \frac{\sum_{i=-m}^{m-k} a_i a_{i+k} - 2a_k}{\sum_{i=-m}^{m} a_i^2 - 2a_0 + 1}.$$

Сравнить значения, полученные после применения этих формул к данным таблицы 46.1 (ряд (10)), с истинными значениями

$$r_1 = -0{,}411, \quad r_2 = -0{,}244, \quad r_3 = 0{,}231,$$
$$r_4 = 0{,}143, \quad r_5 = 0{,}007.$$

ГЛАВА 47

СТАЦИОНАРНЫЕ ВРЕМЕННЫЕ РЯДЫ

47.1 После исключения из временного ряда компонент, соответствующих тренду и сезонным изменениям, остается ряд, представляющий осцилляции около некоторой постоянной величины. Последние могут оказаться пренебрежимо малыми, и тогда следует признать, что на самом деле первоначальный ряд состоял лишь из тренда и сезонной компоненты. Напротив, в ряде может отсутствовать тренд и сезонность, и тогда мы имеем дело с чистыми осцилляциями. В настоящей главе мы будем изучать такие осцилляции, предполагая, что тренд и сезонность или отсутствуют или исключены. Строго говоря, сезонные эффекты следует также рассматривать как осцилляции и не исключать их заранее. Но, как мы позже увидим, осцилляции, носящие сезонный характер, составляют скорее исключение, чем правило, и лучше всего, настолько, насколько это возможно, рассматривать их отдельно.

47.2 Начнем с некоторых интуитивных предположений. Таблица 45.1 (урожайность ячменя) дает нам пример флуктуаций относительно среднего значения, которые происходят в некоторых ограниченных пределах. С другой стороны, можно встретить и ряды, где размах осцилляций систематически растет (как, например, на рис. 47.1, а) или сам претерпевает колебания (как на рис. 47.1, б). Мы исключим такие случаи из рассмотрения и сконцентрируем наше внимание на рядах, у которых амплитуда остается более или менее постоянной во времени. Это не означает, что амплитуда колебаний должна быть всегда одной и той же, мы исключаем лишь случай систематических изменений.

47.3 Чтобы сделать наши рассуждения строгими, рассмотрим упорядоченную последовательность случайных величин $u_1, u_2, \ldots, u_n$. Пусть функция распределения любых n последовательных величин u, скажем $u_{t+1}, \ldots, u_{t+n}$, равна

$$F(u_{t+1}, \ldots, u_{t+n}). \tag{47.1}$$

Если F не зависит от t при всех целых $n > 0$, будем говорить, что F определяет *стационарный ряд*. В этом случае любые n последовательных величин имеют одно и то же распределение независимо от места, которое они занимают в ряде.

Положив в частности $n = 1$, мы видим, что все члены ряда одинаково распределены.

В общей теории случайных процессов, частным случаем которых являются стационарные временные ряды, принято определять стационарность в более широком смысле. Процесс называется *стационарным по среднему*, если все величины u имеют одно и то же математическое ожидание, и *стационарным по дисперсии*, если все u имеют одну и ту же дисперсию. Наиболее часто мы будем использовать следствие свойства стационарности (47.1), касающееся

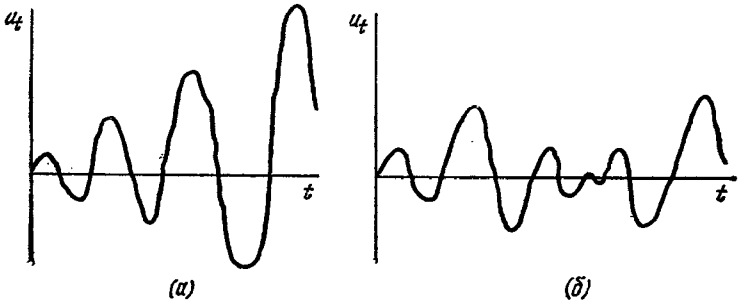

Рис. 47.1 (см. текст).

постоянства средних и дисперсии членов ряда, однако весьма важную роль будут играть и смешанные моменты величин u, а также то обстоятельство, что все величины u имеют одно и то же распределение.

47.4 Итак, мы рассматриваем определенную функцией распределения (47.1), вообще говоря, неограниченную последовательность случайных величин. В этом случае при определении средних значений мы сталкиваемся с новой проблемой. Во-первых, мы можем рассматривать значения какого-либо одного u, например u_1, в различных рядах, которым соответствует (47.1). Во-вторых, можно рассматривать предельное поведение некоторой последовательности величин u (или функции от этих величин), скажем $u_1, \ldots, u_n$ при $n \to \infty$.

В первом случае мы должны рассматривать среднее значений u_1 в различных рядах, полученных любыми способами, не противоречащими (47.1). Каждый такой ряд в принципе возможен, и любой такой практически реализованный ряд мы будем называть *реализацией* процесса. Однако на практике, как правило, мы имеем дело лишь с одной реализацией. Иначе говоря, в определенном смысле нам дана выборка, состоящая лишь из одного наблюдения над процессом. Вскоре мы увидим, почему это обстоятельство не ограничивает серьезно возможности исследования.

47.5 Здесь и в дальнейшем мы будем предполагать, что среднее и дисперсия величин u существуют. Тогда для всех t

имеем

$$\mu = \mathsf{M}u_t = \int_{-\infty}^{\infty} u_t \, dF(u_t), \tag{47.2}$$

$$\sigma^2 = \mathsf{M}(u_t - \mu)^2 = \int_{-\infty}^{\infty} (u_t - \mu)^2 \, dF(u_t). \tag{47.3}$$

Будем также предполагать, что для любой пары величин u существуют *автоковариации*

$$\gamma_j = \mathsf{M}\{(u_t - \mu)(u_{t+j} - \mu)\} = \gamma_{-j} \tag{47.4}$$

и соответствующая *автокорреляция*

$$\rho_j = \gamma_j / \sigma^2 = \rho_{-j}. \tag{47.5}$$

Как мы отмечали в **45.33**, совокупность коэффициентов $\rho_0 (= 1), \rho_1, \rho_2, \ldots$ называют коррелограммой ряда. Мы должны отличать теоретическую коррелограмму, основанную на автокорреляциях, от наблюденной коррелограммы, основанной на сериальных корреляциях некоторого ряда длины n.

47.6 Для любого заданного n теоретическая корреляционная матрица последовательности $u_t, u_{t+1}, \ldots$ является матрицей Лорана

$$\begin{pmatrix} 1 & \rho_1 & \rho_2 & \ldots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \ldots & \rho_{n-2} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \ldots & 1 \end{pmatrix}. \tag{47.6}$$

Элементы любой диагонали, идущей с «северо-запада на юго-восток» равны между собой.

Пример 47.1

Из неотрицательной определенности матрицы Лорана следуют определенные условия *совместимости* коэффициентов автокорреляции. Определитель любого минора, основанного на главной диагонали, не может быть отрицательным. Так, например,

$$\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{vmatrix} = 1 - \rho_1^2 \geqslant 0$$

(тривиальный результат),

$$\begin{vmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{vmatrix} = 1 + 2\rho_1^2(\rho_2 - 1) - \rho_2^2 = (1 - \rho_2)(1 - 2\rho_1^2 + \rho_2) \geqslant 0. \tag{47.7}$$

Таким образом, если $\rho_2 \neq 1$ (и даже в противном случае), то

$$\rho_2 \geqslant 2\rho_1^2 - 1,$$

что уже никоим образом не является тривиальным результатом.

Пример 47.2

В качестве примера схемы, порождающей автокоррелированный стационарный ряд, рассмотрим процесс u_t такой, что

$$u_{t+1} = \rho u_t + \varepsilon_{t+1}, \tag{47.8}$$

где ε — случайная величина с нулевым средним, и величины ε_p и ε_q не коррелированы при $p \neq q$. Тогда

$$\mathsf{M}u_{t+1} = \rho \mathsf{M}u_t,$$

и, возможно, за исключением тривиального случая $\rho = 1$, из условия стационарности следует, что при всех t

$$\mathsf{M}u_t = 0. \tag{47.9}$$

Из (47.8) следует, что u_t зависит от ε_t, ε_{t-1}, ε_{t-2} и т. д., но не зависит от ε_{t+1}. Поэтому

$$\mathsf{M}\{u_t(u_{t+1} - \rho u_t)\} = \mathsf{M}(u_t \varepsilon_{t+1}) = 0,$$

и, следовательно,

$$\operatorname{cov}(u_t, u_{t+1}) = \rho \mathsf{D}u_t = \rho\sigma^2,$$

и корреляция между u_t и u_{t+1} равна ρ:

$$\mathsf{M}\{u_t u_{t+1}\}/\mathsf{M}u_t^2 = \rho. \tag{47.10}$$

Для ρ_k — автокорреляции k-го порядка точно также получаем

$$\mathsf{M}\{u_t(u_{t+k} - \rho u_{t+k-1})\} = \sigma^2(\rho_k - \rho\rho_{k-1}) = 0,$$

и, следовательно,

$$\rho_k = \rho^k. \tag{47.11}$$

Пример 47.3

Из того факта, что матрица Лорана симметрична относительно главной диагонали, следует одно несколько неожиданное свойство схем, построенных в духе предыдущего примера. А именно, такие схемы можно столь же хорошо определять обратными соотношениями, т. е. соотношениями типа

$$u_t = \rho u_{t+1} + \eta_t, \tag{47.12}$$

где η_t — какая-то другая случайная величина. Как и прежде, легко видеть, что u_t не зависит от η_{t-1} и, следовательно, из (47.12), заменив t на $t-1$, получаем, что

$$\mathsf{M}\{u_t(u_{t-1} - \rho u_t)\} = \mathsf{M}(u_t \eta_{t-1}) = 0,$$

откуда снова следует, что первая автокорреляция равна ρ.

47.7 Во временных рядах переход от теоретических распределений к выборочным и изменение направления порождают несколько новых задач, которые мы позже детально рассмотрим. Но прежде всего, без ущерба для дальнейшего обсуждения, приведем несколько примеров выборочных рядов в качестве иллюстраций к тому материалу, с которым чаще всего имеет дело практика. Некоторые из таких примеров содержатся в главе 45; так, на первый взгляд, имеют стационарный характер ряды из таблиц 45.1 (урожайность ячменя) и 45.2 (осадки). Таблица 47.1 содержит знаменитый ряд освобожденных от тренда индексов цен на пшеницу, составленный последним лордом Бевериджем. Этот ряд охватывает 370 лет, что является феноменальной длиной для экономических рядов. В таблице 47.2 приведены отклонения от простых скользящих средних по 11 элементам для долей совершаемых в Англии и Уэльсе браков (на 10 000 населения) за 1843—1896 гг. Таблица 47.3 дает нам искусственный ряд, полученный наложением случайных флуктуаций на простую гармонику. Таблица 47.4 содержит другой искусственный ряд, полученный по более сложной схеме типа схемы примера 47.2.

47.8 Допустим теперь, что нам дан ряд наблюдений $u'_{t_1}, \ldots, u'_{t}$. Штрих обозначает, что имеется лишь одна реализация. Каждое u имеет среднее μ и дисперсию σ^2. Определим временные средние:

$$\mathsf{E}_{t_2-t_1+1}(u') = \frac{1}{t_2 - t_1 + 1} \sum_{t_1}^{t_2} u'_t, \tag{47.13}$$

$$\mathsf{E}(u') = \lim_{t_1 \to -\infty,\ t_2 \to +\infty} \mathsf{E}_{t_2-t_1+1}(u'). \tag{47.14}$$

Обратимся теперь к теореме Биркгофа (1931) и Хинчина (1932), которую приведем без доказательства.

Пусть u_t — стационарный ряд с конечным средним μ. Тогда $\mathsf{E}(u')$ существует для почти всех реализаций, т. е. с вероятностью единица. Кроме того, при выполнении и только при выполнении условия

$$\lim_{n \to \infty} \frac{1}{n} \sum_{j=1}^{n} \rho_j = 0 \tag{47.15}$$

временное среднее $\mathsf{E}(u')$ равно среднему $\mathsf{M}u$, т. е.

$$\mathsf{M}u_t = \mathsf{E}u'_t. \tag{47.16}$$

Это очень важный результат. Он означает, что на практике мы можем оценивать среднее u, вычисляя среднее последовательных значений единственной реализации. Если бы это было не

Таблица 47.1

«Свободные от тренда» индексы цен на пшеницу (европейские цены) на 1500—1869 гг. Ряд составлен последним лордом (тогда сэр Уилльям) Бевериджем

Год	Индекс	Год	Индекс	Год	Индекс	Год	Индекс	Год	Индекс
1500	106	1537	73	1574	113	1611	100	1648	122
01	118	38	86	75	89	12	99	49	134
02	124	39	74	76	87	13	100	50	119
03	94	40	74	77	87	14	94	51	136
04	82	41	76	78	79	15	88	52	102
05	88	42	80	79	90	16	92	53	72
06	87	43	96	80	90	17	100	54	63
07	88	44	112	81	87	18	82	55	76
08	88	45	144	82	83	19	73	56	75
09	68	46	80	83	85	20	81	57	77
10	98	47	54	84	76	21	99	58	103
11	115	48	69	85	110	22	124	59	104
12	135	49	100	86	161	23	106	60	120
13	104	50	103	87	97	24	106	61	167
14	96	51	129	88	84	25	121	62	126
15	110	52	100	89	106	26	105	63	108
16	107	53	90	90	111	27	84	64	91
17	97	54	100	91	97	28	97	65	85
18	75	55	123	92	108	29	109	66	73
19	86	56	156	93	100	30	148	67	74
20	111	57	71	94	119	31	114	68	80
21	125	58	71	95	131	32	108	69	74
22	78	59	81	96	143	33	97	70	78
23	86	60	84	97	138	34	92	71	83
24	102	61	97	98	112	35	97	72	84
25	71	62	105	99	99	36	98	73	106
26	81	63	90	1600	97	37	105	74	134
27	129	64	78	01	80	38	97	75	122
28	130	65	112	02	90	39	93	76	102
29	129	66	100	03	90	40	99	77	107
30	125	67	86	04	80	41	99	78	115
31	139	68	77	05	77	42	107	79	113
32	97	69	80	06	81	43	106	80	104
33	90	70	93	07	98	44	96	81	92
34	76	71	112	08	115	45	82	82	84
35	102	72	131	09	94	46	88	83	86
36	100	73	158	10	93	47	116	84	101

Год	Индекс	Год	Индекс	Год	Индекс	Год	Индекс	Год	Индекс
1685	74	1722	91	1759	91	1796	95	1833	80
86	75	23	94	60	88	97	84	34	78
87	66	24	110	61	100	98	87	35	82
88	62	25	111	62	97	99	120	36	88
89	76	26	103	63	88	1800	139	37	102
90	79	27	94	64	95	01	117	38	117
91	97	28	101	65	101	02	105	39	107
92	134	29	90	66	106	03	94	40	95
93	169	30	96	67	113	04	125	41	101
94	111	31	80	68	108	05	114	42	92
95	109	32	76	69	108	06	98	43	88
96	111	33	84	70	131	07	93	44	92
97	128	34	91	71	136	08	94	45	115
98	163	35	94	72	119	09	94	46	139
99	137	36	101	73	106	10	104	47	90
1700	99	37	93	74	105	11	140	48	80
01	85	38	91	75	88	12	121	49	74
02	72	39	122	76	84	13	96	50	78
03	88	40	159	77	94	14	96	51	86
04	77	41	110	78	87	15	130	52	105
05	66	42	90	79	79	16	178	53	138
06	64	43	81	80	87	17	126	54	141
07	69	44	84	81	88	18	94	55	138
08	125	45	102	82	94	19	86	56	107
09	175	46	102	83	94	20	84	57	82
10	108	47	100	84	92	21	76	58	81
11	103	48	109	85	85	22	77	59	97
12	115	49	104	86	84	23	71	60	116
13	134	50	90	87	93	24	71	61	107
14	108	51	99	88	108	25	69	62	92
15	90	52	95	89	108	26	82	63	79
16	89	53	90	90	86	27	93	64	81
17	89	54	89	91	78	28	114	65	94
18	94	55	85	92	87	29	103	66	119
19	107	56	117	93	85	30	110	67	118
20	89	57	112	94	103	31	105	68	93
21	79	58	95	95	130	32	82	69	102

Таблица 47.2

Доли браков в Англии и Уэльсе: отклонения от простых скользящих средних по 11 точкам за 1843—1896 гг. (на 10 000)

Год	Доля браков	Год	Доля браков	Год	Доля браков	Год	Доля браков	Год	Доля браков
1843	−6	1854	3	1865	8	1876	4	1887	−6
44	1	55	−8	66	9	77	−3	88	−5
45	12	56	−2	67	−2	78	−6	89	1
46	10	57	−3	68	−8	79	−12	90	6
47	−6	58	−7	69	−10	80	−5	91	6
48	−8	59	3	70	−7	81	0	92	2
49	−6	60	4	71	0	82	5	93	−6
50	3	61	−5	72	8	83	7	94	−5
51	4	62	−7	73	12	84	3	95	−6
52	7	63	1	74	7	85	−4	96	1
53	11	64	6	75	5	86	−8		

так, оценивание по одной реализации было бы практически невозможным.

47.9 Если $\mathsf{E}u' = \mathsf{M}u$ и второй момент $\mathsf{M}(u-\mu)^2$ конечен, процесс называется *эргодическим*. Сформулируем снова без доказательства одно важное обобщение теоремы Биркгофа — Хинчина: пусть процесс эргодический, тогда для почти всех реализаций автокорреляция равна корреляции, вычисленной по реализации, т. е. равна

$$\frac{\mathsf{E}\{u_t' - \mathsf{E}u'\}\{u_{t+j}' - \mathsf{E}u'\}}{\mathsf{E}\{u_t' - \mathsf{E}u'\}^2}. \tag{47.17}$$

Этот результат снова позволяет нам оценивать автокорреляции по единственной реализации процесса.

Условие (47.15) не слишком обременительно, но и не является практически всеобъемлющим. Оно выполняется, если, например, автокорреляция стремится к нулю по мере удаления друг от друга членов, которым она соответствует, но оказывается невыполненным, если, скажем, наш ряд гармоничен. В действительности существуют другие условия, выполнения которых достаточно для эргодичности.

47.10 Как мы позже увидим, коррелограмма оказывается полезным инструментом при изучении внутренней структуры временных рядов. Существует еще одна функция, служащая сходным целям. Ее связь с коррелограммой во многом такая же, как связь характеристической функции с плотностью.

Именно, определим функцию

$$W(\alpha) = \alpha + 2\sum_{j=1}^{\infty} \frac{\rho_j \sin \alpha j}{j}. \tag{47.18}$$

Таблица 47.3

Значения ряда $u_t = 10 \sin(\pi t/5) + \varepsilon_t$, где ε_t — случайная величина, равномерно распределенная на отрезке от −5 до +5. (Значения округлены до ближайшей единицы.)

Номер члена	Значение члена	Номер члена	Значение члена	Номер члена	Значение члена	Номер члена	Значение члена	Номер члена	Значение члена
1	3	13	6	25	−5	37	−9	49	−8
2	8	14	4	26	−8	38	−6	50	−1
3	6	15	−3	27	−12	39	−4	51	11
4	2	16	−10	28	−10	40	−2	52	13
5	−4	17	−11	29	−7	41	5	53	12
6	−7	18	−15	30	0	42	12	54	7
7	−9	19	−4	31	1	43	7	55	5
8	−9	20	4	32	8	44	5	56	−1
9	−10	21	11	33	13	45	3	57	−6
10	−1	22	13	34	7	46	−2	58	−14
11	8	23	10	35	4	47	−12	59	−8
12	7	24	6	36	−9	48	−12	60	1

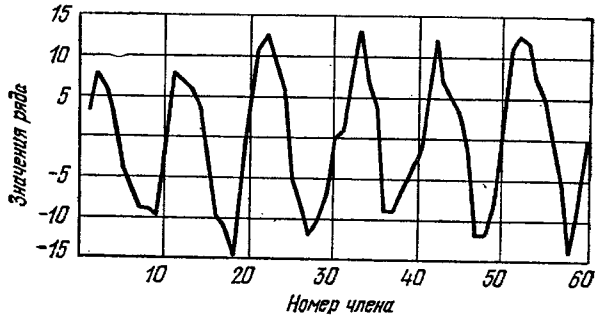

Рис. 47.2. График значений ряда из таблицы 47.3.

Если, начиная с некоторого n и далее $|\rho_{n+k}| < 1$, последний ряд сходится. Мы будем также рассматривать функцию

$$w(\alpha) = \frac{dW}{d\alpha} = 1 + 2\sum_{j=1}^{\infty} \rho_j \cos \alpha j \tag{47.19}$$

в тех случаях, когда последняя существует.

Учитывая, что $\rho_j = \rho_{-j}$ и $\sin\theta$ — нечетная функция от θ, получаем, что

$$w(\alpha) = \sum_{-\infty}^{\infty} \rho_j \cos \alpha j = \tag{47.20}$$

$$= \sum_{-\infty}^{\infty} \rho_j e^{i\alpha j}. \tag{47.21}$$

Таблица 47.4

Значения ряда $u_t = 1{,}1u_{t-1} - 0{,}5u_{t-2} + \varepsilon_t$, где ε_t — случайная величина, равномерно распределенная на отрезке от —9,5 до 9,5. (Значения округлены до ближайшей единицы.)

Номер члена	Значение члена	Номер члена	Значение члена	Номер члена	Значение члена	Номер члена	Значение члена	Номер члена	Значение члена
1	7	14	—7	27	—4	40	—8	53	4
2	6	15	—2	28	3	41	—3	54	—1
3	—6	16	6	29	9	42	—8	55	4
4	—4	17	17	30	4	43	—10	56	7
5	3	18	24	31	—8	44	—16	57	11
6	—4	19	17	32	—6	45	—13	58	0
7	—5	20	4	33	—3	46	1	59	1
8	—1	21	1	34	—2	47	6	60	0
9	10	22	—5	35	0	48	4	61	—5
10	10	23	—4	36	—1	49	11	62	—11
11	6	24	—5	37	—3	50	15	63	—8
12	—4	25	—9	38	3	51	9	64	—3
13	—4	26	—4	39	—1	52	8	65	5

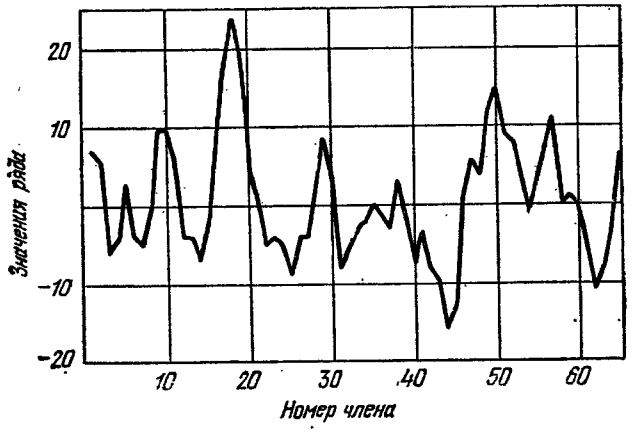

Рис. 47.3. График значений ряда из таблицы 47.4.

Последняя форма определяет $w(\alpha)$ как преобразование Фурье последовательности ρ_j. Умножая (47.20) на $\cos \alpha k$ и почленно интегрируя, получаем, что

$$\int_0^\pi w(\alpha) \cos k\alpha \, d\alpha = \sum_{j=-\infty}^{\infty} \rho_j \int_0^\pi \cos \alpha j \cos \alpha k \, d\alpha = \pi \rho_k,$$

и, следовательно,

$$\rho_j = \frac{1}{\pi} \int_0^\pi w(\alpha) \cos j\alpha \, d\alpha = \frac{1}{\pi} \int_0^\pi \sin \alpha j \, dW(\alpha). \qquad (47.22)$$

Можно также писать, что

$$\rho_j = \frac{1}{\pi} \int_0^\pi w(\alpha) e^{-i\alpha j} \, d\alpha. \qquad (47.23)$$

$W(\alpha)$ называется *спектральной функцией*. Ее производная $w(\alpha)$ называется *спектральной плотностью*. График $w(\alpha)$, как функции от α, называют *спектром* (или *спектром мощности*). $w(\alpha)$ — периодическая функция с периодом 2π. Из (47.18) следует, что $W(0) = 0$, $W(\pi) = \pi$, $W(2\pi) = 2\pi$.

47.11 Спектральную плотность можно также определить непосредственно, не обращаясь к коррелограмме, следующим образом.

Рассмотрим корреляцию или ковариацию u_t и $\cos \alpha t$ при некотором фиксированном α. Если u_t в некотором смысле ритмичен (или псевдоритмичен — давайте, сейчас вопросов не задавать) с частотой α, корреляция будет большой, поскольку u_t и $\cos \alpha t$ колеблются в фазе. Пусть t принимает целые значения и ряд центрирован своим средним. Рассмотрим функции

$$\left.\begin{aligned} a(\alpha) &= \frac{1}{\sqrt{n\pi}} \sum_{t=1}^{n} u_t \cos \alpha t, \\ b(\alpha) &= \frac{1}{\sqrt{n\pi}} \sum_{t=1}^{n} u_t \sin \alpha t. \end{aligned}\right\} \qquad (47.24)$$

Имеем

$$I(\alpha) = a^2(\alpha) + b^2(\alpha) = \frac{1}{n\pi} \left\{ \left(\sum u_t \cos \alpha t \right)^2 + \left(\sum u_t \sin \alpha t \right)^2 \right\} =$$

$$= \frac{1}{n\pi} \left\{ \sum u_t^2 + 2 \sum_{k=1}^{n-1} \sum_{t=1}^{n-k} u_t u_{t+k} \{ \cos \alpha t \cos \alpha (t+k) + \right.$$

$$\left. + \sin \alpha t \sin \alpha (t+k) \} \right\} = \frac{1}{n\pi} \left\{ \sum u_t^2 + 2 \sum_{k=1}^{n-1} \sum_{t=1}^{n-k} u_t u_{t+k} \cos k\alpha \right\} =$$

$$= \frac{s^2}{\pi} \left\{ 1 + 2 \sum_{k=1}^{n-1} r_k \cos k\alpha \right\}, \qquad (47.25)$$

где $s^2 = \sum u_t^2/n$, а $r_k = \sum u_t u_{t+k}/\sum u_t^2$ — коэффициент корреляционного типа. Переходя к пределу, получаем *)

$$I(\alpha) = \frac{\sigma^2}{\pi}\left\{1 + 2\sum_{1}^{\infty} \rho_k \cos k\alpha\right\} = \tag{47.26}$$

$$= \frac{\sigma^2}{\pi}\left\{\sum_{-\infty}^{\infty} \rho_k \cos k\alpha\right\} = \frac{\sigma^2}{\pi} w(\alpha). \tag{47.27}$$

Величина I, которую мы назовем *интенсивностью*, равна, таким образом, спектральной плотности, умноженной на σ^2/π.

47.12 Когда чертят график спектра, обычно по оси ординат откладывают величину I, а по оси абсцисс α. На ранних стадиях развития теории было более принято вычислять величины

$$A = \frac{2}{n}\sum_{t=1}^{n} u_t \cos\frac{2\pi t}{\lambda}, \qquad \lambda = 2\pi/\alpha, \tag{47.28}$$

$$B = \frac{2}{n}\sum_{t=1}^{n} u_t \sin\frac{2\pi t}{\lambda}, \tag{47.29}$$

и затем предельное значение величины

$$S^2 = A^2 + B^2 = \frac{4}{n}\sigma^2 w(\lambda). \tag{47.30}$$

График зависимости S^2 от длины волны λ был назван периодограммой. Мы будем придерживаться той же терминологии, хотя некоторые авторы и называют периодограммой то, что мы называем спектром.

Отметим, что, в то время как делитель в (47.24) имеет порядок $\sqrt{n}$, делитель в (47.29) имеет порядок n. Это объясняется тем, что если u содержит гармоническую компоненту с амплитудой c, скажем $c\cos\alpha t$, а другие компоненты ряда с ней не коррелированы, значение S^2 в точке α равно c^2. У спектра соответствующая ордината будет, по крайней мере для бесконечных рядов, бесконечной. Мы еще вернемся к этому вопросу в главе 49.

47.13 В какой мере полезны коррелограмма или спектр при исследовании внутренней структуры временных рядов, зависит

*) (47.27) надо понимать условно: $I(\alpha)$ — величина случайная, и ее распределение в пределе не вырождено; (47.27) дает в сущности лишь предельное среднее $I(\alpha)$, а замена $\Sigma u_t u_{t+k}/(ns^2)$ на ρ_k неправомерна и приводит к исключению не исчезающих в пределе суммарных флуктуаций. Авторы коснутся этого вопроса в 49.11. (*Прим. перев.*)

до некоторой степени от наших целей и априорных знаний о порождающей эти ряды системе. Вообще говоря, в экономических задачах более наглядными оказываются коррелограммы, в физических — спектры, но существуют области, где предусмотрительный исследователь использовал бы обе характеристики системы (например, океанография, метеорология, некоторые биологические процессы). Как мы уже отмечали, коррелограмма говорит нам нечто о зависимости между членами ряда, разделенными во времени. Спектр же указывает на то, в какой степени ряд подчиняется тому или иному основному ритму. Вычисление спектра похоже на настройку радиоприемника: сигнал большой мощности хорошо слышен тогда, когда частота настройки совпадает с входной частотой. По этой причине пики спектра, если такие есть, иногда отождествляют с гармоническими составляющими порождающей ряд первичной системы. Такого рода интерпретация, однако, требует большой осторожности.

Производящая функция автокорреляций

47.14 В представлении спектральной плотности

$$w(\alpha) = \sum_{-\infty}^{\infty} \rho_j e^{i\alpha j} \tag{47.31}$$

положим

$$z = e^{i\alpha}. \tag{47.32}$$

Тогда $w(\alpha)$ как функция от z представима в виде

$$w(\alpha) = G(z) = \sum_{-\infty}^{\infty} \rho_j z^j. \tag{47.33}$$

Мы получили таким образом производящую функцию автокорреляций $G(z)$. Полезно также определить производящую функцию автоковариаций:

$$C(z) = \sum_{-\infty}^{\infty} \gamma_j z^j. \tag{47.34}$$

Ряды скользящих средних

47.15 Рассмотрим теперь ряд u_t и скользящие средние, определяемые формулой

$$\zeta_t = \sum_{i=0}^{\infty} \alpha_i u_{t-i}. \tag{47.35}$$

Чтобы достигнуть общности, мы здесь приводим скользящее усреднение по бесконечному числу точек. Необходимо отметить,

что ряд ζ_t уже может и не быть эргодическим. Тем не менее он будет таковым, если ряд $\sum \alpha_i^2$ сходится. Последнее условие будем предполагать выполненным. Тогда

$$\mathsf{M}(\zeta_t\zeta_{t+j}) = \mathsf{M}\left\{\sum_{i=0}^{\infty} \alpha_i u_{t-i} \sum_{k=0}^{\infty} \alpha_k u_{t+j-k}\right\} = \sum \alpha_i\alpha_k M(u_{t-i}u_{t+j-k}) =$$

$$= \sum_{i,k=0}^{\infty} \alpha_i\alpha_k\gamma_{i+j-k} = \sum_{i=0}^{\infty}\sum_{l=0}^{\infty} \alpha_i\alpha_{i+j-l}\gamma_l. \qquad (47.36)$$

Легко видеть, что если производящая функция автоковариаций ряда u равна $C(z)$, то производящая функция автоковариаций ряда ζ равна

$$\Gamma(z) = \left(\sum \alpha_j z^j\right)\left(\sum \alpha_j z^{-j}\right) C(z). \qquad (47.37)$$

В частности, если u — «чисто случайный» ряд ($C(z) = 1$), после k-кратного скользящего усреднения автоковариационная функция нового ряда становится равной

$$\left(\sum \alpha_j z^j\right)^k \left(\sum \alpha_j z^{-j}\right)^k. \qquad (47.38)$$

Пример 47.4

Рассмотрим скользящие средние случайного ряда по двум элементам $\left(\alpha_0 = \alpha_1 = \frac{1}{2}\right)$.

$$\Gamma(z) = \frac{1}{4}(1+z)(1+z^{-1}) = \frac{1}{4}(z+2+z^{-1}). \qquad (47.39)$$

Отсюда следует, что $\rho_1 = 1/2$; $\rho_j = 0$, $j \neq 0, 1$. Последнее, впрочем, ясно и из других соображений. Соответствующую автокорреляционную функцию получим, нормируя (47.39) так, чтобы коэффициент при z^0 стал равен единице. Таким образом, $G(z) = \frac{1}{2}(z^{-1}+2+z)$. Положим теперь $z = e^{i\alpha}$ и найдем выражение для спектральной плотности:

$$w(\alpha) = \frac{1}{2}(e^{-i\alpha}+2+e^{i\alpha}) = 1+\cos\alpha. \qquad (47.40)$$

Полученная функция является косинусоидой с максимумом в точке $\alpha = 0$. Если мы повторим усреднение k раз, то получим, что

$$w(\alpha) \propto (1+\cos\alpha)^k. \qquad (47.41)$$

Коэффициент пропорциональности легче всего найти из условия

$$\int_0^{\pi} w(\alpha)\,d\alpha = \pi$$

(см. **47.10**). В нашем случае это приводит к формуле *)

$$w(\alpha) = \frac{\pi^{1/2}\Gamma(k+1)}{\Gamma\left(k+\frac{1}{2}\right)}\left(\frac{1+\cos\alpha}{2}\right)^k.$$

Последнее выражение с ростом k стремится к бесконечности при $\alpha = 0$ и к нулю в любом другом случае. Все ρ_j стремятся к единице. Члены исходного ряда, таким образом, с ростом k сходятся к одной и той же постоянной величине, что легко усмотреть также из того факта, что последовательные итерации сглаживают флуктуации членов ряда.

С другой стороны, если брать последовательные *разности* членов первоначального случайного ряда $\left(\alpha_0 = \frac{1}{2},\ \alpha_1 = -\frac{1}{2}\right)$, то после вычисления k-х разностей

$$w(\alpha) \propto \left(\frac{1-\cos\alpha}{2}\right)^k. \qquad (47.42)$$

$w(\alpha)$ стремится к бесконечности при $\alpha = \pi$. Автокорреляции четного порядка стремятся к $+1$, нечетного к -1. Таким образом, соседние члены предельного ряда равны по абсолютной величине и противоположны по знаку.

Пример 47.5

Случайный ряд многократно подвергается скользящему усреднению с весами

$$\frac{1}{6}[-1, 2, 4, 2, -1].$$

После k итераций производящая функция автоковариаций становится равной

$$6^{-k}\{-z^{-2}+2z^{-1}+4+2z-z^2\}^k.$$

Положив $z = e^{i\alpha}$, находим, что

$$w(\alpha) = c_0(3+2\cos\alpha-2\cos^2\alpha)^k, \qquad (47.43)$$

где c_0 — некоторая постоянная. Функция $w(\alpha)$ принимает максимальное значение при $\cos\alpha = 1/2$ и, следовательно,

$$w(\alpha) = c_0'\left(1-\frac{4}{7}\left(\cos\alpha-\frac{1}{2}\right)^2\right)^k. \qquad (47.44)$$

Таким образом, с ростом k при $\cos\alpha \neq 1/2$ $w(\alpha)$ стремится

*) Символом $\Gamma(z)$ мы обозначили производящую функцию автоковариаций. Если аргумент обозначен другой буквой (не z), символ Γ означает обычную гамма-функцию.

к нулю, в отличие от случая, когда $\cos\alpha = {}^1/_2$. Следовательно, при неограниченном возрастании числа итераций результирующий ряд стремится к периодической «волне» с периодом $2\pi/\arccos\frac{1}{2} = 6$.

Пример 47.6. Теорема Слуцкого о синусоидальном пределе

Вычислим n-кратные скользящие средние по двум элементам случайного ряда, а затем у полученного ряда вычислим разности m-го порядка. Устремим теперь n и m к ∞ так, чтобы m/n стремилось к некоторой постоянной θ, такой, что $0<\theta<1$. Тогда ряд будет стремиться к синусоиде с длиной волны $\lambda = \arccos\frac{1-\theta}{1+\theta}$.

Вычисление m-х разностей эквивалентно m-кратному вычислению первых разностей. Следовательно, производящая функция автоковариаций нового ряда равна

$$\Gamma(z) \propto (1+z)^n(1+z^{-1})^n(1-z)^m(1-z^{-1})^m.$$

Положив $z = e^{i\alpha}$, находим, что

$$w(\alpha) \propto (1-\cos\alpha)^m(1+\cos\alpha)^n. \qquad (47.45)$$

Постоянную можно найти из условия

$$\int_0^\pi w(\alpha)\,d\alpha = \pi.$$

Окончательно получаем

$$w(\alpha) = \frac{\pi\Gamma(m+n+1)}{2^{m+n}\Gamma\left(m+\frac{1}{2}\right)\Gamma\left(n+\frac{1}{2}\right)}(1-\cos\alpha)^m(1+\cos\alpha)^n. \qquad (47.46)$$

Максимальное значение достигается при $\alpha = \alpha_0$, для которого

$$\cos\alpha_0 = \frac{n-m}{n+m} = \frac{1-\theta}{1+\theta}\ ^{*)}. \qquad (47.47)$$

Теорема будет доказана, если мы покажем, что $w(\alpha)\to 0$ при всех α, кроме точки максимума α_0, и что ряд является не только периодическим, но и синусоидальным. Воспользовавшись приближением Стирлинга для Г-функции, получаем

$$w(\alpha) = \frac{(\pi/2)^{1/2}(m+n+1)^{m+n+1/2}(1-\cos\alpha)^m(1+\cos\alpha)^n}{2^{m+n}(m+1/2)^m(n+1/2)^n}.$$

*) Последнее равенство в этом соотношении следует понимать в асимптотическом смысле по $n\to\infty$ и $m\to\infty$, причем $m/n\to\theta$. (*Прим. ред.*)

Если $\cos\alpha = \{(n-m)/(n+m)\} + \varepsilon$, последнее выражение асимптотически эквивалентно выражению

$$\frac{(\pi/2)^{1/2}}{e}(m+n)^{1/2}\left(1-\frac{n+m}{2m}\varepsilon\right)^m\left(1+\frac{n+m}{2n}\varepsilon\right)^n. \qquad (47.48)$$

При $\varepsilon = 0$ это выражение стремится к бесконечности. Логарифмируя и используя разложение для логарифма, нетрудно получить, что (47.48) стремится к нулю равномерно по любому замкнутому отрезку, не содержащему $\varepsilon = 0$. Таким образом, $w(\alpha)$ принимает единственное бесконечное значение при $\alpha_0 = \arccos\frac{1-\theta}{1+\theta}$. Соответственно W — ступенчатая функция с единственным шагом от 0 до π в точке α_0.

Из (47.22) для автокорреляций нового ряда получаем

$$\rho_j = \cos\alpha_0 j. \qquad (47.49)$$

Рассмотрим теперь отрезок полученного ряда $u_1, \ldots u_N$, когда N остается фиксированным при $n\to\infty$. Имеем

$$\mathbf{M}\sum_{i=1}^{N-2}(u_{i+2} - 2\rho_1 u_{i+1} + u_i)^2 = 2(N-2)(1-2\rho_1^2+\rho_2)\,\mathbf{D}u.$$

В силу (47.49) правая часть в пределе равна

$$2(N-2)\,\mathbf{D}u\,\{1-2\cos^2\alpha_0 + \cos 2\alpha_0\} = 0.$$

Следовательно, в пределе

$$u_{i+2} - 2\rho_1 u_{i+1} + u_i = 0. \qquad (47.50)$$

А это и есть разностное уравнение синусоиды.

Некоторые обобщения результата Слуцкого можно найти у Романовского (1932, 1933) и Морэна (1949).

47.16 Если u_t — стационарный ряд, скользящие средние

$$\zeta_t = \beta_0 u_t + \ldots + \beta_h u_{t-h} \qquad (47.51)$$

также стационарны. В частности, если u_t «чисто случайный» процесс, с нулевым средним, автокорреляции нового ряда

$$\rho_j = \begin{cases} \dfrac{\sum_{i=0}^{h-j}\beta_i\beta_{i+j}}{\sum_0^h \beta_i^2}, & j\leqslant h, \\ 0 & j>h. \end{cases} \qquad (47.52)$$

В примере 46.7 мы уже видели, что коррелограмма для таких средних может носить колебательный характер.

47.17 Волд (1938) получил условия, которые следует наложить на постоянные $\rho_1, \ldots, \rho_h$ для того, чтобы они могли быть автокорреляциями скользящих средних случайного ряда.

Рассмотрим производящую функцию

$$G(z) = 1 + \rho_1(z + z^{-1}) + \ldots + \rho_h(z^h + z^{-h}). \qquad (47.53)$$

Положим

$$y = z + z^{-1}. \qquad (47.54)$$

При такой замене $G(z)$ перейдет в полином степени h от y, скажем $H(y)$. Утверждается, что для того, чтобы постоянные ρ_i были автокорреляциями скользящих средних по $(h+1)$ элементам, необходимо и достаточно, чтобы полином $H(y)$ не имел действительных корней нечетной кратности в интервале $-2<y<2$.

Предположим, например, что $\rho_1 \neq 0$, а все остальные ρ равны нулю. Тогда $G(z) = 1 + \rho_1(z + z^{-1})$ и $H(y) = 1 + \rho_1 y$. Последний полином имеет один корень нечетной (единичной) кратности, равный $-1/\rho_1$. Он находится в интервале $(-2,2)$, если $|\rho_1| > 1/2$. Таким образом, нельзя получить скользящие средние по двум элементам, для которых $|\rho_1| > 1/2$, $\rho_2 = \rho_3 = \ldots = 0$, что, впрочем, ясно и из других соображений.

Величины β для скользящих средних определим из соотношения

$$G(z) = \sum_{j=0}^{h} \rho_j (z^j + z^{-j}) = \sum_{j=0}^{h} \beta_j z^j \sum_{j=0}^{h} \beta_j z^{-j}. \qquad (47.55)$$

Вообще говоря, имеется 2^h решений (по числу степеней z имеется h уравнений и каждое из них имеет второй порядок). Однако только при одном из них корни уравнения $\sum \beta_j z^j = 0$ лежат внутри единичного круга. В силу другого результата, который мы рассмотрим позже (47.18), это решение является единственно приемлемым. Из (47.54) следует, что при любом y корни z задаются уравнением

$$z^2 - zy + 1 = 0. \qquad (47.56)$$

Поэтому произведение соответствующих корню y корней уравнения $G(z) = 0$ равно единице и один из корней лежит внутри, а другой вне единичного круга. Из (47.56) получаем, что

$$z = \frac{y}{2} \pm \left(\frac{y^2}{4} - 1\right)^{1/2}. \qquad (47.57)$$

Возможны три случая:

(а) если корень y_1 полинома $H(y)$ является комплексным, то сопряженное число y_1^* также является корнем. Соответствующие числа z_1, z_1^{-1}, z_1^*, $(z_1^*)^{-1}$ являются корнями $G(z)$ и, таким образом, $\sum \beta_j z^j$ делится или на $(z - z_1)(z - z_1^*)$, или на $(z - z_1^{-1})(z - (z_1^{-1})^*)$.

(б) Если действительный корень $H(y)$ по модулю $\geqslant 2$, то из (47.57) следует, что числа z_1 и z_1^{-1} действительны. Следовательно, одно из них является (действительным) корнем уравнения $\sum \beta_j z^j = 0$.

(в) Если действительный корень $H(y)$ по модулю < 2, то z_1 и z_1^{-1} комплексно сопряжены и по модулю равны единице. Тогда выражение $(z - z_1)(z - z_1^{-1})$ является делителем как $\sum \beta_j z^j$, так и $\sum \beta_j z^{-j}$, и, следовательно, корни должны иметь четную кратность.

Отсюда легко следует утверждение теоремы.

Ряды авторегрессии

47.18 Рассмотрим теперь ряды, удовлетворяющие соотношению

$$u_t = -\alpha_1 u_{t-1} - \alpha_2 u_{t-2} \ldots - \alpha_h u_{t-h} + \varepsilon_t, \qquad (47.58)$$

которое (положив $\alpha_0 = 1$) можно записать в более удобной форме:

$$\sum_{j=0}^{h} \alpha_j u_{t-j} = \varepsilon_t. \qquad (47.59)$$

Здесь ε_t — случайная величина, и пока не оговорено противное, будем предполагать, что последовательные величины ε независимы и имеют одну и ту же дисперсию.

Пусть D — такой оператор, что $Du_t = u_{t-1}$. Тогда (47.59) можно переписать в виде

$$\left(\sum \alpha_j D^j\right) u_t = \varepsilon_t.$$

Формальное решение последнего уравнения дается формулой

$$u_t = \frac{1}{\sum \alpha_j D^j} \varepsilon_t = \left(\sum \beta_j D^j\right) \varepsilon_t = \sum_{j=0}^{\infty} \beta_j \varepsilon_{t-j}. \qquad (47.60)$$

Зависимость между постоянными β_j и постоянными α_j определяется из тождества (по D)

$$\frac{1}{\sum \alpha_j D^j} = \sum \beta_j D^j. \qquad (47.61)$$

Однако (47.60) не является полным решением разностного уравнения (47.59). Пусть $z_1, \ldots, z_h$ — корни уравнения

$$z^h + \alpha_1 z^{h-1} + \ldots + \alpha_h = 0. \qquad (47.62)$$

Тогда общее решение уравнения

$$\sum \alpha_j u_{t-j} = 0 \tag{47.63}$$

дается формулой

$$u_t = \sum_{j=1}^{h} A_j z_j^t, \tag{47.64}$$

где A_j — произвольные постоянные.

Мы будем теперь предполагать, что все корни (47.62) различны и $|z_j| < 1$ для всех j, т. е. что все корни (47.62) попадают внутрь единичного круга. Тогда с ростом t решение (47.64) стремится к нулю или, образно говоря, перестает существовать.

Обычно ряд (47.59) рассматривают как ряд, отражающий процесс, начавшийся далеко в прошлом. Тогда вклад решения (47.64) становится пренебрежимо мало, и полное решение фактически совпадает с частным решением (47.60).

Ряды типа (47.58) мы будем называть рядами *авторегрессии* (или авторегрессионными). Они состоят из скользящих средних бесконечного числа элементов. Если ряд ε_t — эргодический, таким же будет и ряд u_t, если $\sum \beta_j^2$ сходится.

Последнее условие в действительности оказывается выполненным, поскольку все корни (47.62) находятся внутри единичного круга (см. упражнение 47.19).

На практике корни уравнения (47.62) ищут редко. Тем не менее Уайз (1956), используя одну теорему Роуза, показал, что условия, налагаемые на эти корни, могут быть выражены в форме алгебраических связей, налагаемых на сами величины α.

47.19 Легко видеть, что u_t зависит от $\varepsilon_t, \varepsilon_{t-1}, \ldots$, но не зависит от $\varepsilon_{t+1}, \varepsilon_{t+2}, \ldots$

Умножим равенство

$$\sum \alpha_j u_{t-j} = \varepsilon_t \tag{47.65}$$

на u_{t-k} и возьмем математические ожидания. Получаем систему уравнений

$$\rho_h + \alpha_1 \rho_{k-1} + \ldots + \alpha_k \rho_{k-h} = 0, \qquad k > 0, \tag{47.66}$$

принадлежащую Юлу (1927) и Г. Уокеру (1931). В частности, поскольку $\rho_{-j} = \rho_j$,

$$\rho_1 + \alpha_1 + \alpha_2 \rho_1 + \ldots \alpha_h \rho_{h-1} = 0,$$
$$\rho_2 + \alpha_1 \rho_1 + \alpha_2 + \ldots \alpha_h \rho_{h-2} = 0$$

и т. д.

Умножая (47.65) на u_{t+k}, $k \geqslant 0$ и беря математические ожидания, получаем

$$\rho_k + \alpha_1 \rho_{k+1} + \ldots + \alpha_h \rho_{k+h} = \frac{D\varepsilon}{Du} \beta_k \tag{47.67}$$

(величины β_j определены в (47.61)).

Эти уравнения получены Волдом (1938).

Пример 47.7. Марковский ряд

Снова рассмотрим ряд примера 47.2. Это простейший случай авторегрессионного ряда (исключая тривиальный случай $h = 0$). Имеем

$$u_t + \alpha_1 u_{t-1} = \varepsilon_t.$$

Последнее удобно записать в виде

$$u_t - \rho u_{t-1} = \varepsilon_t. \tag{47.68}$$

Последний ряд обычно называют марковским.

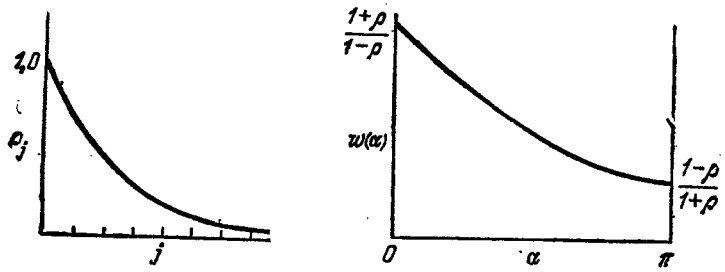

Рис. 47.4. Коррелограмма (слева) и спектр (справа) марковского ряда.

Имеем

$$\frac{1}{1-\rho D} = 1 + \rho D + \rho^2 D^2 + \ldots$$

и, следовательно,

$$\beta_j = \rho^j. \tag{47.69}$$

Отсюда

$$Du = D\left\{\sum_{j=0}^{\infty} \rho^j \varepsilon_{t-j}\right\} = D\varepsilon \sum_{0}^{\infty} \rho^{2j} = \frac{D\varepsilon}{1-\rho^2}. \tag{47.70}$$

Из уравнений (47.66) Юла — Уокера при $h = 1$ получаем

$$\rho_j - \rho_{j-1} \cdot \rho = 0, \qquad j > 0,$$

и, в частности,

$$\rho_1 = \rho \tag{47.71}$$

(вот почему параметр $-\alpha_1$ мы обозначили символом ρ). Далее

$$\rho_k = \rho^k. \tag{47.72}$$

Для спектральной плотности получаем

$$w(\alpha) = \sum_{-\infty}^{\infty} \rho^{|j|} z^j = -1 + \frac{1}{1-\rho z} + \frac{1}{1-\rho z^{-1}} =$$

$$= -1 + \frac{1}{1-\rho e^{i\alpha}} + \frac{1}{1-\rho e^{-i\alpha}} = \frac{1-\rho^2}{1-2\rho\cos\alpha+\rho^2}. \tag{47.73}$$

Коррелограмма и спектр показаны на рис. 47.4.

Пример 47.8. Ряд Юла

Следующий более сложный линейный авторегрессионный ряд носит имя Юла и определяется соотношением

$$u_t + \alpha_1 u_{t-1} + \alpha_2 u_{t-2} = \varepsilon_t. \tag{47.74}$$

Из первых двух уравнений (47.66) Юла — Уокера получаем

$$\rho_1 + \alpha_1 + \alpha_2\rho_1 = 0,$$
$$\rho_2 + \alpha_1\rho_1 + \alpha_2 = 0.$$

Отсюда

$$\rho_1 = -\alpha_1/(1+\alpha_2), \tag{47.75}$$

$$\rho_2 = -\alpha_2 + \alpha_1^2/(1+\alpha_2), \tag{47.76}$$

или

$$\alpha_1 = -\rho_1(1-\rho_2)/(1-\rho_1^2), \tag{47.77}$$

$$\alpha_2 = -(\rho_2-\rho_1^2)/(1-\rho_1^2). \tag{47.78}$$

В этих равенствах параметры α_1 и α_2 представлены в виде функций от первых двух автокорреляций, возможны и обратные соотношения. Вообще, если μ и ν — корни уравнения

$$x^2 + \alpha_1 x + \alpha_2 = 0,$$

то

$$\rho_j = A\mu^j + B\nu^j.$$

Из начальных условий находим, что

$$\rho_0 = 1 = A + B,$$
$$\rho_1 = A\mu + B\nu,$$

и, следовательно,

$$\rho_j = \frac{1}{(\mu-\nu)(1+\mu\nu)}[\mu^{j+1}(1-\nu^2) - \nu^{j+1}(1-\mu^2)]. \tag{47.79}$$

Мы можем придать этому несколько более удобную форму. Положим

$$\mu = pe^{i\theta}, \quad \nu = pe^{-i\theta}. \tag{47.80}$$

Тогда

$$p = |\sqrt{\alpha_2}|, \quad \cos\theta = -\alpha_1/2|\sqrt{\alpha_2}|, \tag{47.81}$$

и (47.79) примет вид

$$\rho_j = \frac{p^j \sin(j\theta+\psi)}{\sin\psi}, \tag{47.82}$$

где

$$\operatorname{tg}\psi = \frac{1+p^2}{1-p^2}\operatorname{tg}\theta. \tag{47.83}$$

Спектральная плотность дается формулой

$$w(\alpha) = \frac{(1-\alpha_2)(1-\alpha_1^2+\alpha_2^2+2\alpha_2)}{(1+\alpha_2)\{1+\alpha_1^2+\alpha_2^2-2\alpha_2+2\alpha_1(1+\alpha_2)\cos\alpha+4\alpha_2\cos^2\alpha\}}. \tag{47.84}$$

На рис. 47.5 показаны типичные коррелограмма и спектр.

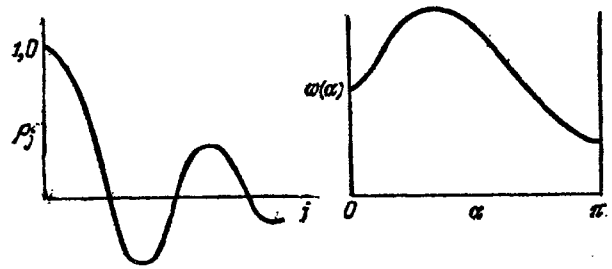

Рис. 47.5. Коррелограмма (слева) и спектр (справа) ряда Юла.

Пример 47.9

Вернемся к вопросу, обсуждавшемуся в конце 47.18. Если корни характеристического уравнения для линейной схемы авторегрессии не лежат внутри единичного круга, процесс не является эргодическим. Если же хотя бы один из корней попадет во внешность такого круга, эволюция членов ряда становится подобной «взрыву». Здесь также возможны колебания, но с резко возрастающей амплитудой.

Если корень лежит на единичной окружности, общее решение не затухает, а содержит гармонический член. Частное решение представляет тогда «блуждающий» ряд. Рассмотрим, например,

простой случай, когда $u_t = u_{t-1} + \varepsilon_t$. Ясно, что частное решение равно

$$u_t = \sum_{j=0}^{\infty} \varepsilon_{t-j}$$

и дисперсия u бесконечна.

Пример 47.10

Рассмотрим схему Юла в предельном случае $\alpha_2 = 1$. Коэффициент p ((47.81)) равен здесь единице и ρ_j в (47.82) «не затухает». Система перестает быть эргодичной.

47.20 Ряды типа ряда

$$u_t = \zeta_t + \varepsilon_t, \tag{47.85}$$

где ζ_t — детерминированный гармонический член, можно рассматривать как гармоники с «наложенными на них ошибками». Иногда такие схемы называют схемами со скрытой периодичностью.

Встречается также несколько иной тип процессов, которые также иногда называют гармониками, хотя здесь этот термин, на наш взгляд, не отражает существа дела. Ряд такого типа может состоять из сумм некоторого числа гармоник, скажем

$$u_t = \sum_j A_j \cos(\alpha_j t), \tag{47.86}$$

где α_j фиксированы, но значения A_j в этом случае могут меняться от одной реализации к другой. В этом случае будут существовать определенные линейные связи типа Юла — Уокера:

$$\sum_{j=0}^{h} c_j \rho_{j-h} = 0. \tag{47.87}$$

Непрерывные ряды

47.21 До сих пор мы касались рядов, которые были определены или наблюдались на некотором дискретном множестве временных точек. Как мы отмечали в главе 45, некоторые ряды «существуют во времени непрерывным образом», и имеются ситуации, в которых мы можем получить непрерывные записи данных; например, график температуры можно получить с помощью

самопишущего прибора, использующего вращающийся валик. То обстоятельство, что в конечном счете явление носит дискретный характер (если это действительно так — см., например, 45.31), не мешает нам рассматривать такие записи как непрерывные. В том случае, когда ряды определены как детерминированные непрерывные функции (например, полиномы или тригонометрические функции), проблема соответствия между предполагаемой непрерывностью реального явления и непрерывностью математической модели редко порождает трудности концептуального характера. Но когда мы переходим к стационарным рядам, где приращение значений между последовательными точками случайны, вопрос о непрерывности должен быть рассмотрен более внимательно. Действительно, существует ли непрерывный ряд, приращения которого случайны, но малы? Нам кажется, что нет, так как понятия случайности и непрерывности в чем-то, по существу, противоположны. Поэтому всякое стремление осуществить привычный с точки зрения математика переход от дискретного к непрерывному случаю, должно тщательно контролироваться. Конечно, может оказаться вполне возможной аппроксимация дискретных выражений непрерывными, например, аппроксимация сумм интегралами, но мы не должны забывать о том, что в этом случае сразу возникает проблема интерпретации.

47.22 Строгое решение этого вопроса невозможно без привлечения теории стохастического интегрирования, которая выходит за рамки этой книги. Тем не менее мы можем дать эвристическое толкование основных результатов примерно в следующем духе.

Рассмотрим непрерывный ряд $u(t)$, определенный на некотором интервале $(-h, h)$. Положив среднее равным нулю, что всерьез не ограничивает общности, мы можем определить дисперсию следующей формулой:

$$\mathrm{D}u = \frac{1}{2h} \int_{-h}^{h} u^2(t)\, dt. \tag{47.88}$$

Если последнее выражение имеет предел при h, стремящемся к бесконечности (что верно для стационарных рядов), дисперсия оказывается определенной и на бесконечном интервале. Аналогично определим автоковариационную функцию. Осуществив нормировку делением на $\mathrm{D}u$, получаем *автокорреляционную функцию*

$$\rho(k) = \lim \frac{1}{2h} \int_{-h}^{h} u(t)\, u(t+k)\, dt. \tag{47.89}$$

Рассмотрим следующее преобразование автокорреляционной функции:

$$\varphi_\rho(p) = \lim \frac{1}{2h} \int_{-h}^{h} \rho(k) e^{ikp}\, dk =$$

$$= \lim \frac{1}{4h^2} \int_{-h}^{h} \int_{-h}^{h} u(t)\, u(t+k)\, dt e^{ikp}\, dk =$$

$$= \lim \frac{1}{4h^2} \int_{-h}^{h} \int_{-h}^{h} u(t)\, e^{-ipt} u(t+k)\, e^{ip(t+k)}\, dt\, dk. \qquad (47.90)$$

Положив $q = t + k$, преобразуем последнее к виду

$$\lim \frac{1}{4h^2} \int_{-h}^{h} \int_{-h}^{h} u(t)\, e^{-ipt} u(q)\, e^{ipq}\, dt\, dq =$$

$$= \lim \frac{1}{4h^2} \int_{-h}^{h} u(t)\, e^{-ipt}\, dt \int_{-h}^{h} u(q)\, e^{ipq}\, dq. \qquad (47.91)$$

Итак, если преобразование ряда $u(t)$ задается формулой

$$\varphi_u(p) = \lim \frac{1}{2h} \int_{-h}^{h} u(t)\, e^{ipt}\, dt = a(p) + ib(p), \qquad (47.92)$$

то, устремляя в (47.91) h к бесконечности для преобразования автокорреляционной функции, получаем

$$\varphi_\rho(p) = a^2(p) + b^2(p) = |\varphi_u(p)|^2. \qquad (47.93)$$

Функция $\varphi_\rho(p)$ есть обобщение на непрерывный случай спектральной плотности (см. (47.21)), которая является преобразованием Фурье последовательности автокорреляций.

47.23 Следует отметить, что даже для непрерывных функций, определенных на бесконечном интервале, автокорреляции не определяют ряд $u(t)$ однозначно. Действительно, при заданной $\varphi_\rho(p)$ из (47.93) получаем

$$\varphi_u(p) = |\varphi_\rho(p)|^{1/2} e^{i\mu}, \qquad (47.94)$$

где μ — произвольная действительная функция. Используя обращение, получаем

$$u(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \varphi_u(p)\, e^{-itp}\, dp = \frac{1}{2\pi} \int_{-\infty}^{\infty} |\varphi_\rho(p)|^{1/2} e^{i\mu - itp}\, dp. \qquad (47.95)$$

Так как значения u действительны, мнимая часть интеграла равна нулю и, следовательно, имеем

$$u(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \sqrt{\varphi_\rho} \cos(\mu - tp)\, dp. \qquad (47.96)$$

(Этот результат принадлежит Винеру (1930).) Следовательно, μ должна быть нечетной функцией от p, но во всем остальном произвольной (исключая, конечно, условие сходимости). Поэтому φ_ρ не определяет u однозначно. Позже мы рассмотрим этот вопрос с точки зрения спектральных функций.

Пример 47.11

Рассмотрим автокорреляционную функцию, которая уже появлялась у нас в связи с рядами Юла в (47.82):

$$\rho(k) = p^k \sin(k\theta + \psi)/\sin\psi.$$

Посмотрим, что произойдет, если мы будем рассматривать ее для всех k, а не только для целых значений. Положим, слегка меняя обозначения,

$$\rho(k) = \frac{e^{-qh} \sin(k\theta + \psi)}{\sin\psi}, \qquad q > 0. \qquad (47.97)$$

При отрицательном k в последнем выражении следует употреблять $|k|$ вместо k. Применяя преобразование, получаем

$$\varphi_\rho(p) = \int_{-\infty}^{\infty} \frac{e^{-|qk|} \sin(k\theta + \psi)}{\sin\psi} e^{ikp}\, dk = \frac{q}{q^2 + (\theta + p)^2} + \frac{q}{q^2 + (\theta - p)^2}. \qquad (47.98)$$

Величину p в этом преобразовании не следует путать с коэффициентом затухания в $\rho(k)$.

Отметим, что спектральная функция (47.98) непрерывна и достигает максимума при $p = 0$. Физик не устоял бы перед искушением рассматривать этот спектр как сходный со спектром белого света, поскольку в обоих представлены все частоты. Однако из последнего обстоятельства отнюдь не следует, что ряд $u(t)$ возникает как сумма гармонических членов со всеми возможными частотами, аналогично белому свету, который можно рассматривать как результат суммарного действия некоторого числа резонаторов, порождающих колебания со всеми возможными длинами волн.

В связи с этим вопросом о белом свете, рассмотрим предельный случай, когда временной ряд определяется как ряд величин, соответствующих небольшим временным интервалам длины Δt.

Предположим, что все автокорреляции равны нулю. Тогда из (47.21) находим, что спектральная плотность равна единице (если в качестве единицы масштаба взять единичный временной интервал), а с единицей масштаба Δt была бы равна приблизительно Δt, т. е. постоянной. Соответствующие физические системы характеризуются постоянной спектральной плотностью и порождают ряды, представляющие собой сумму колебаний, средние амплитуды которых равны, а соседние частоты весьма близки. В теории связи такой случай называют *белым шумом*. При передаче сигнала определенной частоты такой шум служит помехой (как и в радиоприемнике), отражающейся на чистоте приема и вызывающей возмущения типа ошибок, накладывающихся на входной сигнал.

Фильтры и передаточные функции

47.24 Допустим, что нам дан ряд $u(t)$ и совокупность весов $a(t)$. Образуем новый ряд, являющийся линейным временным средним первоначального, по формуле

$$v(t)=\int_0^\infty a(\tau)\,u(t-\tau)\,d\tau. \qquad (47.99)$$

Усреднение проводится по прошлым и настоящему значениям $u(t)$ и не касается будущих значений. Если $u(t)$ определено на дискретном множестве точек, аналогично определяется соответствующая сумма. Чтобы получить спектр, рассмотрим

$$\int_{-\infty}^{\infty} v(t)\,e^{it\alpha}\,dt=\int_0^\infty a(\tau)\int_{-\infty}^{\infty} e^{it\alpha}u(t-\tau)\,dt\,d\tau=$$

$$=\int_0^\infty a(\tau)\,e^{i\tau\alpha}\int_{-\infty}^{\infty} u(t)\,e^{it\alpha}\,dt\,d\tau=\int_0^\infty a(\tau)\,e^{i\tau\alpha}\,d\tau\int_{-\infty}^{\infty} u(t)\,e^{it\alpha}\,dt\,.$$

Вычислив квадраты модулей, получаем

$$|\varphi_v(\alpha)|^2=|\varphi_u(\alpha)|^2\left|\int_0^\infty e^{i\tau\alpha}a(\tau)\,d\tau\right|^2. \qquad (47.100)$$

Функцию

$$T(\alpha)=\int_0^\infty e^{i\tau\alpha}a(\tau)\,d\tau \qquad (47.101)$$

иногда называют частотной характеристикой или *передаточной функцией*. В сущности это х. ф. (преобразование Фурье) весовой функции $a(\tau)$. Равенства (47.93) и (47.100) показывают, что спектральную плотность нового ряда можно получить, умножая спектральную плотность первоначального на модуль передаточной функции. С инженерной точки зрения система представляет собой «входящий» ряд $u(t)$, преобразующее $u(t)$ устройство, работа которого состоит в линейном усреднении, и «выход» $v(t)$.

47.25 С учетом некоторых ограничений передаточную функцию можно выбирать так, чтобы у полученного из $u(t)$ выходного ряда $v(t)$ некоторые частоты были резко выражены. Такую функцию, или такую совокупность весов $a(\tau)$, называют иногда *фильтром*. Вряд ли это самое удачное название, поскольку фильтр скорее удаляет примеси, а не преобразует их, но оно вошло в употребление. Нам нет необходимости ограничиваться использованием лишь средних по прошлому, и мы не будем этого делать. Поэтому обычное скользящее усреднение, которое мы рассматривали в главе 46, также является фильтром в нашем смысле.

Частные автокорреляции

47.26 Запишем линейную схему авторегрессии в виде (47.58)

$$u_t=-\alpha_1u_{t-1}-\alpha_2u_{t-2}-\ldots-\alpha_hu_{t-h}+\varepsilon_t. \qquad (47.102)$$

Можно сказать, что (47.102) позволяет прогнозировать u_t и состоит из членов двух типов: систематических членов, содержащих u_{t-j} и отражающих зависимость u_t от прошлой истории, и случайного элемента ε_t, который можно интерпретировать как возмущение. Тогда можно задать следующий вопрос, который обычно и ставят при регрессионном анализе. Даны автокорреляции ρ_j. Чему тогда равны частные корреляции, отражающие зависимость u_t от каких-то предыдущих членов в том случае, когда влияние на u_t промежуточных членов устранено?

47.27 Прежде всего рассмотрим схему Маркова (47.68),

$$u_t=\rho u_{t-1}+\varepsilon_t. \qquad (47.103)$$

Мы знаем (см. (47.72)), что $\rho_k=\rho^k$. Выпишем частную автокорреляцию $\rho_{13,2}$, отражающую зависимость u_t от u_{t-2} при исключенном «посредничестве» u_{t-1}. В обычных обозначениях имеем (см. (27.5))

$$\rho_{13,2}=\frac{\rho_{13}-\rho_{12}\rho_{32}}{(1-\rho_{12}^2)^{1/2}(1-\rho_{32}^2)^{1/2}},$$

где $\rho_{13}=\rho^2$, $\rho_{12}=\rho_{32}=\rho$. Следовательно,

$$\rho_{13,2}=0. \qquad (47.104)$$

Легко видеть, что в действительности все частные автокорреляции равны нулю. Это очевидно еще и потому, что в (47.103)

регрессия u_t касается лишь u_{t-1}, а ε_t — независимые величины. Итак, в марковской схеме все частные автокорреляции равны нулю. Член u_t зависит лишь от предыдущего члена, и поэтому распределение u_t полностью определяется значением u_{t-1}.

47.28 Рассмотрим теперь схему Юла (47.74)

$$u_t = -\alpha_1 u_{t-1} - \alpha_2 u_{t-2} + \varepsilon_t. \qquad (47.105)$$

Из (47.75)—(47.76) следует, что

$$\rho_1 = -\frac{\alpha_1}{1+\alpha_2}, \qquad \rho_2 = -\alpha_2 + \frac{\alpha_1^2}{1+\alpha_2}.$$

Частная корреляция между u_t и u_{t-2} равна

$$\rho_{13.2} = \frac{\rho_2 - \rho_1^2}{1-\rho_1^2} = -\alpha_2, \qquad (47.106)$$

что и следовало ожидать.

Легко убедиться, что частные корреляции более высоких порядков равны нулю. Например, числитель в выражении для $\rho_{14,23}$ равен $\rho_{14,3} - \rho_{12,3}\rho_{42,3}$. Последнее выражение в свою очередь можно представить в виде дроби, числитель которой равен

$$(1-\rho_1^2)(\rho_3 - \rho_1\rho_2) - (\rho_1 - \rho_1\rho_2)(\rho_2 - \rho_1^2). \qquad (47.107)$$

Рассмотрев определитель, образованный первыми тремя коэффициентами первых трех уравнений (47.66) Юла — Уокера, имеем

$$\begin{vmatrix} \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \\ \rho_3 & \rho_2 & \rho_1 \end{vmatrix} = 0.$$

Разложив его по первому столбцу, получаем, что выражение (47.107) равно нулю.

С точки зрения общей теории регрессии этот результат становится очевидным, если вспомнить, что все, как независимые, так и зависимые, переменные в уравнениях регрессии имеют одну и ту же дисперсию, так что нормированные коэффициенты в этом уравнении равны частным корреляциям.

Бесконечные, полубесконечные и циклические процессы

47.29 Если мы рассмотрим схему линейной авторегрессии

$$\sum_{j=0}^{h} \alpha_j u_{t-j} = \varepsilon_t, \qquad (47.108)$$

как порождающую ряд значений u по заданным значениям ε, мы столкнемся с трудностью, заключающейся в необходимости принятия решения. Мы не можем найти значения u, скажем, в

точке T, не зная значений в точках $T-1, \ldots, T-h$. Мы можем, однако, предположить, что такие значения заданы или известны для некоторой точки T_0. «Двигаясь вперед», мы можем восстановить значения в точках, следующих за T_0. В этом случае наш ряд можно назвать полубесконечным, поскольку мы его рассматриваем как простирающийся до бесконечности в одном направлении, т. е. в направлении возрастания t.

С другой стороны, можно считать, что ряд простирается как в бесконечное будущее, так и в бесконечное прошлое. В этом случае «стартовой точкой» может быть любая, и такой ряд естественно назвать просто бесконечным.

47.30 Исходя из определенных математических соображений, которые станут ясными позднее, мы можем определить циклический *процесс* аналогично тому, как мы определяли циклические коэффициенты корреляции в **45.34**. Конкретно, предположим, что для некоторого N

$$u_{t+N} = u_t,$$
$$u_{t+N+1} = u_{t+1},$$
$$\cdots\cdots$$
$$u_{t+N+h} = u_{t+h}.$$

Пожалуй, трудно вообразить физический процесс, которому соответствует такая схема. Поэтому, насколько это возможно, мы будем ее избегать. Лучшее, что можно сказать об этой схеме, это то, что при возрастании N она, возможно, будет служить хорошим приближением к полубесконечному случаю (конечно, при условии наличия точных результатов для конечного N). Но даже последнее сомнительно, так как вряд ли удастся избежать трудностей при устремлении N к бесконечности. Поэтому нам следует быть очень осторожными при интерпретации результатов, касающихся циклических процессов.

47.31 Теоретические коррелограммы на рис. 47.4 и 47.5 не слишком похожи на выборочные коррелограммы при небольшом количестве наблюдений (под «небольшим» мы здесь понимаем ряд приблизительно из 100 членов). В упражнениях 47.20—22 приведены корреляции, вычисленные по наблюдениям из таблиц 47.1, 47.2 и 47.4. Кроме нерегулярностей выборочного характера, на практике можно столкнуться еще с двумя явлениями: (а) сериальные корреляции оказываются отрицательно смещенными; (б) в схемах Юла и Маркова корреляции высоких порядков «затухают» не столь быстро, как это можно было ожидать.

Теоретическое объяснение этим явлениям будет дано в следующей главе. Вопрос о том, как «подгонять» различные схемы под стационарные ряды и проверять соответствующие гипотезы, мы рассмотрим в главе 50.

УПРАЖНЕНИЯ

47.1 Пусть у стационарного ряда $\rho_1 = \rho_2 = \rho$. Показать, что $\rho \geqslant -1/2$ и

$$\rho_3 \geqslant \frac{4\rho^2 - \rho - 1}{\rho + 1}.$$

47.2 Показать, что семиинварианты марковского ряда

$$u_t = \rho u_{t-1} + \varepsilon_t$$

связаны с семиинвариантами ε соотношением $\varkappa_r(u) = \frac{\varkappa_r(\varepsilon)}{1-\rho^r}$. Выведите отсюда, что если нормированный коэффициент $\lambda_r = \varkappa_r/\varkappa_2^{r/2}$, то

$$\lambda_r(u) = \lambda_r(\varepsilon)\frac{(1-\rho^2)^{r/2}}{1-\rho^r}.$$

Показать, что в общем случае распределение u ближе к нормальному, чем распределение ε, но будет в точности нормальным лишь при условии нормальности последнего распределения.

47.3 Показать, что марковскую схему предыдущего упражнения можно описывать и формулой

$$u_t = \rho u_{t+1} + \eta_t,$$

где

$$\eta_t = -\rho\varepsilon_{t+1} + (1-\rho^2)\sum_{j=0}^{\infty}\rho^j\varepsilon_{t-j}.$$

Вывести отсюда, что если η — случайная величина, то

$$\mathsf{D}\eta = \mathsf{D}\varepsilon,$$

$$\operatorname{cov}(\eta_t, \eta_{t+k}) = 0, \quad k \neq 0.$$

47.4 Убедиться, что для спектральной плотности (47.33)

$$\int_0^{\pi} w(\alpha)\,d\alpha = \pi.$$

47.5 Проверить формулу (47.84) для схемы авторегрессии Юла. Убедиться, что и здесь интеграл от $w(\alpha)$ от 0 до π равен π.

47.6 Показать, что существуют четыре и только четыре скользящих средних случайного ряда с коррелограммой

$$\rho_1 = -\frac{42}{85}, \quad \rho_2 = \frac{4}{17}, \quad \rho_3 = -\frac{8}{85}, \quad \rho_4 = \rho_5 = \ldots = 0.$$

Эти значения таковы:

$$\frac{1}{5}[8, -4, 2, -1], \ \frac{1}{5}[-1, 2, -4, 8], \ \frac{1}{5}[2, -1, 8, -4], \ \frac{1}{5}[-4, 8, -1, 2].$$

(Волд, 1938.)

47.7 Показать, что в общей схеме авторегрессии со случайным членом ε

$$\mathsf{D}\left(\sum_{j=0}^{h} a_j u_{t+j}\right) = \mathsf{D}\varepsilon.$$

47.8 Показать, что в схеме авторегрессии Юла

$$\frac{\mathsf{D}u}{\mathsf{D}\varepsilon} = \frac{1+\alpha_2}{(1-\alpha_2)\{(1+\alpha_2)^2 - \alpha_1^2\}}.$$

47.9 Показать, что автокорреляции m-х разностей случайного ряда даются формулой

$$\rho_j = (-1)^j \frac{m^{(j)}}{(m+j)^{(j)}}.$$

47.10 Показать, что если какой-либо ряд «подгоняется» под схему Юла с автокорреляциями ρ_j, автокорреляции остатков равны

$$\sigma_j = \frac{(1+\alpha_1^2+\alpha_2^2)\rho_j + \alpha_1(1+\alpha_2)(\rho_{j+1}+\rho_{j-1}) + \alpha_2(\rho_{j+2}+\rho_{j-2})}{1+\alpha_1^2+\alpha_2^2+2\alpha_1(1+\alpha_2)\rho_1 + 2\alpha_2\rho_2}.$$

47.11 Показать, что ряд, у которого автокорреляционная функция равна

$$v(j) = (\sin\lambda j)/\lambda j,$$

имеет спектр с единственным скачком в точке λ.

47.12 Показать, что любой ряд линейной авторегрессии можно представить в виде комбинированной последовательности рядов Юла и Маркова, в которой «ошибка» ε у одного ряда будет значением другого.

47.13 Пусть

$$A = \frac{2}{n}\sum_{j=1}^{n} u_j \cos\beta_j,$$

$$B = \frac{2}{n}\sum_{j=1}^{n} u_j \sin\beta_j.$$

Показать, что если $u_t = a\sin\alpha t + b_t$, где b_t — компонента, не коррелированная с $\sin\alpha t$, то

$$A = \frac{a}{n}\left[\frac{\sin\left\{\frac{1}{2}n(\alpha-\beta)\right\}\sin\left\{\frac{1}{2}(n+1)(\alpha-\beta)\right\}}{\sin\left\{\frac{1}{2}(\alpha-\beta)\right\}} + \frac{\sin\left\{\frac{1}{2}n(\alpha+\beta)\right\}\sin\left\{\frac{1}{2}(n+1)(\alpha+\beta)\right\}}{\sin\left\{\frac{1}{2}(\alpha+\beta)\right\}}\right],$$

и аналогичное выражение верно и для B. Показать отсюда, что при возрастании n $S^2 = A^2 + B^2$ остается малым, если только не мало $\alpha - \beta$. В последнем случае $S^2 = a^2$.

47.14 Показать, что у «непрерывного» ряда с автоковариационной функцией

$$\rho(k) = e^{-a|k|}$$

спектральная плотность равна

$$w(\alpha) = \frac{1}{\pi(a^2+\alpha^2)}.$$

(Ср. с характеристической функцией распределения Коши.)

47.15 Пусть ряд u_t таков, что

$$u_t = u_{t-1} + \varepsilon_t,$$

где ε_t — случайный ряд с единичной дисперсией. Ряд разбивается на последовательные группы из m членов, пусть v_t — арифметическое среднее t-й группы. Показать, что

$$\mathrm{D}(\Delta v_t) = (2m^2 + 1)/3m,$$

и

$$\operatorname{cov}(\Delta v_t, \Delta v_{t+1}) = \frac{m^2 - 1}{2(2m^2 + 1)}.$$

[(Уоркинг, 1960.)

47.16 Пусть U есть $(N \times N)$-матрица

$$\begin{pmatrix} 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & 0 & \dots & 0 \end{pmatrix}.$$

Показать, что последовательные степени U получаются смещением состоящей из единиц диагонали вправо и что $U^N = 0$. Получить отсюда, что если u — вектор-столбец $(u_t, u_{t-1}, \dots, u_{t-N})$, то схему авторегрессии можно записать в виде

$$\left(\sum_{j=0}^{h} \alpha_j U^j\right) u = \varepsilon.$$

Показать далее, что матрица рассеяния u равна

$$V(u) = \left(\sum \alpha_j U^j\right)^{-1} \left(\sum \alpha_j U'^j\right)^{-1}.$$

(Уиттл, 1951.)

47.17 Показать, что матрица, обратная к матрице рассеяния u, равна

$$V^{-1} = \left(\sum \alpha_j U'^j\right)\left(\sum \alpha_j U^j\right).$$

Получите отсюда, что для рядов линейной авторегрессии третьего порядка со случайными ошибками матрица, обратная к автоковариационной матрице, равна

$$V^{-1} = \begin{pmatrix} 1 & \alpha_1 & \alpha_2 & \alpha_3 \dots \\ \alpha_1 & 1 + \alpha_1^2 & \alpha_1(1 + \alpha_2) & \alpha_2 + \alpha_1\alpha_3 \dots \\ \alpha_2 & \alpha_1(1 + \alpha_2) & 1 + \alpha_1^2 + \alpha_2^2 & \alpha_1 + \alpha_1\alpha_2 + \alpha_2\alpha_3 \dots \\ \alpha_3 & \alpha_2 + \alpha_1\alpha_3 & \alpha_1 + \alpha_1\alpha_2 + \alpha_2\alpha_3 & 1 + \alpha_1^2 + \alpha_2^2 + \alpha_3^2 \dots \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{pmatrix}.$$

(Уайз, 1955.)

47.18 Показать, что хотя автокорреляционная матрица ряда является матрицей Лорана, обратная к ней таковой не является.

(Уиттл, 1951.)

47.19 Раскладывая на элементарные дроби левую часть уравнения (47.65), показать, что $\sum \beta_j^2$ сходится (см. 47.18), если корни (47.62) различны и лежат внутри единичного круга.

47.20 Ниже приведены сериальные корреляции для данных таблицы 47.1 (цены на пшеницу). Построить коррелограмму:

Порядок корреляции k	r_k	Порядок корреляции k	r_k	Порядок корреляции k	r_k	Порядок корреляции k	r_k
1	0,562	16	0,158	31	0,060	46	—0,036
2	0,103	17	0,109	32	—0,008	47	—0,013
3	—0,075	18	0,002	33	—0,039	48	0,042
4	—0,092	19	—0,075	34	0,007	49	0,062
5	—0,082	20	—0,062	35	0,056	50	0,065
6	—0,136	21	—0,021	36	0,010	51	0,050
7	—0,211	22	—0,062	37	—0,004	52	0,009
8	—0,261	23	—0,088	38	—0,015	53	—0,027
9	—0,192	24	—0,084	39	—0,047	54	—0,053
10	—0,070	25	—0,076	40	—0,047	55	—0,073
11	—0,003	26	—0,091	41	0,008	56	—0,106
12	—0,015	27	—0,052	42	0,034	57	—0,084
13	—0,012	28	—0,032	43	0,065	58	—0,019
14	0,047	29	—0,012	44	0,099	59	0,003
15	0,101	30	—0,059	45	0,009	60	0,010

47.21 Ниже приведены сериальные корреляции для данных таблицы 47.2 (доли браков). Построить коррелограмму:

Порядок корреляции k	r_k	Порядок корреляции k	r_k	Порядок корреляции k	r_k	Порядок корреляции k	r_k
1	0,563	6	—0,025	11	—0,080	16	—0,126
2	—0,089	7	0,353	12	—0,136	17	—0,036
3	—0,498	8	0,396	13	—0,132	18	0,131
4	—0,631	9	0,254	14	—0,058	19	0,209
5	—0,467	10	0,104	15	—0,095	20	0,205

47.22 Ниже приведены сериальные корреляции для данных таблицы 47.4 (искусственный ряд). Построить коррелограмму:

Порядок корреляций k	r_k	Порядок корреляций k	r_k	Порядок корреляций k	r_k	Порядок корреляций k	r_k
1	0,70	9	0,12	17	0,12	24	—0,43
2	0,29	10	0,05	18	0,29	25	—0,57
3	0,01	11	—0,05	19	0,33	26	—0,56
4	—0,17	12	—0,17	20	0,22	27	—0,26
5	—0,27	13	—0,27	21	0,05	28	0,02
6	—0,25	14	—0,31	22	—0,12	29	0,17
7	—0,13	15	—0,30	23	—0,28	30	0,27
8	0,07	16	—0,18				

ГЛАВА 48

ВЫБОРОЧНАЯ ТЕОРИЯ СЕРИАЛЬНЫХ КОРРЕЛЯЦИЙ

Теория для больших выборок

48.1 В **45.32** мы определили сериальную корреляцию с лагом (или запаздыванием) k и отметили, что для определенных целей с математической и вычислительной точек зрения удобнее основываться на более простых определениях. При больших n все имеющиеся в виду определения асимптотически эквивалентны. Поэтому в теории больших выборок мы будем рассматривать стандартную ошибку формы

$$r_j = \frac{1}{n}\sum_{i=1}^{n-j} u_i u_{i+j} \Big/ \frac{1}{n}\sum_{i=1}^{n} u_i^2 = c/v. \qquad (48.1)$$

Как обычно, в окончательных выражениях безразлично использовать ли теоретические или выборочные формы. Для автокорреляций мы, как правило, будем использовать символы ρ_j.

Применяя обычную процедуру, имеем

$$\delta r_j = \frac{\delta c}{v} - \frac{c\,\delta v}{v^2},$$

$$\mathsf{D} r_j = \frac{\mathsf{D}c}{v^2} - \frac{2c\,\mathrm{cov}\,(c,\,v)}{v^3} + \frac{c^2\,\mathsf{D}v}{v^4}.$$

Положив, без потери общности, $\mathsf{M}v = 1$, получим *)

$$\mathsf{D} r_j = \mathsf{D}c - 2c\,\mathrm{cov}\,(c,\,v) + c^2\,\mathsf{D}v. \qquad (48.2)$$

Для получения развернутого представления этого выражения нам понадобится один общий результат, касающийся ковариации двух ковариационных членов. Имеем

$$\mathrm{cov}\left\{\frac{1}{n}\sum u_a u_{a+s},\ \frac{1}{n}\sum u_b u_{b+s+t}\right\} = \frac{1}{n^2}\mathsf{M}\left\{\sum u_a u_{a+s}\sum u_b u_{b+s+t}\right\} - \rho_s\rho_{s+t} = \frac{1}{n^2}\mathsf{M}\left\{\sum_{a,\,b}(u_a u_{a+s} u_b u_{b+s+t})\right\} - \rho_s\rho_{s+t}. \qquad (48.3)$$

*) Оговорив в подзаголовке, что речь идет о *больших* выборках, автор при выводе и записи ряда соотношений (см. (48.2), (48.5)—(48.9) и др.) пренебрегает членами порядка $o(1/n)$, а также зачастую не различает выборочных и соответствующих теоетических характеристик распределения. (*Прим. ред.*)

Чтобы иметь возможность вычислить в этом выражении смешанные моменты четвертого порядка, сделаем еще несколько предположений. Предположим, что все u имеют совместное нормальное распределение, так что их характеристическая функция имеет вид

$$\exp\left\{-\frac{1}{2}(\theta_a^2 + \theta_{a+s}^2 + \theta_b^2 + \theta_{b+s+t}^2 + 2\rho_s\theta_a\theta_{a+s} + \ldots)\right\}. \qquad (48.4)$$

Для $\mathsf{M}(u_a u_{a+s} u_b u_{b+s+t})$ получаем выражение

$$\rho_s\rho_{s+t} + \rho_{b-a}\rho_{b-a+t} + \rho_{b-a+s+t}\rho_{b-a-s}.$$

Суммируя по a и b, получаем, что ковариация (48.3) равна

$$\frac{1}{n^2}\left[n^2\rho_s\rho_{s+t} + n\sum_{i=-\infty}^{\infty}\rho_i\rho_{i+t} + n\sum_{i=-\infty}^{\infty}\rho_{i+s+t}\rho_{i-s}\right] - \rho_s\rho_{s+t} = \frac{1}{n}\left\{\sum\rho_i\rho_{i+t} + \sum\rho_{i+s+t}\rho_{i-s}\right\}. \qquad (48.5)$$

Перейдем к конкретным формулам. Положив $s = t = 0$, получаем

$$\mathsf{D}v = \frac{2}{n}\sum_{-\infty}^{\infty}\rho_i^2. \qquad (48.6)$$

Положив $t = 0$, получаем

$$\mathsf{D}c = \frac{1}{n}\sum_{-\infty}^{\infty}(\rho_i^2 + \rho_{i+s}\rho_{i-s}). \qquad (48.7)$$

Положив $s = 0$ и заменяя t на s, получаем

$$\mathrm{cov}\,(v,\,c) = \frac{2}{n}\sum_{-\infty}^{\infty}\rho_i\rho_{i+s}. \qquad (48.8)$$

Наконец, подставляя последние выражения в (48.2), получаем

$$\mathsf{D}r_j = \frac{1}{n}\sum_{i=-\infty}^{\infty}[\rho_i^2 + \rho_{i-j}\rho_{i+j} - 4\rho_j\rho_i\rho_{i+j} + 2\rho_i^2\rho_j^2]. \qquad (48.9)$$

Эта формула получена Бартлеттом (1946).

48.2 Последний результат говорит о том, что даже при больших выборках и упрощающем дело предположении о нормальности дисперсия r_j зависит от автокорреляций ряда. Это очень неудобно, поскольку мы не можем оценить последние, основываясь на конечных рядах. Тем не менее можно получить достаточно хорошие приближения, если действовать в духе следующих примеров.

Пример 48.1

Рассмотрим сначала тот простой случай, когда все теоретические автокорреляции равны нулю (случайный ряд). Из (48.9) получаем

$$\mathsf{D}r_j = \frac{1}{n}. \tag{48.10}$$

Последнее подтверждает, что дисперсия имеет порядок $1/n$. В сущности, в этом случае выборочная формула сводится к выборочной формуле для обычного коэффициента корреляции двумерной нормальной выборки. Последнее очевидным образом следует из того, что ряд случаен.

Пример 48.2

Если ρ_j и все последующие ρ малы, (48.9) можно заменить приближенной формулой

$$\mathsf{D}r_j = \frac{1}{n}\sum_{-(j-1)}^{j-1} \rho_i^2. \tag{48.11}$$

Можно проверить (мы оставляем это до упражнения 48.1), что при некоторых простых предположениях

$$\operatorname{cov}(r_j, r_{j+t}) = \frac{1}{n}\sum_{i=-(j-1)}^{(j-1)} \rho_i \rho_{i+t}. \tag{48.12}$$

Пример 48.3

Если ряд подчиняется уравнению Маркова (47.68) с параметром ρ, из (48.11) и (47.72) для больших j получаем

$$\mathsf{D}r_j = \frac{1}{n}\left\{\sum_{-\infty}^{\infty} \rho^{|2i|}\right\} = \frac{1}{n}\,\frac{1+\rho^2}{1-\rho^2}. \tag{48.13}$$

Более точное выражение для малых j приведено в упражнении 48.8.

Пример 48.4

Приближения (48.11) и (48.12) можно получить непосредственно из производящей функции автокорреляций.

Действительно, в обозначениях **47.14**

$$w(\lambda) = G(z) = \sum_{-\infty}^{\infty} \rho_i z^i,$$

и, следовательно,

$$\{G(z)\}^2 = \sum_{\alpha=-\infty}^{\infty} z^\alpha \sum_i \rho_i \rho_{\alpha-i} = \sum_\alpha z^\alpha \Big(\sum_i \rho_i \rho_{\alpha+i}\Big), \tag{48.14}$$

так что G^2 — производящая функция интересующей нас последовательности сумм.

Например, для марковского процесса

$$G(z) = -1 + \frac{1}{1-\rho z} + \frac{1}{1-\rho z^{-1}},$$

$$G^2(z) = 1 + (1-\rho z)^{-2} + (1-\rho z^{-1})^{-2} - 2(1-\rho z)^{-1} - \\ - 2(1-\rho z^{-1})^{-1} + 2(1-\rho z)(1-\rho z^{-1}).$$

Отсюда получаем, что коэффициент при z^0 равен

$$1 + 1 + 1 - 2 - 2 + 2(1 + \rho^2 + \rho^4 + \ldots) = \frac{1+\rho^2}{1-\rho^2},$$

что совпадает с (48.13). Коэффициент при z^k равен

$$(k+1)\rho^k - 2\rho^k + 2(\rho^k + \rho^{k+2} + \ldots) = \rho^k\left\{k - 1 + \frac{2}{1-\rho^2}\right\}.$$

Следовательно, ковариация между r_j и r_{j+k} в схеме Маркова в силу (48.12) равна

$$\frac{1}{n}\rho^k\left\{k - 1 + \frac{2}{1-\rho^2}\right\}. \tag{48.15}$$

Используя (48.13), получаем, что соответствующая корреляция равна

$$\frac{\rho^k\{(k+1) - (k-1)\rho^2\}}{1+\rho^2}. \tag{48.16}$$

Метод может быть распространен на схемы более высоких порядков (см. Кенуй (1947а) и упражнение 48.14).

Смещение в оценках автокорреляций

48.3 Если r имеет форму $A/\sqrt{BC}$, то, обозначив a, b, c отклонения от средних, имеем

$$r_j = \frac{\mathsf{M}A + a}{\{[\mathsf{M}B + b][\mathsf{M}C + c]\}^{1/2}}.$$

Используя биномиальное разложение с точностью до второго порядка, находим

$$\mathsf{M}r_j = \frac{\mathsf{M}A}{\{\mathsf{M}B\mathsf{M}C\}^{1/2}}\left\{1 - \frac{\mathsf{M}(ab)}{2\mathsf{M}A\mathsf{M}B} - \frac{\mathsf{M}(ac)}{2\mathsf{M}A\mathsf{M}C} + \frac{\mathsf{M}(bc)}{4\mathsf{M}B\mathsf{M}C} + \right. \\ \left. + \frac{3\mathsf{M}b^2}{8\mathsf{M}^2B} + \frac{3\mathsf{M}c^2}{8\mathsf{M}^2C} + \ldots\right\}. \tag{48.17}$$

Положим

$$B=\frac{1}{n-j}\sum_{i=1}^{n-j}u_i^2-\frac{1}{(n-j)^2}\left(\sum_{i=1}^{n-j}u_i\right)^2, \qquad (48.18)$$

$$C=\frac{1}{n-j}\sum_{i=1}^{n-j}u_{i+j}^2-\frac{1}{(n-j)^2}\left(\sum_{i=1}^{n-j}u_{i+j}\right)^2, \qquad (48.19)$$

так что $\mathsf{M}B=\mathsf{M}C$. Отсюда после небольших упрощений, использующих асимптотическую эквивалентность B и C, получаем

$$\mathsf{M}r_j \doteqdot \frac{\mathsf{M}A}{\mathsf{M}B}-\frac{\operatorname{cov}(a,b)}{\mathsf{M}^2B}+\frac{\mathsf{M}A\,\mathsf{D}b}{\mathsf{M}^3B}. \qquad (48.20)$$

Пусть дисперсия ряда равна единице. Положим $\nu=n-j$. Тогда, взяв математические ожидания (48.18—19), получаем

$$\mathsf{M}B=\mathsf{M}C=\frac{1}{\nu}\left\{\nu-1-\frac{2}{\nu}\sum_{i=1}^{\nu-1}(\nu-i)\rho_i\right\}. \qquad (48.21)$$

Пусть теперь

$$A=\frac{1}{n-j}\sum_{i=1}^{n-j}u_iu_{i+j}-\frac{1}{(n-j)^2}\sum_{i=1}^{n-j}u_i\sum_{i=1}^{n-j}u_{i+j}. \qquad (48.22)$$

Имеем

$$\mathsf{M}A=\frac{1}{\nu}\left\{\nu\rho_j-\frac{1}{\nu}\sum_{i=0}^{\nu-1}(\nu-i)\rho_{j+i}-\frac{1}{\nu}\sum_{i=0}^{j}(\nu-i)\rho_{j-i}-\right.$$
$$\left.-\frac{1}{\nu}\sum_{i=1}^{\nu-j-1}(\nu-j-i)\rho_i\right\}, \qquad j<0. \qquad (48.23)$$

$\mathsf{D}b$ и $\operatorname{cov}(a,b)$ мы уже вычислили в (48.6) и (48.8). Подставляя все полученные величины в (48.20), получаем $\mathsf{M}r_j$. Не выписывая эту громоздкую формулу, рассмотрим несколько частных случаев.

Пример 48.5

Если ряд случаен, то $\rho_j=0$, за исключением $\rho_0=1$. Из (48.20) получаем, что

$$\mathsf{M}r_j \doteqdot -\frac{1}{\nu}. \qquad (48.24)$$

Оказывается, что в этом случае мы можем точно вычислить смещение, используя определение

$$r_j=\frac{n}{n-j}\,\frac{\sum_1^{n-j}(u_i-\bar u)(u_{i+j}-\bar u)}{\sum_1^{n}(u_i-\bar u)^2}.$$

Положим

$$z_i=u_i-\bar u.$$

Тогда

$$\mathsf{M}r_j=\frac{n}{n-j}\mathsf{M}\frac{\sum z_iz_{i+j}}{\sum z_i^2}=n\mathsf{M}\frac{z_iz_{i+j}}{\sum z_i^2}=$$
$$=\frac{n}{n(n-1)}\mathsf{M}\frac{\sum_{i,k=1}^{n}z_iz_k}{\sum z_i^2}=\frac{1}{n-1}\mathsf{M}\frac{(\sum z_i)^2-\sum z_i^2}{\sum z_i^2}=$$
$$=-\frac{1}{n-1}\qquad (i\neq k), \qquad (48.25)$$

так как $\Sigma z_i\equiv 0$. Это согласуется с (48.24) с точностью до порядка n^{-1}. Замечательно, что r_j обладает отрицательным смещением даже для случайных рядов.

Пример 48.6

Если ряд таков, что

$$\rho_1=\rho;\qquad \rho_j=0,\qquad j\neq 0,1;$$

из (48.20) находим, что

$$\mathsf{M}r_1=\rho+(1+\rho)(4\rho^2-2\rho-1)/\nu, \qquad (48.26)$$
$$\mathsf{M}r_2=-(1+2\rho+2\rho^2)/\nu, \qquad (48.27)$$
$$\mathsf{M}r_j=-(1+2\rho)/\nu,\ j>2. \qquad (48.28)$$

Пример 48.7

Для марковского ряда (47.68) с параметром ρ сходным образом находим, что

$$\mathsf{M}r_1=\rho-\frac{1+3\rho}{\nu}, \qquad (48.29)$$

$$\mathsf{M}r_j=\rho^j-\frac{1}{\nu}\left\{\frac{1+\rho}{1-\rho}(1-\rho^j)+2k\rho^j\right\}, \qquad j>1. \qquad (48.30)$$

Во всех этих случаях имеется отрицательное смещение. Очевидно, что оно может быть весьма серьезным. Для марковского ряда из 25 членов при $\rho = {}^1/_2$ среднее значение r_1 было бы приблизительно равно 0,4, а не 0,5.

Поправка Кенуя

48.4 Мы можем воспользоваться упрощенным вариантом метода Кенуя и устранить смещение в духе (40.28), разбивая ряд на два. Пусть r — сериальный коэффициент всего ряда, а $r_{(1)}$ и $r_{(2)}$ — коэффициенты «половинок». Используем статистику

$$R = 2r - \frac{1}{2}(r_{(1)} + r_{(2)}), \tag{48.31}$$

которая является несмещенной с точностью до порядка n^{-1}.

Дальнейшие результаты, касающиеся смещения, можно найти у Мэрриотта и Поупа (1954), Кендалла (1954) и Кенуя (1956) (см. **17.10** т. 2). Для марковского случая Уайт (1961) получил результаты с точностью до порядка n^{-3}.

Некоторые точные результаты

48.5 Теперь мы перейдем к рассмотрению некоторых точных результатов теории распределений сериальных корреляций. Как и следует ожидать, точность можно приобрести лишь ценой некоторых упрощений в постановке. Обычно предполагают нормальность теоретического распределения, иногда также упрощают определения, исследуя более простые статистики. Прежде всего получим некоторые результаты (принадлежащие Морэну) «методом математического ожидания». Затем мы остановимся на некоторых распределениях (полученных Р. Л. Андерсоном и рядом других авторов, в основном Даниэлсом), которые отражают совершенно новые стороны теории распределений.

48.6 Рассмотрим случай нулевых автокорреляций и определим первую сериальную корреляцию формулой

$$r_1 = \frac{n\sum_{i=1}^{n-1}(u_i - \bar{u})(u_{i+1} - \bar{u})}{(n-1)\sum_{i=1}^{n}(u_i - \bar{u})^2}. \tag{48.32}$$

Мы уже показали в (48.25), что

$$\mathsf{M}r_1 = -1/(n-1). \tag{48.33}$$

Пусть

$$I = \frac{n-1}{n} r_1, \qquad z_i = u_i - \bar{u},$$

так что

$$\mathsf{M}I = -\frac{1}{n}.$$

Тогда имеем

$$\mathsf{M}I^2 = \mathsf{M}\left\{\frac{\sum_{i=1}^{n-1} z_i z_{i+1}}{\sum z_i^2}\right\}^2 =$$
$$= \mathsf{M}\left\{\left(\sum z_i^2\right)^{-2}\left[\sum_1^{n-1} z_i^2 z_{i+1}^2 + 2\sum_1^{n-2} z_i z_{i+1}^2 z_{i+2} + \right.\right.$$
$$\left.\left. + \sum z_i z_{i+1} z_k z_{k+1}\right]\right\}, \tag{48.34}$$

где i, $i+1$, k, $k+1$ не совпадают. Итак,

$$\mathsf{M}I^2 = \mathsf{M}\left\{\left(\sum z_i^2\right)^{-2}[(n-1)z_1^2 z_2^2 + 2(n-2)z_1 z_2^2 z_3 + \right.$$
$$\left. + (n-2)(n-3)z_1 z_2 z_3 z_4]\right\},$$

или на языке расширенных симметрических функций (**12.5** т. 1)

$$\mathsf{M}I^2 = \mathsf{M}\left\{[2]^{-2}\left[\frac{1}{n}[2^2] + \frac{2}{n(n-1)}[21^2] + \frac{1}{n(n-1)}[1^4]\right]\right\}. \tag{48.35}$$

Используя таблицу 10 из приложения, выразим расширенные симметрические функции в терминах сумм степеней. Так как здесь $(1) \equiv 0$, получим

$$\mathsf{M}I^2 = \mathsf{M}\left\{\frac{1}{n}\left\{1 - \frac{(4)}{(2)^2}\right\} + \frac{2}{n(n-1)}\left\{\frac{2(4)}{(2)^2} - 1\right\} + \right.$$
$$\left. + \frac{1}{n(n-1)}\left\{3 - \frac{6(4)}{(2)^2}\right\}\right\}. \tag{48.36}$$

Если выборка нормальна, k_4/k_2^2 не зависит от k_2^2 (см. **37.27**), и в силу (12.28)

$$\frac{k_4}{k_2^2} = \frac{n-1}{(n-2)(n-3)}\left\{n(n+1)\frac{(4)}{(2)^2} - 3(n-1)\right\}. \tag{48.37}$$

Следовательно, поскольку выборка нормальна, математическое ожидание k_4/k_2^2 равно нулю, и в силу (48.37)

$$\mathsf{M}\left\{\frac{(4)}{(2)^2}\right\} = \frac{3(n-1)}{n(n+1)}. \tag{48.38}$$

Подстановка в (48.36) и дальнейшие упрощения приводят к формуле

$$\mathsf{M}I^2 = \frac{n^2 - 3n + 3}{n^2(n-1)}. \tag{48.39}$$

Следовательно, в нормальном случае

$$\mathsf{D}I = \frac{(n-2)^2}{n^2(n-1)},$$
$$\mathsf{D}r_1 = \frac{(n-2)^2}{(n-1)^3} = \frac{1}{n+1} + o\left(\frac{1}{n}\right). \tag{48.40}$$

Так как в силу неравенства Коши $n(4) \geqslant (2)^2$, $\mathsf{M}\{(4)/(2)^2\} \geqslant n^{-1}$ при любом теоретическом распределении; и, как указал Морэн (1967), из (48.36) следует совершенно общее утверждение

$$\mathsf{D}r_1 \leqslant \frac{n(n-2)}{(n-1)^3} = \frac{1}{n-1} + o\left(\frac{1}{n}\right). \tag{48.41}$$

Значение (48.40), полученное для нормального случая, почти достигает указанного максимального значения. У распределений с медленно убывающими хвостами $(4)/(2)^2$ велико и $\mathsf{D}r_1$ соответственно мала.

Морэн (1947—8) получил результат для случая циклического определения, а также третьи и четвертые моменты (см. упражнения 48.6, 48.12). Дженкинс (1954b, 1956) использовал сходные методы для получения совместного распределения сериальных корреляций (см. 48.26 и упражнения 48.18 и 48.19).

Распределение Р. Л. Андерсона

48.7 По причинам, которые станут ясны позже, мы рассмотрим теперь распределение первой циклической сериальной корреляции

$$r_1 = \frac{u_1u_2 + u_2u_3 + \ldots + u_{n-1}u_n + u_nu_1 - n\bar{u}^2}{\sum_{i=1}^{n}(u_i - \bar{u})^2}. \tag{48.42}$$

Следуя Р. Андерсону (1942), рассмотрим распределение этой статистики в случае, когда $u_1, \ldots, u_n$ — независимые нормальные величины. Индекс у r_1 будем опускать.

Будем искать линейное преобразование величин $\{u_1, \ldots, u_n\}$ в такие $\xi_1, \ldots, \xi_n$, при котором r принимает форму $\sum \lambda_i \xi_i^2 / \sum \xi_i^2$. Преимущество циклического определения состоит в том, что в этом случае все λ можно определить точно.

Знаменатель в определении r принимает требуемую форму при любом ортогональном преобразовании.

Числитель равен величине

$$q = \sum u_iu_{i+1}\left(1 - \frac{2}{n}\right) - \frac{1}{n}\sum u_i^2 - \frac{2}{n}\sum u_iu_{i+k}, \qquad k > 1. \tag{48.43}$$

Рассмотрим

$$q - \lambda \sum u_i^2. \tag{48.44}$$

Как и в случае главных компонент (см. **43.6**), если мы определим λ, как величину, максимизирующую (48.44), мы придем к желаемой сумме $\sum \lambda_i \xi_i^2$. Нам нет нужды искать само преобразование: величины λ — это все, что нам нужно. Из (48.43—44) получим, что они являются корнями уравнения

$$\begin{vmatrix} -\left(\lambda + \frac{1}{n}\right) & \frac{1}{2}\left(1 - \frac{2}{n}\right) & -\frac{1}{n} & \ldots & -\frac{1}{n} & \frac{1}{2}\left(1 - \frac{2}{n}\right) \\ \frac{1}{2}\left(1 - \frac{2}{n}\right) & -\left(\lambda + \frac{1}{n}\right) & \frac{1}{2}\left(1 - \frac{2}{n}\right) & \ldots & -\frac{1}{n} & -\frac{1}{n} \\ -\frac{1}{n} & \frac{1}{2}\left(1 - \frac{2}{n}\right) & -\left(\lambda + \frac{1}{n}\right) & \ldots & -\frac{1}{n} & -\frac{1}{n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \frac{1}{2}\left(1 - \frac{2}{n}\right) & -\frac{1}{n} & -\frac{1}{n} & \ldots & \frac{1}{n}\left(1 - \frac{2}{n}\right) & -\left(\lambda + \frac{1}{n}\right) \end{vmatrix} = 0. \tag{48.45}$$

Это циркулянтный определитель, и он может быть разложен на множители. Действительно, рассмотрим циркулянт

$$D = \begin{vmatrix} a_1 & a_2 & \ldots & a_n \\ a_n & a_1 & \ldots & a_{n-1} \\ a_{n-1} & a_n & \ldots & a_{n-2} \\ \cdot & \cdot & \cdot & \cdot \\ a_2 & a_3 & \ldots & a_1 \end{vmatrix}. \tag{48.46}$$

Пусть $\omega_1, \omega_2, \ldots, \omega_{n-1}, \omega_n (= 1)$ — корни n-й степени из единицы. Умножая j-й столбец D на ω_k^{j-1} и суммируя по j, получаем, что $\sum_j a_j \omega_k^{j-1}$ является делителем D. Поскольку это верно для всех k

$$D = \prod_{k=1}^{n}\left\{\sum_{j=1}^{n} a_j \omega_k^{j-1}\right\}. \tag{48.47}$$

Далее

$$\sum_{j=1}^{n} \omega_k^{j-1} = \begin{cases} 0, & k \neq n, \\ n, & k = n, \end{cases}$$

и, следовательно,

$$\sum_{j=3}^{n-1} \omega_k^{j-1} = \begin{cases} n - 3, & k = n, \\ -(1 + \omega_k + \omega_k^{-1}), & k \neq n. \end{cases} \tag{48.48}$$

Подставляя соответствующие значения величин a из (48.45) в (48.47) и учитывая (48.48), после некоторых преобразований получаем

$$D = \lambda \prod_{k=1}^{n-1}\left\{-\lambda + \frac{1}{2}(\omega_k + \omega_k^{-1})\right\}.$$

Так как $(\omega_k^{-1}+\omega_k)/2=\cos[2\pi k/n]$,

$$D=\lambda\sum_{k=1}^{n-1}\left(-\lambda+\cos\frac{2\pi k}{n}\right). \qquad (48.49)$$

Таким образом, получаем корни $\lambda=0$, $\lambda=\cos 2\pi k/n$, $k=1, \ldots, n-1$. Решающим моментом является то, что корни появляются парами, за исключением, возможно, одного, равного -1. Итак,

$$q=\begin{cases}\sum\limits_{i=1}^{(n-1)/2}\lambda_i v_i, & \text{если } n \text{ нечетно}, \qquad (48.50)\\ \sum\limits_{i=1}^{(n-2)/2}\lambda_i v_i - v, & \text{если } n \text{ четно}, \qquad (48.51)\end{cases}$$

где v_i имеют χ^2-распределение с двумя ст. св., а v — с одной ст. св. Знаменатель r (обозначим его p) имеет χ^2-распределение с $(n-1)$ ст. св. Итак, задача сведена к нахождению распределения величины

$$r=\begin{cases}\sum\limits_{i=1}^{(n-1)/2}\lambda_i v_i/\sum v_i, & \text{если } n \text{ нечетно}, \qquad (48.52)\\ \left(\sum\lambda_i v_i - v\right)/\left(\sum v_i + v\right), & \text{если } n \text{ четно}. \qquad (48.53)\end{cases}$$

48.8 Эта задача о распределении была решена Р. Андерсоном, чьим выкладкам мы и следовали. Проиллюстрируем метод на частном случае $n=6$, а затем приведем общий результат.

При $n=6$ имеем

$$r=\frac{\lambda_1 v_1+\lambda_2 v_2 - v}{v_1+v_2+v}, \qquad p_6=v_1+v_2+v, \qquad (48.54)$$

где

$$\lambda_1=\cos\frac{2\pi}{6}=\frac{1}{2}, \qquad \lambda_2=-\frac{1}{2}. \qquad (48.55)$$

Совместное распределение величин v дается формулой

$$dF(v_1, v_2, v)=\frac{1}{4(2\pi)^{1/2}}v^{-1/2}e^{-v/2}e^{-v_1/2}e^{-v_2/2}\,dv\,dv_1\,dv_2= \qquad (48.56)$$

$$=\frac{1}{4(2\pi)^{1/2}}v^{-1/2}e^{-p_6/2}\,dv\,dv_1\,dv_2. \qquad (48.57)$$

Мы вынуждены рассмотреть два случая:

$$\lambda_2\leqslant r\leqslant\lambda_1 \quad \text{и} \quad -1\leqslant r\leqslant\lambda_2.$$

(Как следует из (48.52—3), $r\leqslant\lambda_1$, где λ_1 — наибольший корень.) Рассмотрим первый случай.

Согласно (48.54) имеем

$$v_2=\frac{1}{\lambda_1-\lambda_2}\{p_6(\lambda_1-r)-v(1+\lambda_1)\}, \qquad (48.58)$$

$$v_1=\frac{1}{\lambda_1-\lambda_2}\{p_6(r-\lambda_2)+v(1+\lambda_2)\}. \qquad (48.59)$$

Якобиан преобразования $\frac{\partial(v_1, v_2)}{\partial(p_6, r)}=\frac{p_6}{\lambda_1-\lambda_2}$. Отсюда и из (48.57) для распределения p_6, r и v имеем

$$dF=\frac{1}{4(2\pi)^{1/2}}\frac{v^{-1/2}p_6e^{-p_6/2}}{\lambda_1-\lambda_2}\,dp_6\,dr\,dv. \qquad (48.60)$$

Интегрируя по v в пределах, определяемых (48.58) и (48.59), получаем

$$dF=\frac{1}{4(2\pi)^{1/2}}\frac{p_6e^{-p_6/2}}{\lambda_1-\lambda_2}\left[2v^{1/2}\right]_0^{p_6\left(\frac{\lambda_1-r}{1+\lambda_1}\right)}dp_6\,dr=$$

$$=\frac{1}{2(2\pi)^{1/2}(\lambda_1-\lambda_2)(1+\lambda_1)^{1/2}}p_6^{3/2}e^{-p_6/2}(\lambda_1-r)^{1/2}\,dp_6\,dr. \qquad (48.61)$$

Наконец, интегрируя по p_6 от 0 до ∞, получаем

$$dF(r)=\frac{3}{2}\frac{(\lambda_1-r)^{1/2}}{(\lambda_1-\lambda_2)(1+\lambda_1)^{1/2}}\,dr, \qquad \lambda_2\leqslant r\leqslant\lambda_1. \qquad (48.62)$$

Второй случай рассматривается аналогично. При $-1\leqslant r\leqslant\lambda_2$ имеем

$$dF(r)=\left\{\frac{3}{2}\frac{(\lambda_1-r)^{1/2}}{(\lambda_1-\lambda_2)(1+\lambda_1)^{1/2}}+\frac{3}{2}\frac{(\lambda_2-r)^{1/2}}{(\lambda_2-\lambda_1)(1+\lambda_2)^{1/2}}\right\}dr. \qquad (48.63)$$

48.9 Для таких распределений типично, что их аналитические представления принимают различную форму в зависимости от того, между какими критическими значениями λ лежит аргумент r. Кривые плотности этих распределений непрерывны, но производные более высоких порядков уже могут быть разрывны. Также типично, что явное выражение функции распределения легко выписать. Например, в силу (48.62) и (48.63)

$$\mathsf{P}(R>r)=\begin{cases}\dfrac{(\lambda_1-r)^{3/2}}{(\lambda_1-\lambda_2)(1+\lambda_1)^{1/2}}, & \lambda_2\leqslant r\leqslant\lambda_1, \qquad (48.64)\\[2ex] \dfrac{(\lambda_1-r)^{3/2}}{(\lambda_1-\lambda_2)(1+\lambda_1)^{1/2}}+\dfrac{(\lambda_2-r)^{3/2}}{(\lambda_2-\lambda_1)(1+\lambda_2)^{1/2}}, & -1\leqslant r\leqslant\lambda_2. \qquad (48.65)\end{cases}$$

Приведем общий вид «дополнения» к функции распределения:

$$\mathbf{P}(R > r) =$$

$$= \begin{cases} \sum_{i=1}^{m} \dfrac{(\lambda_i - r)^{(n-3)/2}}{\prod_{j=1,\, i \neq j}^{(n-1)/2} (\lambda_i - \lambda_j)}, & \lambda_{m+1} \leqslant r \leqslant \lambda_m,\ n \text{ нечетно}, \quad (48.66) \\ \sum_{i=1}^{m} \dfrac{(\lambda_i - r)^{(n-3)/2}}{\prod_{j=1,\, i \neq j}^{(n-2)/2} (\lambda_i - \lambda_j)(1 + \lambda_i)^{1/2}}, & \lambda_{m+1} \leqslant r \leqslant \lambda_m,\ n \text{ четно}. \quad (48.67) \end{cases}$$

Плотность имеет различный вид на $(n-1)/2$ или $(n-2)/2$ интервалах в зависимости от четности n.

48.10 Теперь бегло отметим несколько особенностей этого распределения, представляющих скорее математический, чем статистический, интерес:

(а) Вычисляя r_l при $l \neq 1$, мы приходим к циркулянту с сомножителями типа $\lambda_k - \cos(2\pi lk/n)$.

Если l не имеет общих делителей с n, циркулянт остается тем же самым, и распределение не меняется. В противном случае аналитическое выражение для распределения изменится.

(б) Существуют иные способы вычислений, также приводящие к циркулянту, даже в том случае, когда коэффициент определен не циклично. Например, при $n = 2m$ для коэффициента

$$(u_1u_2 + \ldots + u_{m-1}u_m + u_{m+1}u_{m+2} + \ldots + u_{n-1}u_n)/\sum u^2$$

характеристическое свойство парности корней λ сохраняется и, следовательно, мы снова приходим к распределению Андерсона. Другие случаи рассмотрены Дербином и Ватсоном (1950—1).

48.11 Перейдем к вычислению характеристической функции q и p (числителя и знаменателя r). Временно символами s и t будем обозначать аргументы функции. Совместная х. ф. q и p

$$\varphi(s, t) \propto \int \exp\left\{-\frac{1}{2}\left[\sum u^2 - 2it \sum (u - \bar{u})^2 - \right.\right.$$
$$\left.\left. - 2is(u_1u_2 + \ldots + u_nu_1 - n\bar{u}^2)\right]\right\} du = \Delta^{-1/2}, \quad (48.68)$$

где

$$\Delta = \begin{vmatrix} 1 - 2it\left(1 - \frac{1}{n}\right) + \frac{2is}{n} & 1 + \frac{it}{n} & \ldots & \frac{it}{n} - is + \frac{2}{n} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{it}{n} - is + \frac{2}{n} & 1 + \frac{it}{n} & \ldots & 1 - 2it\left(1 - \frac{1}{n}\right) + \frac{2is}{n} \end{vmatrix}. \quad (48.69)$$

Это циркулянт того же типа, что и в (48.45), и, следовательно, как и в (48.49),

$$\Delta = \prod_{k=1}^{n-1} \left\{1 - 2i\left(t + s\cos\frac{2\pi k}{n}\right)\right\}. \quad (48.70)$$

Логарифмируя и вычисляя коэффициенты, получаем, что семиинварианты q и p

$$\varkappa_{ij} = i!\,j! \sum_{k=1}^{n-1} \frac{2^{i+j-1}}{i+j} \binom{i+j}{i} \left(\cos\frac{2\pi k}{n}\right)^i = \quad (48.71)$$

$$= 2^r r! \sum \left(\cos\frac{2\pi k}{n}\right)^i, \qquad r = i + j - 1. \quad (48.72)$$

Далее

$$\sum \cos\frac{2\pi k}{n} = -1 = \sum \cos^3\frac{2\pi k}{n} = \sum \cos^5\frac{2\pi k}{n} = \ldots \quad (48.73)$$

$$\sum \cos^2\frac{2\pi k}{n} = \frac{1}{2}(n-2), \quad \sum \cos^4\frac{2\pi k}{n} = \frac{3n-8}{8}. \quad (48.74)$$

Подставляя, находим

$$\varkappa_{10} = -1, \quad \varkappa_{01} = n - 1, \quad \varkappa_{20} = n - 2, \quad \varkappa_{11} = -2, \quad \varkappa_{02} = 2(n-1);$$
$$\varkappa_{30} = -8, \quad \varkappa_{21} = 4(n-2), \quad \varkappa_{12} = -8; \quad \varkappa_{03} = 8(n-1). \quad (48.75)$$

Семиинварианты $\varkappa_{20}$ и $\varkappa_{02}$ имеют порядок n, в то время как семиинварианты «более высоких порядков не имеют более высокого порядка». Таким образом, после стандартной нормировки семиинварианты более высоких порядков стремятся к нулю, а соответствующее распределение сходится к двумерному нормальному. Более того, после стандартной нормировки $\varkappa_{11}$ стремится к нулю, так что q и p асимптотически некоррелированы, а следовательно, в силу нормальности и асимптотически независимы.

В действительности при нормальном распределении r и p независимы и, следовательно,

$$\mathbf{M}(r^m p^m) = \mathbf{M}r^m \mathbf{M}p^m = \mathbf{M}q^m,$$
$$\mathbf{M}r^m = \mathbf{M}q^m / \mathbf{M}p^m. \quad (48.76)$$

Тогда мы можем вычислить моменты r, используя (48.75). Находим

$$\mu_1(r) = -1/(n-1),$$
$$\mu_2(r) = \frac{n(n-3)}{(n+1)(n-1)^2}, \quad (48.77)$$
$$\mu_3(r) = \frac{2(2n-1)(n-6)}{(n-1)^3(n+1)(n+3)}. \quad (48.78)$$

Среднее согласуется с точным результатом (48.25) для нециклического случая. Дисперсия согласуется с «нециклической» дисперсией (48.40) только с точностью до порядка n^{-1}.

48.12 Диксон (1944), искусно сглаживая характеристическую функцию (48.68—70), получил приближенное представление распределения Андерсона. Пусть временно

$$\alpha = 1 - 2it, \qquad \beta = -2is, \qquad \theta_k = 2\pi k/n. \tag{48.79}$$

Тогда приближенно имеем

$$\begin{aligned}
\varphi(s, t) &\propto \prod_{k=1}^{n-1} (\alpha + \beta\cos\theta_k)^{-1/2} \doteqdot \\
&\doteqdot (\alpha+\beta)^{1/2} \exp\left[-\frac{1}{2}\log\prod_{k=1}^{n}(\alpha+\beta\cos\theta_k)\right] \doteqdot \\
&\doteqdot (\alpha+\beta)^{1/2} \exp\left[-\frac{n}{4\pi}\cdot\frac{2\pi}{n}\sum_{k=1}^{n}\log(\alpha+\beta\cos\theta_k)\right] \doteqdot \\
&\doteqdot (\alpha+\beta)^{1/2} \exp\left[-\frac{n}{4\pi}\int_0^{2\pi}\log(\alpha+\beta\cos\theta)\,d\theta\right] \doteqdot \\
&\doteqdot (\alpha+\beta)^{1/2} \exp\left[-\frac{n}{2}\log\left\{\frac{1}{2}\left(\alpha+(\alpha^2-\beta^2)^{1/2}\right)\right\}\right] \doteqdot \\
&\doteqdot 2^{n/2}(\alpha+\beta)^{1/2}\left\{\alpha+(\alpha^2-\beta^2)^{1/2}\right\}^{-n/2}.
\end{aligned} \tag{48.80}$$

Теперь, последовательно дифференцируя, можно найти моменты q и p. (На самом деле дифференцируя по α и интегрируя по β, можно найти даже моменты типа $\mathsf{M}(q^k/p^k)$.) Итак, приходим к неожиданно простым выражениям:

$$\mu_1' = -\frac{1}{n-1}, \tag{48.81}$$

$$\mu_2' = \frac{1}{n+1}, \tag{48.82}$$

$$\mu_3' = -\frac{3}{(n-1)(n+3)}, \tag{48.83}$$

$$\mu_4' = \frac{3}{(n+1)(n+3)}. \tag{48.84}$$

Можно показать (Диксон, 1944), что на самом деле эти формулы точны. Можно убедиться, что моменты величины r^2, вплоть до порядка $n/2$, совпадают с моментами распределения

$$\frac{\Gamma\left(\frac{n}{2}+\frac{1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{1}{2}\right)} r^{-1}(1-r^2)^{(n-2)/2}\,dr^2. \tag{48.85}$$

Итак, учитывая (16.62), получаем, что квадрат сериальной корреляции имеет приблизительно то же распределение, что и квадрат обыкновенного коэффициента корреляции, полученного по выборке объема $n+2$ из некоррелированной нормальной генеральной совокупности.

Диксон (1944) изучал также случай, когда ряд имеет известное (нулевое) среднее, и случай, когда запаздывание (в коэффициенте) $l > 1$. Справедливы те же результаты вплоть до моментов порядка $n/2m$, где m — наибольший общий делитель l и n. В случае известного среднего надо лишь n заменить на $n+1$. (См. упражнение 48.20.)

48.13 Купмэнс (1942) получил тот же результат иным путем. Он выразил х. ф. p и q в форме контурного интеграла и до интегрирования «сгладил» значения λ, распределив их равномерно по кругу. Это привело его к выражению

$$h(r) = \frac{\left(\frac{n}{2}-1\right)2^{n/2}}{\pi}\int_0^{\arccos r}(\cos\alpha - r)^{\frac{1}{2}n-\alpha}\sin\frac{n\alpha}{2}\sin\alpha\,d\alpha. \tag{48.86}$$

Х. Рубин (1945), показав с помощью интегрирования по частям, что

$$\frac{\partial}{\partial r}h(r, n) = -nrh(r, n-2),$$

и используя потом индукцию, вычислил этот интеграл.

48.14 Теперь нам лучше остановиться и попытаться кратко оценить уже достигнутое. Мы получили в явном виде формулы для больших выборок, однако эти формулы не слишком удобны для применения. Отметим также, что, за исключением физики и метеорологии, временные ряды большой длины встречаются довольно редко. В случае малых выборок нами получены точные результаты для случайных рядов, но использовать построенные нами критерии мы можем лишь после того, как убедимся в «случайности» ряда с помощью более простых критериев. Кроме того, точные результаты нами получены в предположении нормальности и большей частью цикличности в определении коэффициентов. Тем не менее все эти результаты весьма наглядны и порождают ряд других интересных задач.

Распределение Мэдоу — Лейпника

48.15 Если статистики $t_1, \ldots, t_k$ достаточны для $\theta_1, \ldots, \theta_k$ в обычных обозначениях, имеем

$$P(u_1, \ldots, u_n \mid \theta) = Q(t \mid \theta)\,R(u),$$

и, следовательно,

$$P(t \mid \theta) = P(t, \theta_0)\frac{Q(t \mid \theta)}{Q(t \mid \theta_0)}. \tag{48.87}$$

Мэдоу (1945), используя этот результат, исследовал связь между приближенным распределением сериальных корреляций для «ненулевого» случая и тем же распределением для «нулевого» случая $\theta = \theta_0$.

Предположим, что $u_1, \ldots, u_n$ имеют совместное нормальное распределение следующего вида:

$$\log L = \text{const} - \frac{1}{2}\left[A\sum_{i=1}^{n}(u_i-\mu)^2 + 2B\sum_{i=1}^{n}(u_i-\mu)(u_{i+1}-\mu)\right]. \tag{48.88}$$

Среднее здесь равно μ.

Как и прежде, будем предполагать, что r определен циклично с q и p в качестве числителя и знаменателя. Тогда $\bar{u}$, p и q достаточны для μ, A и B. В качестве «нулевой» гипотезы возьмем случай Андерсона $A = 1$, $B = 0$. Тогда из (48.87) получаем

$$P(r, p \mid A, B) \propto P(r, p \mid 1, 0)\frac{e^{-(Ap+2Brp)/2}}{e^{-p/2}}.$$

Так как r и p независимы, а p кратно величине χ^2 с $n-1$ ст. св., то

$$P(r, p \mid A, B) \propto p^{(n-3)/2}e^{-p(A+2Br)/2}f(r). \tag{48.89}$$

Интегрируя по p от 0 до ∞, получаем

$$P(r \mid A, B) \propto \frac{1}{\left(\frac{1}{2}A + Br\right)^{(n-1)/2}}f(r). \tag{48.90}$$

Таким образом, «ненулевое» распределение отличается от «нулевого» множителем вида $\left(\frac{A}{2}+Br\right)^{-(n-1)/2}$.

В том случае, когда известно, что все u имеют нулевое среднее, также справедлива формула (48.90) с показателем степени в знаменателе, равным $n/2$.

48.16 Рассмотрим частный случай марковского ряда

$$u_t = \rho u_{t-1} + \varepsilon_t.$$

Известно (см. (47.70)), что

$$\mathbf{D}u = \frac{1}{1-\rho^2}\mathbf{D}\varepsilon$$

и

$$\operatorname{cov}(u_t, u_{t+1}) = \rho\,\mathbf{D}u = \frac{\rho}{1-\rho^2}\mathbf{D}\varepsilon.$$

Тогда функция правдоподобия для ε дается формулой

$$\log L = \text{const} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\varepsilon_i^2 = \text{const} - \frac{1}{2\sigma^2}\sum(u_i - \rho u_{i-1})^2,$$

где $\sigma^2 = \mathbf{D}\varepsilon$. Это приблизительно эквивалентно равенству

$$\log L = \text{const} - \frac{1}{2\sigma^2}(1+\rho^2)\sum u_i^2 + \frac{\rho}{\sigma^2}\sum u_i u_{i+1}. \tag{48.91}$$

Отсюда для A и B из (48.88) получаем

$$A = \frac{1+\rho^2}{\sigma^2}, \qquad B = -\frac{\rho}{\sigma^2},$$

и, следовательно,

$$\frac{1}{2}A + Br \propto (1 - 2\rho r + \rho^2). \tag{48.92}$$

Таким образом, заменяя в (48.90) показатель степени в знаменателе на $n/2$, а $f(r)$ — на приближенное выражение (48.85), получаем

$$P(r \mid \rho) = \frac{\Gamma\left(\frac{1}{2}n+1\right)}{\Gamma\left(\frac{n}{2}+\frac{1}{2}\right)\Gamma\left(\frac{1}{2}\right)}\cdot\frac{(1-r^2)^{(n-1)/2}}{(1-2\rho r+\rho^2)^{n/2}}, \qquad -1 \leqslant r \leqslant 1. \tag{48.93}$$

48.17 Это замечательное представление изучалось Лейпником (1947), Кенуем (1948), Дженкинсом (1956) и Кендаллом (1957). Получить прямым интегрированием моменты распределения (48.93) не так просто, как это кажется на первый взгляд.

Обозначив символом C постоянную, для момента k-го порядка относительно начала координат (нецентральный момент) получаем

$$\mu_k' = C\int_{-1}^{1}\frac{(1-r^2)^{(n-1)/2}r^k}{(1+\rho^2-2\rho r)^{n/2}}dr,$$

$$-\frac{1}{n}\frac{\partial\mu_k'}{\partial\rho} = \int_{-1}^{1}\frac{\rho r r^{k-1}}{(1+\rho^2-2\rho r)^{\frac{n}{2}+1}}dF - \int_{-1}^{1}\frac{(1+\rho^2)-(1+\rho^2-2\rho r)}{2\rho(1+\rho^2-2\rho r)^{\frac{n}{2}+1}}dF =$$

$$= \frac{1}{2}\int\frac{(1+\rho^2)-(1+\rho^2-2\rho r)}{(1+\rho^2-2\rho r)^{\frac{n}{2}+1}}r^{k-1}dF -$$

$$-\frac{1+\rho^2}{2\rho}\int\frac{r^k\,dF}{(1+\rho^2-2\rho r)^{\frac{n}{2}+1}} + \frac{1}{2\rho}\mu_k' =$$

$$= \frac{1}{2\rho}\mu_k' - \frac{1}{2}\mu_{k-1}' + \frac{1+\rho^2}{2\rho}\int\frac{(\rho-r)r^{k-1}}{(1+\rho^2-2\rho r)^{\frac{n}{2}+1}}dF.$$

Отсюда находим

$$\left(\frac{1}{n}\frac{\partial}{\partial\rho}+\frac{1}{2\rho}\right)\mu'_k=\left(\frac{1+\rho^2}{2\rho n}\frac{\partial}{\partial\rho}+\frac{1}{2}\right)\mu'_{k-1}. \qquad (48.94)$$

Легко проверить индукцией, что μ'_k — полином k-го порядка относительно ρ. Более того, четные моменты содержат лишь четные степени ρ, нечетные моменты — лишь нечетные.

Пусть

$$\mu'_k=\sum_{m=0}^{k}a_{km}\rho^m. \qquad (48.95)$$

Дифференцируя (48.94) m раз и положив $\rho=0$, получаем

$$a_{km}=\frac{1}{n+2m}\{(m+1)a_{k-1,\,m+1}+(n+m-1)a_{k-1,\,m-1}\}. \qquad (48.96)$$

Так как $\mu'_0=1$, $a_{00}=1$, то

$$a_{11}=\frac{n}{n+2}$$

и, следовательно,

$$\mu'_1=\frac{n\rho}{n+2}. \qquad (48.97)$$

Последовательное использование (48.96) приводит к формулам

$$\mu'_2=\frac{1}{n+2}+\frac{n(n+1)}{(n+2)(n+4)}\rho^2, \qquad (48.98)$$

$$\mu'_3=\frac{3n\rho}{(n+2)(n+4)}+\frac{n(n+1)}{(n+4)(n+6)}\rho^3, \qquad (48.99)$$

$$\mu'_4=\frac{3}{(n+2)(n+4)}+\frac{6n(n+1)\rho^2}{(n+2)(n+4)(n+6)}+\frac{n(n+1)(n+3)}{(n+4)(n+6)(n+8)}\rho^4 \qquad (48.100)$$

и т. д. В частности, моменты относительно среднего (центральные моменты)

$$\mu_2=\frac{1}{n+2}-\frac{n(n-2)}{(n+2)^2(n+4)}\rho^2\sim\frac{1-\rho^2}{n}, \qquad (48.101)$$

$$\mu_3=\frac{-6n\rho}{(n+2)^2(n+4)}+\frac{2n(n-2)(3n-2)}{(n+2)^3(n+4)(n+6)}\rho^3\sim\frac{-6\rho(1-\rho^2)}{n^2}, \qquad (48.102)$$

$$\mu_4\sim\frac{3(1-\rho^2)^2}{n^2}. \qquad (48.103)$$

Отметим, в частности, что при стандартной нормировке μ_3 стремится к нулю, а μ_4 — к 3, что указывает на сходимость распределения к нормальному.

Уайт (1957) и Лейпник (1958) вычисляли моменты в виде полиномов типа Гегенбауэра. Лейпник получил выражение для характеристической функции и показал, что последняя сходится к нормальной.

48.18 Приближенное выражение для дисперсии в (48.101)

$$\mathrm{D}r=\frac{1-\rho^2}{n}$$

(заметим мимоходом, что оно отличается от выражения для смешанных моментов) подсказывает нам нормализующее преобразование

$$r=\sin z, \qquad \rho=\sin\zeta.$$

Последнее использовал Дженкинс (1954а), который для $x=z-\zeta$ получил

$$\mu'_1(x)=-\frac{3}{2}\frac{\rho}{(1-\rho^2)^{1/2}}\cdot\frac{1}{n}+O(n^{-2}), \qquad (48.104)$$

$$\mu'_2(x)=\frac{1}{n}-\frac{2-5\rho^2}{2(1-\rho^2)}\frac{1}{n^2}+O(n^{-3}), \qquad (48.105)$$

$$\gamma_1=\frac{-3\rho}{(1-\rho^2)^{1/2}}n^{-1/2}+O(n^{-1}), \qquad (48.106)$$

$$\gamma_2=\frac{2(7\rho^2-1)}{1-\rho^2}n^{-1}+O(n^{-2}). \qquad (48.107)$$

Очевидно, что эти формулы не работают при ρ, близком к единице, но при уменьшении ρ они вполне удовлетворительны. (См. упражнение 48.17.)

Приближения Даниэлса

48.19 Значительное развитие теория выборочных распределений сериальных корреляций получила благодаря Даниэлсу (1956), который впервые систематически использовал метод перевала *). Детальное описание этого метода можно найти в книге Джефри и Джефри «Методы математической физики» (Кембридж, 1956). Кратко суть метода состоит в следующем. В комплексной плоскости интеграл от аналитической функции по замкнутому контуру, ограничивающему область, не содержащую особых точек, равен нулю. Следовательно, вычисляя интеграл по любому пути (не обязательно замкнутому), можно один путь заменять другим (фиксируя начальную и конечную точки), если область, ограниченная двумя путями, не содержит особых

*) В оригинале — Saddle-point method (метод седловых точек). Подробное описание этого метода читатель может найти в монографии М. А. Лаврентьев и Б. В. Шабат, Методы теории функций комплексного переменного, «Наука», 1973. (*Прим. ред.*)

точек. Предположение, на котором основывается метод, состоит в том, что значения функций резко убывают после прохождения через точку достижения максимума и, следовательно, в районе седловой точки наблюдается наиболее высокая концентрация значений функции. Поэтому нам следует выбрать путь, проходящий через седловую точку поверхности и наиболее круто спускающийся после прохождения через эту точку.

Заменяя полное интегрирование интегрированием в районе «перевала», мы получаем хорошее приближение к истинному значению интеграла.

На практике метод оказывается полезным при вычислении средних или отношений средних различных величин.

48.20 Пусть, как и прежде, $r = q/p$. Пусть $M(\theta_1, \theta_2)$ — производящая функция моментов величин p и q, т. е.

$$M(\theta_1, \theta_2) = \int \exp\{\theta_1 p + \theta_2 q\}\, dF. \qquad (48.108)$$

Используя обратное преобразование Фурье, имеем

$$f(p, q) = \frac{1}{(2\pi i)^2} \int\int M(\theta_1, \theta_2) \exp\{-(\theta_1 p - \theta_2 q)\}\, d\theta_1\, d\theta_2. \qquad (48.109)$$

Интегрирование ведется вдоль мнимых осей плоскостей θ_1 и θ_2. В частности,

$$f(p, rp) = \frac{1}{(2\pi i)^2} \int\int M(\theta_3 - r\theta_2, \theta_2)\, e^{-p\theta_3}\, d\theta_3\, d\theta_2,$$

где интегрирование по $\theta_3 = \theta_1 + r\theta_2$ ведется вдоль мнимой оси плоскости θ_3 или по любому иному пути, переход к которому допускается. Обращение относительно θ_3 приводит к тому, что

$$\int_0^\infty f(p, rp)\, e^{\theta_3 p}\, dp = \frac{1}{2\pi i} \int M(\theta_3 - r\theta_2, \theta_2)\, d\theta_2. \qquad (48.110)$$

Поэтому, если дифференцирование допустимо, то

$$\int_0^\infty f(p, rp)\, e^{\theta_3 p} p\, dp = \frac{1}{2\pi i} \int \frac{\partial M(\theta_3 - r\theta_2, \theta_2)}{\partial \theta_3}\, d\theta_2, \qquad (48.111)$$

и, поскольку при $\theta_3 = 0$ левая часть равна плотности распределения r,

$$h(r) = \frac{1}{2\pi i} \int \frac{\partial M(\theta_3 - r\theta_2, \theta_2)}{\partial \theta_3}\Big|_{\theta_3=0} d\theta_2. \qquad (48.112)$$

Это эквивалентно результату Гири ((11,78) т. 1). Этот результат иным способом получен в упражнении 11.24.

Часто оказывается нужным преобразовать q, а следовательно, и θ_2. Если это преобразование таково, что $\theta_2 = \theta_2(\theta_4, \theta_3)$, то (48.112) переходит в формулу

$$h(r) = \frac{1}{2\pi i} \int \frac{\partial}{\partial \theta_3}\left[\{M(\theta_3 - r\theta_2, \theta_2)\} \frac{\partial \theta_2}{\partial \theta_4}\right]_{\theta_3=0} d\theta_4. \qquad (48.113)$$

48.21 Прежде всего рассмотрим циклично определенные коэффициенты циклического марковского процесса в случае, когда известно, что среднее равно нулю, а дисперсия — единице. Совместное распределение величин u запишется в виде (см. 48.91))

$$dF = \frac{1-\rho^n}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2}\left\{(1+\rho^2)\sum_{i=1}^n u_i^2 - 2\rho \sum_{i=1}^n u_i u_{i+1}\right\}\right] du_1 \ldots du_n, \qquad (48.114)$$

где, конечно, $u_1 = u_{n+1}$.

П. ф. м. (ср. с (48.69)) равна

$$M(\theta_1, \theta_2) = (1-\rho^n)\,\Delta^{-1/2}, \qquad (48.115)$$

где

$$\Delta = \begin{vmatrix} 1+\rho^2-2\theta_1 & -(\rho+\theta_2) & 0 & \cdots & -(\rho+\theta_2) \\ -(\rho+\theta_2) & 1+\rho^2-2\theta_1 & -(\rho+\theta_2) & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ -(\rho+\theta_2) & 0 & 0 & \cdots & 1+\rho^2-2\theta_1 \end{vmatrix}. \qquad (48.116)$$

Это циркулянт, который преобразуется к виду

$$\Delta = \{1+\rho^2-2\theta_1-2(\rho+\theta_2)\} \prod_{k=1}^{n-1}\left\{1+\rho^2-2\theta_1-2(\rho+\theta_2)\cos\frac{2\pi k}{n}\right\}.$$

Положим

$$\frac{1+\rho^2-2\theta_1}{\rho+\theta_2} = z + \frac{1}{z}. \qquad (48.117)$$

Тогда Δ преобразуется к виду

$$\Delta = \frac{(\rho+\theta_2)^n}{z^n} \prod\left(z^2 - 2z\cos\frac{2\pi k}{n} + 1\right) = \frac{(\rho+\theta_2)^n (1-z^n)^2}{z^n}. \qquad (48.118)$$

Следовательно,

$$M(\theta_3 - r\theta_2, \theta_2) = \frac{1-\rho^n}{1-z^n} \cdot \frac{(1-2rz+z^2)^{n/2}}{(1-2\rho r+\rho^2-2\theta_3)^{n/2}}, \qquad (48.119)$$

где

$$\theta_2 = -\rho + \frac{z(1-2\rho r+\rho^2-2\theta_3)}{1-2rz+z^2}. \qquad (48.120)$$

Тогда

$$\left(\frac{\partial \theta_2}{\partial z}\right)_{\theta_3} M = \frac{(1-\rho^n)(1-z^2)(1-2rz+z^2)^{\frac{n}{2}-2}}{(1-z^n)(1-2\rho r+\rho^2-2\theta_3)^{\frac{n}{2}-1}}, \qquad (48.121)$$

и (48.113) принимает вид

$$h(r)=\frac{(n-2)(1-\rho^n)}{2\pi i(1-2\rho r+\rho^2)^{n/2}}\int\frac{(1-z^2)(1-2rz+z^2)^{\frac{n}{2}-2}}{1-z^n}\,dz. \quad (48.122)$$

Нам осталось определить путь интегрирования. Рассмотрим пару преобразований, которые вместе с (48.117) составляют

$$z+\frac{1}{z}=\zeta,\qquad \zeta-r=\frac{1-2\rho r+\rho^2}{2(\rho+\theta_2)}. \quad (48.123)$$

Область $|z|<1$ переходит во всю плоскость θ_2 с вырезанной на действительной оси внешностью интервала

$$\left(-\frac{(1+\rho)^2}{2(1+r)},\ \frac{(1-\rho)^2}{2(1-r)}\right).$$

Путь в плоскости θ_2, идущий из «точки $\tau-i\infty$» в «точку $\tau'+i\infty$» и проходящий через лакуну на действительной оси, соответствует пути в области $|z|<1$, соединяющему точки $e^{i\varphi}$ и $e^{-i\varphi}$, где $\cos\varphi=r$, и проходящему через точку $z=r$. Следовательно, путь интегрирования в (48.122) должен иметь такой вид.

Рассмотрим множитель

$$(1-2rz+z^2)^{\frac{1}{2}n-2}=(1-r^2-(z-r)^2)^{\frac{1}{2}n-2}$$

в подинтегральной функции из (48.122). Его седловая точка (точка максимума) соответствует $z=r$ (действительное значение). Путь, наиболее «круто» спускающийся из седловой точки, соответствует перпендикуляру к действительной оси. (Доказательство этого факта мы опускаем.) Таким образом, мы должны интегрировать вдоль отрезка, соединяющего $e^{i\varphi}$ и $e^{-i\varphi}$.

До сих пор наши формулы были точны. Заменим теперь в (48.122) ρ^n на нуль и $1-z^n$ — на единицу, полагая, что вклад этих выражений мал. Тогда получим, что

$$h(r)\sim\frac{n-2}{2\pi i(1-2\rho r+\rho^2)^{n/2}}\int(1-z^2)(1-2rz+z^2)^{\frac{n}{2}-2}\,dz.$$

Положим

$$z=r+iw(1-r^2)^{1/2}. \quad (48.124)$$

Тогда $-1\leqslant w\leqslant 1$ и

$$h(r)=\frac{(n-2)(1-r^2)^{(n-1)/2}}{2\pi(1-2\rho r+\rho^2)^{n/2}}\int_{-1}^{1}(1+w^2)(1-w^2)^{\frac{n}{2}-2}\,dw=$$

$$=\frac{\Gamma\left(\frac{n}{2}+1\right)(1-r^2)^{(n-1)/2}}{\pi^{1/2}\Gamma\left(\frac{n}{2}+\frac{1}{2}\right)(1-2\rho r+\rho^2)^{n/2}}. \quad (48.125)$$

Мы снова пришли к распределению Мэдоу — Лейпника (48.93).

48.22 Если замена множителя $(1-z^n)^{-1}$ в (48.122) на единицу нас не устраивает, мы не можем вычислить интеграл в явном виде. Разлагая, однако, этот множитель по степеням z и интегрируя соответствующий ряд, можно получить, что

$$h(r)=\frac{\Gamma\left(\frac{n}{2}+1\right)(1-\rho^n)}{\sqrt{\pi}(1-2\rho r+\rho^2)^{n/2}}\left\{\frac{(1-r^2)^{(n-1)/2}}{\Gamma\left(\frac{n}{2}+\frac{1}{2}\right)}-\right.$$

$$-\frac{3}{2^n\Gamma\left(\frac{3n}{2}+\frac{1}{2}\right)}\frac{d^n}{dr^n}(1-r^2)^{(3n-1)/2}+$$

$$\left.+\frac{5}{2^{2n}\Gamma\left(\frac{5n}{2}+\frac{1}{2}\right)}\frac{d^{2n}}{dr^{2n}}(1-r^2)^{(5n-1)/2}-\ldots\right\}. \quad (48.126)$$

Даниэлс (1956), которому принадлежит этот результат, оценил сверху ошибку, получающуюся при (48.122), первым членом (48.126). Эта ошибка мала при ρ, близком к нулю, но оказывается значимой, например, при ρ, близком к 0,5 и n порядка 20.

48.23 Используя описанный метод, Даниэлс получил еще ряд результатов, которые мы приведем без подробных доказательств.

В случае циклического марковского процесса с неизвестным средним для циклического сериального коэффициента корреляции

$$h(r)=\frac{(n-3)(1-\rho^n)}{2\pi i(1-\rho)(1-2\rho r+\rho^2)^{(n-1)/2}}\int\frac{(1-z)(1-z^2)(1-2rz+z^2)^{(n-5)/2}}{1-z^n}\,dz. \quad (48.127)$$

Снова «игнорируя» z^n, получаем, что

$$h(r)=\frac{n\Gamma\left(\frac{n}{2}-\frac{1}{2}\right)(1-r^2)^{\frac{n}{2}-1}}{2\pi^{1/2}\Gamma\left(\frac{n}{2}\right)(1-\rho)(1-2\rho r+\rho^2)^{(n-1)/2}}\left\{1-r-\frac{1}{n}(1+r)\right\}. \quad (48.128)$$

Следует отметить, что мы можем получить моменты этого распределения, основываясь на моментах распределения Мэдоу — Лейпника (48.93).

Альтернативная форма может быть получена из следующих соображений. Последний член в (48.128) имеет относительно первого порядок $1/n$. Если мы заменим в нем r на ρ, результат останется верным (с точностью до порядка $1/n$). Следовательно, мы можем этим членом пренебречь, но тогда постоянные в

(48.128) следует пересчитать, чтобы интеграл от $h(r)$ по всей области определения был равен единице. Окончательно

$$h(r)=\frac{\Gamma\left(\frac{n}{2}+\frac{3}{2}\right)}{2\pi^{1/2}\Gamma\left(\frac{n}{2}\right)}\frac{(1-r)(1-r^2)^{\frac{n}{2}-1}}{(1-2r\rho+\rho^2)^{(n-1)/2}}\{1+O(n^{-3/2})\}. \qquad (48.129)$$

48.24 В случае нециклического марковского процесса с известным средним рассмотрим нециклическую статистику

$$r=\frac{u_1u_2+\ldots+u_{n-1}u_n}{\frac{1}{2}u_1^2+u_2^2+\ldots+u_{n-1}^2+\frac{1}{2}u_n^2}. \qquad (48.130)$$

Даниэлс показал, что

$$h(r)=\frac{n}{2\pi^{1/2}}\frac{\Gamma\left(\frac{n}{2}-1\right)}{\Gamma\left(\frac{n}{2}-2\right)}\frac{(1-\rho^2)^{1/2}(1-r^2)^{\frac{n}{2}-1}}{(1-\rho r)(1-2\rho r+\rho^2)^{\frac{n}{2}-1}}\{1+O(n^{-1})\}. \qquad (48.131)$$

Возможно также эквивалентное приближение

$$h(r)=\frac{\Gamma\left(\frac{N}{2}+1\right)}{\pi^{1/2}\Gamma\left(\frac{N}{2}+2\right)}\cdot\frac{(1-r^2)^{(N-1)/2}}{(1-2\rho r+\rho^2)^{N/2}}\{1+O(n^{-3/2})\}, \qquad (48.132)$$

где

$$N=n-1+\frac{\rho^2}{1-\rho^2}. \qquad (48.133)$$

48.25 Для нециклического марковского процесса с неизвестным средним

$$h(r)=\frac{\Gamma\left(\frac{N}{2}+\frac{3}{2}\right)}{2\pi^{1/2}\Gamma\left(\frac{N}{2}\right)\{N(1-\rho)-(1+\rho)\}}\times$$
$$\times\frac{(1-r)(1-r^2)^{\frac{N}{2}-1}}{(1-2\rho r+\rho^2)^{(N-1)/2}}\{1+O(n^{-3/2})\}, \qquad (48.134)$$

где r определено, как и в (48.130), а N, как и в (48.133).

48.26 Полученные результаты можно значительно улучшить. Даниэлс (1956) изучал общий циклично определенный процесс авторегрессии.

В случае процессов более высокого порядка, чем марковский, представляют интерес совместные распределения двух и более сериальных корреляций, а также частных корреляций. Впервые эти вопросы рассматривал Кенуй (1945b), затем ими занимались Дженкинс (1954b, 1956), Ватсон (1956) и Даниэлс (1956). Ради экономии места мы не будем входить в детальное обсуждение результатов. Выяснилось, что, за исключением марковского случая только циклично определенные статистики и процессы поддаются удовлетворительному изучению. Метод Даниэлса в принципе, может быть распространен на нециклический случай, но пока еще никто не нашел в себе силы взяться за труд, который с этим связан.

48.27 Распределению сериальных корреляций для случая, отличного от нормального, пока уделялось мало внимания. Выборочные эксперименты Кокса (1966) и Кенуя (1948) дали основание предполагать, что в «нулевом» случае нормальная теория может служить хорошим приближением для широкого класса распределений, даже когда $n=10$ или 20. Однако для распределений с медленно убывающими хвостами дисперсия r_1 оказывается меньше. См. 48.6 выше и работу Морэна (1967), где детально исследуются случаи экспоненциальных теоретических распределений *).

УПРАЖНЕНИЯ

48.1 Проверить равенство (48.12).

48.2 Показать, что в схеме Юла

$$u_t-1.1u_{t-1}+0.5u_{t-2}=\varepsilon_t.$$

$\mathsf{D}r_1$ приблизительно равна $2{,}44/n$.

(Бартлетт, 1946.)

48.3 Показать, что для рядов с $\rho_s=0$ при $s\geqslant 4$

$$\mathsf{M}r_j=-\frac{1}{n-j}(1+2\rho_1+2\rho_2+2\rho_3), \qquad j\geqslant 3.$$

48.4 Пусть выборочный коэффициент корреляции определен циклично и дается формулой

$$r_1=\frac{\sum_{i=1}^{n}u_iu_{i+1}-n\bar{u}^2}{\sum u_i^2-n\bar{u}^2}.$$

Показать, что в схеме Маркова

$$\mathsf{M}r_1=\rho-\frac{1+4\rho}{n}.$$

(Кендалл, 1954.)

48.5 Пусть коэффициент корреляции определен так же, как и в предыдущем примере. Показать, что для нормального случайного ряда

$$\mathsf{D}r_1=\frac{n(n-3)}{(n+1)(n-1)^2}.$$

(Морэн, 1947.)

*) Читатель может пополнить сведения по статистическим выводам в анализе сериальных корреляций, воспользовавшись, например, монографией: Э. Хениан, Многомерные временные ряды, «Мир», 1974, гл. VI. (*Прим. ред.*)

48.6 В качестве продолжения 48.6 показать, что

$$\mu_3(r_1) = \frac{4(n-4)(n-3)}{(n-1)^4(n+3)}$$

и что при цикличном определении

$$\mu_3 = \frac{2(2n-1)(n-6)}{(n-1)^3(n+1)(n+3)}.$$

(Морэн, 1947.)

48.7 Вычислить (48.9) для схемы Маркова и сопоставить результат с (48.101).

48.8 Поступая в духе 48.1, показать, что в марковском случае для больших выборок

$$\mathsf{D}v = \frac{(\varkappa_4+2)(1+\rho^2)}{n(1-\rho^2)},$$

$$\mathsf{D}c = \frac{1}{n}\left[\frac{(1+\rho^2)(1-\rho^{2j})}{1-\rho^2} + \rho^{2j}\left\{2j + \frac{(\varkappa_4+2)(1+\rho^2)}{1-\rho^2}\right\}\right],$$

$$\operatorname{cov}(c, v) = \frac{\rho^j}{n}\left[2j + \frac{(\varkappa_4+2)(1+\rho^2)}{1-\rho^2}\right].$$

(Величины c и v соответствуют r_j (см. 48.1).) Получить отсюда, что независимо от $\varkappa_4$

$$\mathsf{D}r_j = \frac{1}{n}\left[\frac{(1+\rho^2)(1-\rho^{2j})}{1-\rho^2} - 2j\rho^{2j}\right].$$

(Бартлетт, 1946.)

48.9 Аналогично (48.9) показать, что для больших выборок

$$\operatorname{cov}(r_j, r_{j+k}) = \frac{1}{n}\sum_{i=-\infty}^{\infty}\{\rho_i\rho_{i+j} + \rho_i\rho_{i+2j+k} + 4\rho_j\rho_{j+k}\rho_i^2 - 2\rho_j\rho_i\rho_{i+j+k} - 2\rho_{j+k}\rho_i\rho_{i+k}\}.$$

(Бартлетт, 1946.)

48.10. Показать (см. пример 48.4), что для больших выборок дисперсии r_j из (48.9) дается формулой

$$n\mathsf{D}r_j = \left[\left(1+2\rho_j^2\right)\text{кф}(z^0) + \text{кф}(z^{2j}) - 4\rho_j\text{кф}(z^j)\right] \quad \text{в} \quad G^2(z),$$

где $G(z)$ — производящая функция автокорреляций, а символ «кф» означает «коэффициент при».

Сопоставить формулу с результатом упражнения 48.8.

48.11 В качестве продолжения упражнения 48.4 показать, что

$$\mathsf{M}r = \rho^j - \frac{1}{n}\left\{\frac{1+\rho}{1-\rho}(1-\rho^j) + 3j\rho^j - j\rho^{n-j}\right\}.$$

(Кендалл, 1954.)

48.12 Пусть первая сериальная корреляция (при известном нулевом среднем) равна

$$r_1 = \frac{n}{n-1}\,\frac{\sum_{i=1}^{n-1} u_i u_{i+1}}{\sum_{i=1}^{n} u_i^2}.$$

Показать, что для случайного ряда r_1 и знаменатель в последней формуле независимы. Получить отсюда, что

$$\mu_1(r_1) = 0 = \mu_3(r_1),$$

$$\mu_2(r_1) = \frac{n}{(n-1)(n+2)},$$

$$\mu_4(r_1) = \frac{3n^3(n^2+4n-9)}{(n-1)^4(n+2)(n+4)(n+6)}.$$

(Морэн, 1948.)

48.13 Пусть (см. 48.16)

$$r = \operatorname{th} z, \qquad \rho = \operatorname{th}\zeta, \qquad z - \zeta = x.$$

Отталкиваясь от распределения Мэдоу — Лейпника и используя разложение, показать, что плотность распределения величины x равна

$$h(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{x^2}{2\sigma^2}\right\}X,$$

где

$$\sigma^2 = \operatorname{ch}^2\zeta/n = 1/\{n(1-\rho^2)\}$$

и

$$X = 1 - \rho x + \left\{\frac{1}{4n} - \frac{x^2(1-\rho^2)}{2} + \frac{nx^4(1-\rho^2)(1-3\rho^2)}{12} + \frac{\rho^2x^2}{2}\right\} - \rho x\left\{\frac{1}{4n} - \frac{5x^2(1-\rho^2)}{6} + \frac{nx^4(1-\rho^2)(1-3\rho^2)}{12} + \frac{\rho^2x^2}{6}\right\} + O\left(\frac{1}{n^2}\right).$$

Вычислить моменты и показать, что z распределено приближенно нормально со средним

$$\zeta - \frac{\rho}{n(1-\rho^2)} + \frac{\rho(1+\rho^2)}{n^2(1-\rho^2)^2}$$

и дисперсией

$$\frac{1}{n(1-\rho^2)} - \frac{2\rho^2}{n^2(1-\rho^2)^2}.$$

(Кенуй, 1948.)

48.14 Показать, что в схеме Юла из примера 48.2 производящая функция автокорреляций дается формулой

$$\sigma^2\sum_{-\infty}^{\infty}\rho_i z_i^i = \frac{1}{(1-1{,}1z+0{,}5z^2)(1-1{,}1z^{-1}+0{,}5z^{-2})} = 1 + \frac{0{,}7333-0{,}5z}{1-1{,}1z+0{,}5z^2}z + \frac{0{,}7333-0{,}5z^{-1}}{1-1{,}1z^{-1}+0{,}5z^{-2}}z^{-1},$$

где $\sigma^2 = \mathrm{D}u$. Возводя в квадрат и используя разложение, показать, что

$$\sum_{-\infty}^{\infty} \rho_i^2 = 2{,}44.$$

Вновь убедиться, что для больших выборок $\mathrm{D}r_j$ равна приблизительно $2{,}44/n$

(Кенуй, 1947a.)

48.15 Показать, что для общей схемы линейной авторегрессии $\sum_{j=1}^{k} \alpha_j u_{t-j} = \varepsilon_t$

$$\lim_{m\to\infty} \frac{\mathrm{D}\Delta^m u_t}{\binom{2m}{m}} = \mathrm{D}\varepsilon \left\{ \sum_{j=1}^{k} (-1)^j \alpha_j \right\}^{-2}.$$

(Муртейра, 1951.)

48.16 Заменив в схеме предыдущего примера ε на $\sum_{l=1}^{p} \beta_l \varepsilon_{t-l}$, показать, что тот же предел равен

$$\mathrm{D}\varepsilon \left\{ \frac{\sum_{i=1}^{p} (-1)^i \beta_i}{\sum_{j=1}^{k} (-1)^j \alpha_j} \right\}^2.$$

(Муртейра, 1951.)

48.17 Пусть (см. 48.16)

$$r = \sin z, \qquad \rho = \sin \zeta, \qquad z - \zeta = x.$$

Используя распределение Мэдоу — Лейпника (48.93), показать, что

$$h(x) = \sqrt{\frac{n}{2\pi}} \exp\left(-\frac{nx^2}{2}\right) \left\{ 1 - \frac{n\rho}{2(1-\rho^2)^{1/2}} x^3 + \frac{1}{4n} - \frac{1}{24} \frac{(2+13\rho^2)}{1-\rho^2} nx^4 + \right.$$
$$+ \frac{1}{8} \cdot \frac{\rho^2}{1-\rho^2} n^2 x^6 - \frac{1}{8} \frac{\rho(1+5\rho^2)}{(1-\rho^2)^{3/2}} nx^5 - \frac{1}{8} \frac{\rho}{(1-\rho^2)^{1/2}} x^3 +$$
$$\left. + \frac{1}{48} \frac{\rho(2+13\rho^2)}{(1-\rho^2)^{3/2}} n^2 x^7 - \frac{1}{48} \frac{\rho^3}{(1-\rho^2)^{3/2}} n^3 x^9 + O(n^{-2}) \right\}.$$

Вывести отсюда (48.104) и (48.105).

(Дженкинс, 1954a.)

48.18 Показать, что для случайного ряда с нулевым средним и единичной дисперсией характеристическая функция циклично определенных величин

$$p = \frac{1}{2} \sum u_i^2, \qquad q = \frac{1}{2} \sum u_i u_{i+1}, \qquad r = \frac{1}{2} \sum u_i u_{i+2}$$

равна циркулянту

$$\prod_{k=1}^{n} \left(1 - \theta_p - \theta_q \cos\frac{2\pi k}{n} - \theta_r \cos\frac{4\pi k}{n}\right)^{-1/2}.$$

Доказать, что если

$$\mu_{st} = \mathrm{M}\left(\frac{q^s r^t}{p^{s+t}}\right),$$

то

$$\mu_{10} = \mu_{11} = 0, \qquad \mu_{20} = \mu_{02} = \frac{1}{n+2},$$
$$\mu_{11} = \mu_{12} = \mu_{31} = 0,$$
$$\mu_{21} = \frac{2}{(n+2)(n+4)},$$
$$\mu_{22} = \frac{n+12}{(n+2)(n+4)(n+6)}.$$

(Дженкинс, 1954b.)

48.19 Показать, что если в статистике предыдущего примера ввести среднее $\bar{u}$ так, чтобы, например

$$p = \frac{1}{2} \sum (u_i - \bar{u})^2,$$

то характеристическая функция p, q и r окажется равной

$$\prod_{k=1}^{n-1} \left(1 - \theta_p - \theta_q \cos\frac{2\pi k}{n} - \theta_r \cos\frac{4\pi k}{n}\right)^{-1/2}.$$

Показать, что

$$\mu_{11} = -\frac{1}{n^2-1}.$$

(Дженкинс, 1954b. Там же приведены значения моментов более высоких порядков.)

48.20 Опустим в определении (48.42) статистики r_1 величину $\bar{u}$, считая, что среднее известно. Показать, что выражение для х. ф., соответствующее (48.80), остается тем же самым, с тем исключением, что множитель $(\alpha+\beta)^{1/2}$ теперь будет отсутствовать. Получить отсюда, что нечетные моменты равны нулю и приближенно

$$\mu_{2k} = \frac{1 \cdot 3 \cdot 5 \ldots (2k-1)}{(n+2)(n+4)\ldots(n+2k)}.$$

Проверить справедливость (48.85) с заменой n на $n+1$.

(Диксон, 1944.)

48.21 Пусть $x_1, \ldots, x_{2n+1}$ — независимые случайные величины, а r — сериальный коэффициент корреляции между x_{2s} и $\frac{1}{2}(x_{2s-1} + x_{2s+1})$, $s = 1, \ldots, n$. Показать, что $\mathrm{M}r = 0$, а $\mathrm{D}r = (n-1)^{-1}$.

(Морэн, 1967. Ср. 31.19 т. 2, где обсуждался обыкновенный коэффициент корреляции.)

ГЛАВА 49

СПЕКТРАЛЬНАЯ ТЕОРИЯ

Гармонический анализ

49.1 В главе 47 мы рассматривали спектр и связанные с ним функции как преобразования автоковариационной функции, но указывали, что эти функции могут появляться и в ином аспекте. А именно: спектр может служить обусловленной корреляциями характеристикой близости между временным рядом и определенными гармониками. Перейдем к более полному изучению предмета именно в этом направлении.

Изучая тепловые потоки, Фурье пришел к следующему разложению функций в ряд гармонических членов:

$$f(x)=\sum_{r=1}^{\infty} a_r \sin rx+\frac{1}{2} b_0+\sum_{r=1}^{\infty} b_r \cos rx. \tag{49.1}$$

Несмотря на то, что отдельные члены ряда периодичны, весьма широкий класс непериодических функций, рассматриваемых на конечном интервале, может быть представлен таким образом. Для справедливости такого разложения, например, достаточно, чтобы на интервале от $-\pi$ до π $f(x)$ была однозначной, непрерывной, за исключением лишь конечного числа точек разрыва, функцией и имела бы лишь конечное число точек максимума и минимума. Ряд в правой части (49.1) называется рядом Фурье. Замечательно, что члены этого ряда ортогональны. Действительно,

$$\left.\begin{aligned}\int_{-\pi}^{\pi} \cos rx \sin sx\, dx&=\begin{cases}0, & r\neq s,\\ \pi, & r=s,\end{cases}\\ \int_{-\pi}^{\pi} \sin rx \sin sx\, dx&=\begin{cases}0, & r\neq s,\\ \pi, & r=s.\end{cases}\end{aligned}\right\} \tag{49.2}$$

Следовательно, умножая (49.1) последовательно на $\sin rx$ и $\cos rx$ и интегрируя, имеем

$$a_r=\frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin rx\, dx, \tag{49.3}$$

$$b_r=\frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos rx\, dx. \tag{49.4}$$

Ряд можно также записать в форме

$$f(x)=\sum_{r=0}^{\infty} c_r \sin(rx+\varphi_r), \tag{49.5}$$

где φ_r — фазовый угол.

Так как все члены в (49.1) периодичны с периодом 2π, выражение для $f(x)$ имеет тот же период. Если $f(x)$ определена на интервале от $-L$ до L, мы можем использовать разложение с членами типа $\sin(\pi rx/L)$ и $\cos(\pi rx/L)$. Вопрос здесь, конечно, состоит лишь в изменении масштаба и переходе от интервала длины 2π к интервалу длины $2L$.

49.2 Углы измеряются, как обычно, в радианах и имеют нулевую размерность. Следовательно, величина αt в $\sin \alpha t$ имеет нулевую размерность, а α измеряется в радианах в единицу времени. Последнюю величину иногда называют *угловой частотой*. Там, где это не может вызвать недоразумения, мы будем называть ее просто частотой. Однако значение $\sin \alpha t$ повторяется через период $2\pi/\alpha$ и, следовательно, число циклов за единицу времени равно $\alpha/2\pi$. Последнее можно также рассматривать как частоту. Период $2\pi/\alpha$ имеет размерность времени (t) и называется также «длиной волны», хотя в нашем контексте «длиной» является период времени.

49.3 Итак, функцию можно разложить в ряд синусов и косинусов; последовательные члены в (49.1) будут иметь периоды 2π, $2\pi/2$, $2\pi/3$ и т. д., соответствующие угловые частоты будут равны 1, 2, 3..., а циклические частоты — $1/2\pi$, $2/2\pi$, $3/2\pi$, ... В общем случае, когда $f(x)$ определена на интервале длины $2L$, угловые частоты будут иметь вид $\pi r/L$. Таким образом, имеется одна основная частота π/L, а все остальные кратны ей. Такое представление было бы искусственным, если бы мы знали, что $f(x)$ равна сумме гармонических членов с несоизмеримыми частотами. Поэтому, вообще говоря, следует рассмотреть более общий гармонический ряд

$$f(x)=\sum_{j=0}^{\infty} a_j \sin(\alpha_j x)+\sum_{j=0}^{\infty} b_j \cos(\alpha_j x), \tag{49.6}$$

где все α_j могут принимать любые действительные значения. Простого способа вычисления a_j, b_j, аналогичного описанному в (49.3), (49.4), пока не существует. Задача оценивания a_j и b_j рассматривалась в девятнадцатом веке физиками и метеорологами; и, несмотря на обилие накопленных впоследствии фактов, методика, по существу, осталась прежней. Изменился, однако, взгляд на предмет. Ранние авторы искали скрытые гармоники. Более современный подход состоит в понимании спектра как некоторой характеристики временного ряда, независимо от того, является ли последний на самом деле суммой гармоник или нет.

Частота Найквиста и «неразличимые» частоты

49.4 При наблюдении ряда через равные единичные промежутки времени проявляются две важные особенности гармонического анализа. Ясно, что колебания с периодом, меньшим единицы, нами не будут замечены. Например, сезонные изменения останутся скрытыми, если мы будем проводить наблюдения каждое 1 января. Необходимо по крайней мере два наблюдения в год, чтобы обнаружить такую периодичность. В общем случае, если между двумя наблюдениями проходит время t_0, мы не можем выявить периоды, меньшие чем $2t_0$ или угловые частоты, большие чем π/t_0. Эта предельная величина и известна как частота Найквиста.

При определении в **47.10** спектральной плотности

$$w(\alpha) = \sum_{-\infty}^{\infty} \rho_j e^{i\alpha j} \tag{49.7}$$

подразумевалось, что временной интервал равен единице и α меняется от 0 до π. Ордината в точке π дает нам значение спектральной плотности, соответствующее частоте Найквиста.

49.5 Второй эффект, который хотелось бы отметить, также связан с интервалом между наблюдениями. Предположим, что этот интервал равен единице, и рассмотрим член $\sin(2\pi t/3)$ при $t = 1, 2, 3$, и т. д. Соответствующие значения равны $\sqrt{3}/2$, $-\sqrt{3}/2$, 0, $\sqrt{3}/2$ и т. д. Но эти значения дает нам и член $\sin(8\pi t/3)$ или $\sin(14\pi t/3)$ и т. д. Длина интервала между наблюдениями не позволяет отличить угловую частоту $2\pi/3$ от любой частоты типа $(2\pi/3) + 2\pi j$, $j = 1, 2, 3$, и т. д. Эти частоты естественно назвать *неразличимыми* *), поскольку их нельзя различить на основе анализа наблюдений. Ясно, что в указанном смысле результаты эксперимента будут безразличны к проверяемым гипотезам.

*) В оригинале указанным частотам соответствует термин «aliase». (*Прим. перев.*)

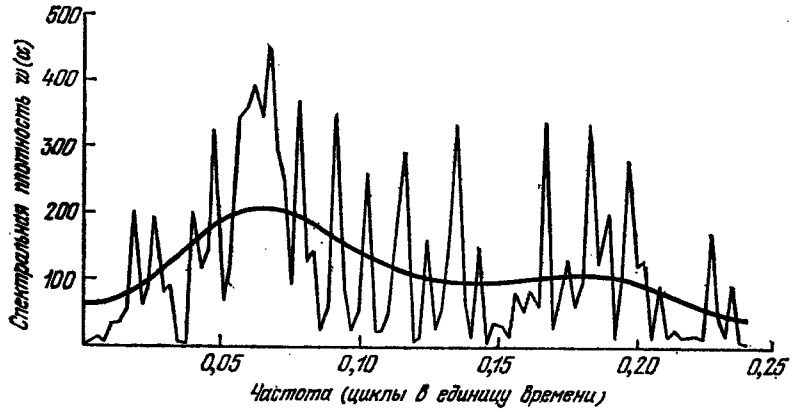

Рис. 49.1. Спектр для ряда Бевериджа индексов цен на пшеницу (таблица 47.1).

Для простоты рассмотрены частоты до 0.24. Оставшаяся часть спектра пренебрежимо мала. Сглаживание проведено с помощью ядра Парзена (упражнение 49.7). См. также рис. 49.4.

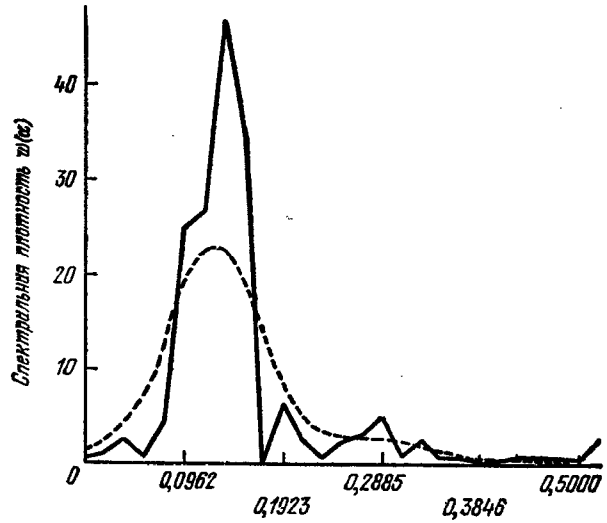

Рис. 49.2. Спектр для данных таблицы 47.2 (доли браков).

Максимум при частоте 0,1346 соответствует периоду 7,4 года. Пунктирная линия — результат сглаживания с помощью «окна» Парзена (упражнение 49.7) и первых пятнадцати сериальных корреляций.

49.6 В **47.11** мы определили функции

$$a(\alpha)=\frac{1}{\sqrt{n\pi}}\sum_{t=1}^{n}u_t\cos\alpha t, \tag{49.8}$$

$$b(\alpha)=\frac{1}{\sqrt{n\pi}}\sum_{t=1}^{n}u_t\sin\alpha t. \tag{49.9}$$

Мы показали, что интенсивность $I(\alpha)$, определенная как сумма квадратов $a(\alpha)$ и $b(\alpha)$ в пределе равна спектральной плотности $w(\alpha)$, умноженной на σ^2/π. На рис. 47.4 и 47.5 были

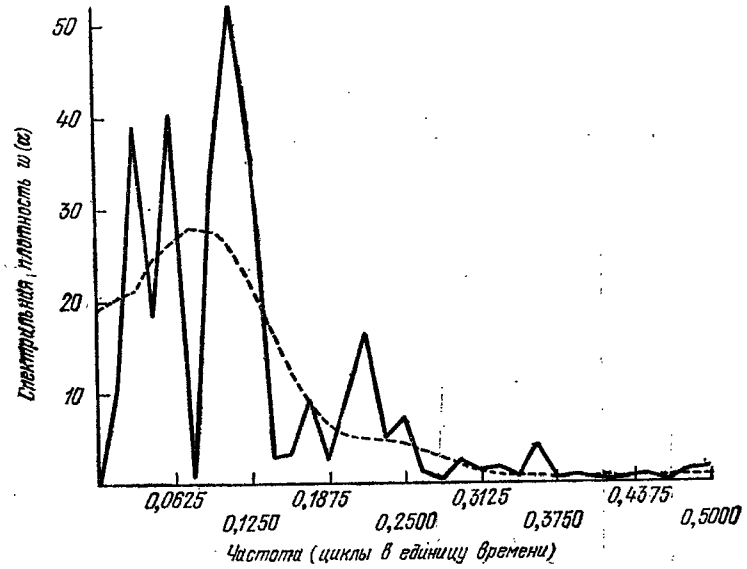

Рис. 49.3. Спектр для данных таблицы 47.4 (схема авторегрессии второго порядка).

Пунктирная линия—результат сглаживания с помощью окна Парзена (упражнение 49.7) и первых пятнадцати ковариаций.

приведены графики $w(\alpha)$ для схем Юла и Маркова. Рассмотрим еще несколько примеров. Графики спектральных функций на рис. 49.1—49.3 даны для сравнения с коррелограммами из упражнений 47.20—47.22 (ряд Бевериджа цен на пшеницу, доли браков из таблицы 47.2 и искусственный ряд Юла таблицы 47.4).

Наблюденный материал часто содержит весьма резкие флуктуации и впоследствии мы увидим, почему это так.

49.7 Иногда удобно вычислять не само $w(\alpha)$, а $\log w(\alpha)$, так как при этом большие значения интенсивности выделяются не столь резко. Кроме того, такая операция имеет то преимущество, что, как мы увидим позже, в этом случае для определенных классов оценок ширина соответствующих доверительных интервалов остается постоянной (см. **49.15**). Большей частью мы будем использовать $w(\alpha)$, но в некоторых примерах следующей главы мы будем иметь дело и с $\log w$.

Суммы $a(\alpha)$ и $b(\alpha)$ являются взвешенными суммами величин с одинаковой дисперсией, и их распределение для стационарных рядов близко к нормальному. Сначала мы рассмотрим поведение спектра при наличии гармонического члена или тренда.

49.8 Предположим, что ряд равен сумме гармонического члена с угловой частотой α и других членов, не коррелированных с первым:

$$u_t=c\sin\alpha t+g. \tag{49.10}$$

Вычисляя

$$\sum_{t=1}^{n}\sin\alpha t e^{i\beta t},$$

получаем, что это выражение равно

$$\frac{1}{\sin\left\{\frac{\alpha-\beta}{2}\right\}}\left[\cos\left\{\left(n+\frac{1}{2}\right)(\alpha-\beta)\right\}-\cos\left\{\frac{1}{2}(\alpha-\beta)\right\}-\right.$$
$$\left.-i\sin\left\{\left(n+\frac{1}{2}\right)(\alpha-\beta)\right\}+i\sin\left\{\frac{1}{2}(\alpha-\beta)\right\}\right]+$$
$$+\text{ тот же член с заменой }+\beta\text{ на }-\beta. \tag{49.11}$$

При β, близком α, первый член доминирует. Величина суммы $\sum ge^{i\beta t}$, по предположению, пренебрежимо мала. Следовательно, интенсивность $I(\beta)$ равна сумме квадратов действительной и мнимой частей (49.11), умноженной на $c^2/n\pi$. Итак,

$$I(\beta)=\frac{c^2\sin^2\left\{\frac{1}{2}n(\alpha-\beta)\right\}}{4\pi n\sin^2\left\{\frac{1}{2}(\alpha-\beta)\right\}}. \tag{49.12}$$

Ордината соответствующей периодограммы (см. (47.30)) равна

$$S^2(\beta)=\frac{c^2\sin^2\left\{\frac{1}{2}n(\alpha-\beta)\right\}}{n^2\sin^2\left\{\frac{1}{2}(\alpha-\beta)\right\}}. \tag{49.13}$$

Пусть теперь

$$\alpha - \beta = \frac{2\pi m}{n}, \quad n \text{ велико}, \quad m \text{ конечно}. \tag{49.14}$$

Тогда приближенно, но с малой ошибкой

$$S^2(\beta) = \frac{c^2 \sin^2 m\pi}{(m\pi)^2}. \tag{49.15}$$

Следовательно, при $\beta = \alpha$ периодограмма имеет пик амплитуды величиной c^2, и он будет с обеих сторон окружен меньшими пиками стремящейся к нулю интенсивности, находящимися от первого пика на расстояниях $^1/_2$, $^3/_2$, $^5/_2$, $^7/_2$ и т. д.

Сходным образом ведет себя спектральная плотность с тем исключением, что при $\beta = \alpha$ ордината становится бесконечной теоретически и очень большой на практике. Мы выбрали в (49.8) и (49.9) в качестве делителя $\sqrt{n\pi}$, потому что хотим, чтобы ордината давала нам значение спектральной плотности или соответствующей оценки. Если u — чисто случайный ряд, в пределе *)

$$I(\alpha) = \sigma^2/\pi, \tag{49.16}$$

так что ординаты спектра одинаковы для всех частот. В периодограмме ордината теоретически равнялась бы нулю.

49.9 Появление «боковых пиков» (см. (49.15)) или в спектре или в периодограмме зависит до некоторой степени от того, сколь близки частоты, для которых мы вычисляем I или S. Если в периодограмме по оси абсцисс откладывать период (что предписывает определение), «основания боковых пиков» становятся шире с ростом периода. Действительно, пусть

$$\lambda = \frac{2\pi}{\alpha}, \quad \mu = \frac{2\pi}{\beta}. \tag{49.17}$$

Тогда в силу (49.14)

$$\frac{1}{\lambda} - \frac{1}{\mu} = \frac{m}{n}$$

или приближенно

$$\mu - \lambda = \frac{m\lambda^2}{n}. \tag{49.18}$$

Следовательно, ширина оснований боковых пиков зависит от λ.

*) В общем случае $I(\alpha)$ — случайная величина (см. замечание на стр. 566). Авторы рассмотрят этот вопрос в 49.11. (*Прим. перев.*)

На рис. 49.4 для сравнения с рис. 49.1 приведена периодограмма ряда Бевериджа. Сначала Беверидж вычислял значения

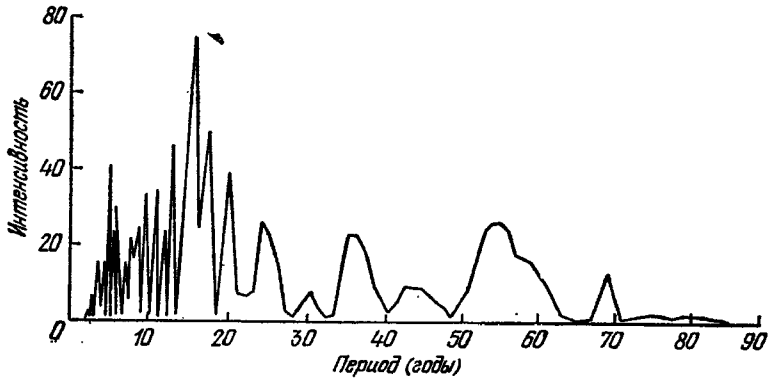

Рис. 49.4. Пернодограмма для ряда Бевериджа индексов цен на пшеницу. Дана для сравнения со спектром на рис. 49.1.

приблизительно через равные интервалы, а затем дополнил их значениями в районе точек, где обозначены пики.

Пример 49.1

Иногда замена сумм интегралами избавляет нас от утомительных вычислений и существенно облегчает получение хороших приближений. Например, при больших n

$$\sum_{t=1}^{n} \sin \alpha t \sin \beta t$$

может быть заменена на

$$\int_0^T \sin \alpha t \sin \beta t \, dt,$$

где T — длина ряда. Интеграл оказывается равным

$$\frac{1}{2}\left[\frac{\sin\{(\alpha-\beta)T\}}{\alpha-\beta} - \frac{\sin(\alpha+\beta)T}{\alpha+\beta}\right].$$

Аналогично с той же степенью точности

$$\sum_{t=1}^{n} \sin \alpha t \cos \beta t = \frac{1}{2}\left[-\frac{\cos\{(\alpha-\beta)T\}-1}{\alpha-\beta} - \frac{\cos\{(\alpha+\beta)T\}-1}{\alpha+\beta}\right].$$

Тогда интенсивность вблизи $\alpha = \beta$ равна

$$I(\beta) = \frac{\sin^2\left\{\frac{1}{2}(\alpha-\beta)T\right\}}{\pi T(\alpha-\beta)^2}. \qquad (49.19)$$

Предельный случай стремления $\alpha - \beta$ к нулю можно рассматривать так же, как и раньше.

Негармонические колебания

49.10 Нужно помнить, что пик спектра, интерпретируемый как характеристика гармонического члена, не связан с остальными пиками, только если всем пикам соответствуют синусоидальные или косинусоидальные члены. Если же имеется периодический член не гармонического характера, то в спектре ему могут соответствовать несколько пиков.

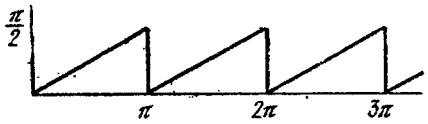

Рис. 49.5 (см. текст).

Рассмотрим в некотором смысле крайний тип колебаний, показанный на рис. 49.5.

Это в сущности периодически продолженный график $x/2$ в области $\{0 \leqslant x \leqslant \pi\}$. Рассмотрим разложение Фурье:

$$\frac{x}{2} = \sin x - \frac{1}{2}\sin 2x + \frac{1}{3}\sin 3x - \ldots, \quad 0 \leqslant x < \pi. \qquad (49.20)$$

Тогда в спектре пики интенсивности будут приходиться на частоты 1, 2, 3 и т. д. В периодограмме «интенсивности» образовали бы ряд с амплитудами, уменьшающимися пропорционально 1, 1/4, 1/9 и т. д. Таким образом, при наличии негармонического периодического элемента та часть, которую мы выделяем в качестве основной (в разложениях типа (49.10)), может оказать влияние на вид спектра вдоль всей его области определения.

Пример 49.2

Посмотрим теперь, что случится, если ряд будет содержать в качестве компоненты линейный тренд. Или, еще проще, рассмотрим чистый тренд $u_t = t$ и применим к нему спектральный анализ. Используя приближения примера 49.1, получаем

$$\int_0^T t\sin\alpha t\,dt = \left[-\frac{t\cos\alpha t}{\alpha}\right]_0^T + \int_0^T \frac{\cos s\alpha t}{\alpha}\,dt =$$

$$= -\frac{T\cos\alpha T}{\alpha} + \frac{\sin\alpha T}{\alpha^2}. \qquad (49.21)$$

Аналогично

$$\int_0^T t\cos\alpha t\,dt = \frac{T\sin\alpha T}{\alpha} + \frac{\cos\alpha T - 1}{\alpha^2}. \qquad (49.22)$$

Итак, интенсивность

$$I(\alpha) = \frac{1}{\pi T}\left\{\frac{T^2}{\alpha^2} + O(T)\right\} = \frac{T}{\pi\alpha^2} + O(1). \qquad (49.23)$$

Спектральная функция (T — постоянная) будет представлена кривой типа $y = 1/x^2$ с наибольшей интенсивностью в начале координат. С другой стороны, периодограмма дается функцией

$$S^2(\alpha) = \frac{4}{\alpha^2} = \frac{\lambda^2}{\pi^2}, \qquad (49.24)$$

где λ — длина волны.

Результаты можно истолковать, исходя из общих соображений. Тренд аналогичен очень длинной волне, что эквивалентно низкой частоте. Очевидно, что если низкие частоты находятся вне наших интересов, нам следует использовать любые попытки исключения тренда прежде, чем применять спектральный анализ.

Статистическое исследование ординаты спектра

49.11 Гармоническим составляющим исследуемого временного ряда соответствует спектральная плотность, задаваемая ординатами, восстановленными из дискретного набора точек; однако на практике эти ординаты оказываются «размазанными» в силу конечности используемой в расчетах траектории временного ряда. Рассмотрим теперь спектры недетерминированных стационарных рядов, которые могут быть представлены как взвешенные суммы (конечные или бесконечные) рядов случайных величин.

Прежде всего рассмотрим суммы $a(\alpha)$ и $b(\alpha)$ из (49.8) и (49.9) для случайного ряда u_t с нулевой автокорреляцией и дисперсией σ^2. Так как при увеличении n и $\alpha \neq 0, \pi$

$$\frac{1}{n}\sum_{k=1}^{n}\cos^2\alpha k \to \frac{1}{2}, \qquad \frac{1}{n}\sum_{k=1}^{n}\sin^2\alpha k \to \frac{1}{2}, \qquad (49.25)$$

$$\frac{1}{n}\sum_{k=1}^{n}\cos\alpha k \sum_{k=1}^{n}\sin\alpha k \to 0, \qquad (49.26)$$

то a и b в пределе являются независимыми нормальными величинами с нулевыми средними и дисперсиями, равными $\sigma^2/2\pi$.

Следовательно, $2\pi I/\sigma^2 = 2\pi(a^2+b^2)/\sigma^2$ имеет χ^2-распределение с двумя степенями свободы. Точно так же сумма S^2 в периодограмме имеет распределение

$$dF = \frac{n}{4\sigma^2}\exp\left(-\frac{nS^2}{4\sigma^2}\right)dS^2. \qquad (49.27)$$

Следовательно, асимптотически

$$\mathsf{M}I = \sigma^2/\pi, \qquad \mathsf{M}w = 1, \qquad (49.28)$$

$$\mathsf{D}I = \sigma^4/\pi^2, \qquad \mathsf{D}w = 1 = \{\mathsf{M}w\}^2. \qquad (49.29)$$

При $\alpha = 0$ и π дисперсия удваивается.

Таким образом, для случайных рядов стандартное отклонение ординаты спектра (или просто значение спектра) имеет тот же порядок, что и сама ордината.

Распределение (49.27) использовалось при построении примеров критерия для наблюденных значений ординат периодограммы. Вероятность того, что S^2 превысит величину $4\sigma^2\varkappa/n$, равна $e^{-\varkappa}$. В 1914 г. Д. Уокер показал, что если $e^{-\varkappa}$ мало, вероятность того, что m независимых ординат не превзойдут $4\sigma^2\varkappa/n$, равна $(1-e^{-\varkappa})^m$, так что вероятность того, что хотя бы одна из них превзойдет эту величину, равна $1-(1-e^{-\varkappa})^m$.

Дэвис (1941) затабулировал эту функцию. Фишер (1929), заметив, что критерий зависит от σ^2, модифицировал его в духе Стьюдента. Дэвис затабулировал также и функцию, появляющуюся в этом случае.

49.12 Замечательно, что при больших выборках (49.29) справедливо для недетерминированных стационарных рядов довольно общего типа. Этот результат, по существу, принадлежит Бартлетту (см., например, его книгу 1955 г.). Удобно объединить (49.8) и (49.9) в единую формулу

$$J(\alpha) = a(\alpha) + ib(\alpha) = \frac{1}{\sqrt{n\pi}}\sum_{t=1}^{n} u_t e^{i\alpha t}. \qquad (49.30)$$

Математическое ожидание $J(\alpha)$ равно нулю. Если u_t — случайный ряд с нулевыми автокорреляциями и дисперсией σ^2, то

$$\mathsf{M}\{J(\alpha)J(\beta)\} = \frac{\sigma^2}{n\pi}\sum_{t=1}^{n} e^{i\alpha t+i\beta t} = \frac{\sigma^2}{n\pi}\,\frac{e^{i(\alpha+\beta)}\{1-e^{in(\alpha+\beta)}\}}{1-e^{i(\alpha+\beta)}}. \qquad (49.31)$$

Последнее равно нулю, если α и β имеют вид $2\pi p/n$, где p — целое. Аналогично получаем, что $J(\alpha)$ не коррелировано с сопряженной величиной $J^*(\beta) = a(\beta) - ib(\beta)$. Итак, в этом случае $a(\alpha)$ и $b(\alpha)$ не коррелированы, а из (49.33) (см. ниже) следует, что будут асимптотически не коррелированы и соответствующие спектральные ординаты.

Поскольку $I(\alpha) = J(\alpha)J^*(\alpha)$, положив в (49.31) $\beta = -\alpha$, мы вновь получаем

$$\mathsf{M}\{I(\alpha)\} = \frac{\sigma^2}{\pi}.$$

Далее

$$\mathsf{M}\{I(\alpha)I(\beta)\} = \frac{1}{n^2\pi^2}\,\mathsf{M}\left\{\sum_{t=1}^{n} u_t e^{i\alpha t}\sum_{s=1}^{n} u_s e^{-i\alpha s}\sum_{k=1}^{n} u_k e^{i\beta k}\sum_{l=1}^{n} u_l e^{-i\beta l}\right\} =$$

$$= \frac{1}{n^2\pi^2}\sum \mathsf{M}(u_t u_s u_k u_l)\exp\{i(\alpha t - \alpha s + \beta k - \beta l)\}. \qquad (49.32)$$

Математические ожидания здесь не равны нулю, если только t, s, k, l все равны между собой (в этом случае появляется четвертый момент u), или если индексы равны попарно. Если $t = s$ и $k = l$, соответствующий член равен $\mathsf{M}I(\alpha)\cdot\mathsf{M}I(\beta)$. Отсюда получаем

$$\operatorname{cov}\{I(\alpha), I(\beta)\} =$$

$$= \frac{\varkappa_4}{n\pi^2} + \frac{\sigma^4}{n^2\pi^2}\left\{\frac{1-\cos\{n(\alpha+\beta)\}}{1-\cos(\alpha+\beta)} + \frac{1-\cos\{n(\alpha-\beta)\}}{1-\cos(\alpha-\beta)}\right\}. \qquad (49.33)$$

При $\alpha = \beta$, учитывая, что $1-\cos\theta = \theta^2/2$ при малых θ, имеем

$$\mathsf{D}I(\alpha) = \frac{\sigma^4}{\pi^2} + O(n^{-1}), \qquad (49.34)$$

что подтверждает (49.29).

Если распределение u отлично от нормального, ковариация $I(\alpha)$ и $I(\beta)$ имеет порядок n^{-1}. Если u нормально, ковариация имеет порядок n^{-2} и даже равна нулю, если α и β имеют вид $2\pi p/n$, где p целое.

49.13 Рассмотрим теперь случай, когда u_t является взвешенным средним случайных величин ε, скажем,

$$u_t = \sum_{0}^{\infty} g_s \varepsilon_{t+s}. \qquad (49.35)$$

Имеем

$$J_u(\alpha) = \frac{1}{\sqrt{n\pi}}\sum_{s=0}^{\infty}\sum_{t=1}^{n} g_s\varepsilon_{t+s}e^{i\alpha t} = \frac{1}{\sqrt{n\pi}}\sum_{s=0}^{\infty}\sum_{t=1}^{n}\varepsilon_{t+s}e^{i\alpha(t+s)}e^{-i\alpha s}g_s \doteqdot$$

$$\doteqdot \frac{1}{\sqrt{n\pi}}\sum_{k=0}^{\infty}\varepsilon_k e^{i\alpha k}\sum_{s=0}^{\infty} e^{-i\alpha s}g_s.$$

Отсюда получаем, что приближенно

$$J_u(\alpha) = J_\varepsilon(\alpha)\,h^*(\alpha), \qquad (49.36)$$

где $h(\alpha)$ — преобразование g_s, а именно

$$h(\alpha) = \sum_{s=0}^{\infty} g_s e^{i\alpha s}. \qquad (49.37)$$

Следовательно,

$$I_u(\alpha) = I_\varepsilon(\alpha)\, h(\alpha)\, h^*(\alpha). \qquad (49.38)$$

Этот результат аналогичен полученному в **47.24** результату, касающемуся передаточной функции для непрерывных рядов. Далее

$$\mathbf{M} I_u(\alpha) = h(\alpha)\, h^*(\alpha)\, \mathbf{M} I_\varepsilon(\alpha) \qquad (49.39)$$

и асимптотически

$$\mathbf{D} I_u(\alpha) = \mathbf{M}\{I_u(\alpha)\}^2. \qquad (49.40)$$

Как и в (49.29), правая часть (49.40) удваивается при $\alpha = 0$ или π.

Сглаживание спектра

49.14 Полученные результаты порождают новые задачи оценивания. Если ряд имеет длину n, дисперсия наблюденного значения спектра имеет порядок не $1/n$, а w^2.

Более того, значения спектра, соответствующие значениям $\alpha = 2\pi p/n$, не коррелированы (в точности в нормальном случае и приближенно во всех иных).

Следовательно, наблюденный спектр будет содержать резкие флуктуации (хорошей иллюстрацией служит рис. 49.1) и явится весьма ненадежной оценкой теоретического спектра.

Мы попытаемся преодолеть эту трудность, сглаживая спектр, а именно, заменяя $I(\alpha)$ взвешенной суммой соседних значений. Это приведет нас к хорошей оценке в том смысле, что последняя будет иметь небольшую дисперсию, однако в качестве платы мы должны примириться со смещенностью.

49.15 Пусть весовая функция $h(u)$ такова, что

$$h(u) = h(u + 2\pi), \qquad (49.41)$$

$$\int_{-\pi}^{\pi} h(u)\, du = 1. \qquad (49.42)$$

Такую функцию называют «ядром» или «спектральным окном». Пусть $I(\alpha)$ — оценка интенсивности. Построим сглаженную функцию

$$I_A(\alpha) = \int_{-\pi}^{\pi} h(u)\, I(\alpha - u)\, du = \int_{-\pi}^{\pi} h(\alpha - u)\, I(u)\, du. \qquad (49.43)$$

Взяв математическое ожидание, мы видим, что если $I(\alpha)$ — несмещенная оценка, то оценка $I_A(\alpha)$, будучи взвешенным средним, вообще говоря, смещена. Чтобы уменьшить смещение, можно выбрать $h(u)$, сосредоточенную в узком интервале, окружающем точку $u = 0$. В этом случае интеграл в правой части (49.43) дает нам почти несмещенную оценку. Однако потери все равно неизбежны до тех пор, пока $h(u)$ не будет равна нулю всюду, кроме точки $u = 0$. Если значения $h(u)$ сильно концентрированы, можно считать, что «эффективная» область определения функции много у́же, чем вся область определения. Такую эффективную область определения иногда называют «полосой пропускания» спектрального окна.

Интеграл (49.43) можно приблизить суммой

$$I_A(\alpha) = \frac{2\pi}{n} \sum h(u_j)\, I(\alpha - u_j), \qquad u_j = 2\pi j/n. \qquad (49.44)$$

Поскольку значения I не коррелированы,

$$\mathbf{D} I_A(\alpha) = \frac{4\pi^2}{n^2} \sum h^2(u_j)\, \mathbf{D} I(\alpha - u_j).$$

Используя (49.40), получаем, что приближенно

$$\mathbf{D} I_A(\alpha) \doteqdot \frac{4\pi^2}{n^2} \sum h^2(u_j)\, I^2(\alpha - u_j) \doteqdot \frac{2\pi}{n} \int_{-\pi}^{\pi} h^2(u)\, I^2(\alpha - u)\, du. \qquad (49.45)$$

Если $h(u)$ сконцентрировано в узкой полосе пропускания, из (49.45) получаем, что приближенно

$$\mathbf{D} I_A(\alpha) = \frac{2\pi}{n} I^2(\alpha) \int_{-\pi}^{\pi} h^2(u)\, du. \qquad (49.46)$$

Последнее вновь удваивается при $\alpha = 0, \pi$. Таким образом, если последний интеграл конечен, дисперсия имеет теперь порядок n^{-1}. Можно также показать, что с той же степенью точности $\mathbf{D} \log I_A(\alpha)$ постоянна, и, следовательно, $\log I(\alpha)$ имеет доверительные интервалы постоянной ширины. Можно также показать, что корреляция между $I_A(\alpha)$ и $I_A(\beta)$ приблизительно равна

$$\int_{-\pi}^{\pi} h(u)\, h(u + \alpha - \beta)\, du \Big/ \int_{-\pi}^{\pi} h^2(u)\, du. \qquad (49.47)$$

Поскольку последнее выражение положительно при любой приемлемой весовой функции, термин «сглаживание» вполне естествен.

Вычисление значений спектра

49.16 На практике, чтобы построить спектр, не нужно вычислять суммы (49.8) и (49.9) для различных α, а затем подсчитывать интенсивность или значение спектральной плотности. Действительно,

$$I(\alpha)=\frac{1}{n\pi}\left\{\left(\sum_1^n u_t\cos\alpha t\right)^2+\left(\sum_1^n u_t\sin\alpha t\right)^2\right\}=$$

$$=\frac{1}{n\pi}\sum_{s,\,t=1}^{n}u_s u_t(\cos\alpha t\cos\alpha s+\sin\alpha t\sin\alpha s)=\frac{2}{\pi}\sum_0^{n-1}c_k\cos k\alpha, \quad (49.48)$$

где c_k — выражение типа ковариации:

$$c_k=\frac{1}{n}\sum_{t=1}^{n-k}u_t u_{t+k}. \quad (49.49)$$

При бесконечном n это сводится к известному выражению (см. (47.27))

$$I(\alpha)=\frac{\sigma^2}{\pi}w(\alpha)=\frac{\sigma^2}{\pi}\sum_{-\infty}^{\infty}\rho_k\cos k\alpha. \quad (49.50)$$

Вычисление значений спектра обычно основывается на (49.48). Однако в любом случае мы не можем вычислить c_k при $k>n-1$; кроме того, в практических ситуациях мы вряд ли возьмем на себя труд вычислить все $n-1$ сериальных корреляций. Рассмотрим оценку

$$I_q(\alpha)=\frac{1}{\pi}\sum_{-q}^{q}\lambda_k c_k\cos k\alpha, \quad (49.51)$$

основанную на q сериальных корреляциях *). (Постоянные λ выбираются прежде всего в целях улучшения оценки, в остальном из соображений удобства).

Оценка (49.51) соответствует форме

$$I_q(\alpha)=\frac{\sigma^2}{\pi}\sum_{-q}^{q}\lambda_k\rho_k\cos k\alpha$$

или

$$w_q(\alpha)=\sum_{-q}^{q}\lambda_k\rho_k\cos k\alpha. \quad (49.52)$$

*) В (49.51) подразумевается, что $c_{-k}=c_k$. (*Прим. ред.*)

Но в силу (47.22)

$$\rho_k=\frac{1}{\pi}\int_0^{\pi}w(u)\cos ku\,du.$$

Подставляя это выражение в (49.52), получаем

$$w_q(\alpha)=\frac{1}{\pi}\sum_{-q}^{q}\lambda_k\int_0^{\pi}w(u)\cos ku\cos k\alpha\,du=$$

$$=\int_{-\pi}^{\pi}w(u)\left\{\frac{1}{2\pi}\sum_{-q}^{q}\lambda_k\cos ku\cos k\alpha\right\}du. \quad (49.53)$$

Следовательно, использование (49.51) эквивалентно сглаживанию спектра весовой функцией

$$h(\beta)=\frac{1}{2\pi}\sum_{-q}^{q}\lambda_k\cos k\beta\cos k\alpha. \quad (49.54)$$

При $\lambda_0=1$ она удовлетворяет условиям (49.41), (49.42).

Итак, мы имеем

$$\int_{-\pi}^{\pi}h^2(\beta)\,d\beta=\frac{1}{2\pi}\sum_{-q}^{q}\lambda_k^2\cos^2 k\alpha. \quad (49.55)$$

Пример 49.3

Возьмем сначала все λ равными единице. Тогда

$$h(\beta)=\frac{1}{2\pi}\sum_{-q}^{q}\cos k\beta\cos k\alpha=\frac{1}{\pi}\sum_{-q}^{q}[\cos[(\alpha+\beta)k]+\cos[(\alpha-\beta)k]]=$$

$$=\frac{1}{\pi}\left[\frac{\sin\left\{\left(q+\frac{1}{2}\right)(\alpha+\beta)\right\}}{\sin\left\{\frac{1}{2}(\alpha+\beta)\right\}}+\frac{\sin\left\{\left(q+\frac{1}{2}\right)(\alpha-\beta)\right\}}{\sin\left\{\frac{1}{2}(\alpha-\beta)\right\}}\right], \quad (49.56)$$

причем

$$\int_{-\pi}^{\pi}h^2(\beta)\,d\beta=\frac{1}{2\pi}\left[q+\frac{1}{2}+\frac{\sin\{(2q+1)\alpha\}}{2\sin\alpha}\right]. \quad (49.57)$$

Пример 49.4 (Бартлетт, 1950)

Пусть

$$\lambda_k=1-\frac{k}{q}. \quad (49.58)$$

Тогда

$$h(\beta)=\frac{1}{\pi}\left\{\frac{\sin^2\left\{\frac{1}{2}q(\alpha+\beta)\right\}}{q\sin^2\left\{\frac{1}{2}(\alpha+\beta)\right\}}+\frac{\sin^2\left\{\frac{1}{2}q(\alpha-\beta)\right\}}{q\sin^2\left\{\frac{1}{2}(\alpha-\beta)\right\}}\right\} \tag{49.59}$$

и

$$\int_{-\pi}^{\pi} h^2(\beta)\,d\beta=\frac{1}{2\pi}\left\{\frac{(q-1)(2q-1)}{6q}+\frac{1}{2q\sin^2\alpha}-\frac{\sin 2q\alpha\cos\alpha}{4q^2\sin^3\alpha}\right\}. \tag{49.60}$$

Пример 49.5 (Даниэлл, 1946)

Пусть

$$\lambda_k=\frac{\sin kh}{kh},\qquad h>0. \tag{49.61}$$

Известно, что

$$\int_0^\infty \frac{\sin px\cos qx}{x}\,dx=\begin{cases}\frac{1}{2}\pi, & |p|>|q|,\\ \frac{1}{4}\pi, & |p|=|q|,\\ 0, & |p|<|q|.\end{cases} \tag{49.62}$$

Тогда весовая функция

$$h(\beta)=\frac{1}{2\pi}\left\{1+2\sum_{k=1}^{q}[\cos\beta k\cdot\sin hk\cdot\cos\alpha k]\cdot\frac{1}{hk}\right\}.$$

Последнее выражение оценим интегралом

$$\frac{1}{2\pi}\int_0^\infty\frac{\sin hx}{hx}\left[\cos\left\{\frac{1}{2}(\alpha+\beta)x\right\}+\cos\left\{\frac{1}{2}(\alpha-\beta)x\right\}\right]dx=$$

$$=\begin{cases}\frac{1}{2h}, & h\geqslant\alpha-\beta\geqslant-h,\\ 0 & \text{в иных случаях.}\end{cases} \tag{49.63}$$

Предлагались также и многие другие ядра, в основном Блэкмэном и Тьюки (1958) и Парзеном (1961). См. упражнения 49.5—7 и обзор Дженкинса (1961).

Оценивание спектральных плотностей

49.17 Задаче оценивания спектральных плотностей уделялось большое внимание. Полное изложение всех результатов (которые сами по себе далеко не полны) заняло бы слишком много места. Удовлетворимся поэтому кратким изложением основных принципов.

Цель оценивания состоит в получении хороших оценок значений спектра вдоль всей области определения. (Обычно это не является конечной целью анализа, но на этом вопросе мы здесь останавливаться не будем.) Мы уже видели, что по разным причинам идеал недостижим, поэтому введем в рассмотрение ядро или *спектральное окно*, чтобы сгладить грубые иррегулярности наблюденного спектра. «Хорошее» ядро будет иметь относительно узкую «эффективную» область определения, но ни одно ядро не будет совершенным, и на его значения могут оказывать влияние случайные пики спектра *). В этом случае говорят об «утечке» на краях «спектрального окна», «через которое» мы просматриваем часть спектра. Таким образом, мы приходим к вопросу об эффективности сглаживания различных ядер, и, следовательно, к вопросу о надежности (которая уменьшается при усреднении), а отсюда и к смещению. В связи со сказанным могут оказаться полезными процедуры, описанные у Уиттла (1957), где рассматривается априорное распределение значений спектра, Блэкмена и Тьюки (1958), где обсуждаются вопросы использования предварительного анализа данных, и Парзена (1961), который среди других вопросов рассматривал задачу о предварительном определении скорости затухания автокорреляций.

49.18 Утечка, о которой мы говорили выше, приводит к серьезным неприятностям: среднее искажается, если значения ядра сами по себе малые в крайних точках области определения умножаются на «случайно» большие значения спектра. По этой причине Блэкмен и Тьюки (1958) предложили использовать процесс предварительного *выбеливания*, цель которого — отфильтровать ряд так, чтобы пики сгладились. Например, если первоначальный спектр имеет пик в α_1 и мы можем преобразовать первоначальный ряд так, чтобы этот пик сгладился, на оценку значения спектра в другой точке α_2 уже не будет оказывать влияния случайный пик в α_1. Очевидно, что это довольно опасная процедура, но, к счастью, мы можем после «закрасить» спектр (терминология Нерлова (1964)). Последняя идея основывается на результате **47.24**, состоящем в том, что если $v(t)$ — ряд, полученный с помощью линейной фильтрации $u(t)$, то

$$w_v(\alpha)=w_u(\alpha)\cdot|f(\alpha)|^2, \tag{49.64}$$

где $f(\alpha)$ — передаточная функция фильтра. Зная фильтр, мы всегда можем восстановить спектр первоначального ряда по оцененному спектру отфильтрованного. Процедура исследова-

*) (49.43) можно интерпретировать двояко. Первое равенство описывает воздействие ядра на значения I, второе — воздействие I на значение ядра. (*Прим. перев.*)

лась Хекстом (1964). Отметим также, что иногда полезно использовать различные процедуры для различных частей спектра.

49.19 Даниэлс (1962) использовал несколько иной подход. Сначала он сглаживал спектр некоторым ядром, удовлетворяющим заранее заданному критерию (можно, например, потребовать, чтобы оно обеспечивало некоторый минимум разрешающей способности). Затем стандартным методом проводилась операция, обратная сглаживанию и улучшающая разрешающую способность (за счет выборочной дисперсии) до уровня, на котором полезные изменения в картине спектра перестают быть ощутимыми. Предлагались два пути проведения последней процедуры. Первый основывается на локальной аппроксимации спектра полиномом, второй — на разностях спектра. Метод носит эмпирический характер в том смысле, что здесь форма оценок зависит от характера данных и требует трудоемких вычислений. Тем не менее он позволяет последовательными приближениями прийти к устойчивому решению.

49.20 В связи с равенством (49.64) можно отметить работы Хеннана (1960) и Дербина (1961), касающиеся влияния на спектр сезонных изменений и исключения тренда. Мы уже говорили о том, что исключение тренда искажает остальные компоненты. С помощью (49.64), по крайней мере теоретически, можно восстановить спектр остальных компонент, устранив искажения, связанные с исключением тренда. Это не означает конечно, что мы можем восстановить сами остаточные компоненты.

49.21 Следует всегда помнить, что периодограммы и спектр интересуют нас не сами по себе, за исключением, возможно, некоторых задач из физики и электроники, где величину спектра можно интерпретировать как мощность колебаний данной частоты. В большинстве же статистических задач поиск спектра служит «диагностическим» целям и помогает построить модель системы, порождающей наблюденный ряд. Поэтому наше внимание должно быть в большей степени сконцентрировано на проверке гипотез, касающихся моделей, а не на исследовании отдельных значений спектра или периодограммы. Указанную тему мы рассмотрим в следующей главе. Более пространное изложение спектрального анализа можно найти в книгах Блэкмэна и Тьюки (1958), Гренандера и Розенблатта (1957), Грэнджера и Хатанаки (1964), а также в докладах симпозиума по анализу временных рядов под редакцией Розенблатта (1963).

Неравные временные интервалы

49.22 В заключение мы сделаем несколько замечаний по поводу того практически важного случая, когда наблюдения проводятся ежедневно или ежемесячно. Пусть u_1, u_2 и т. д. — результаты наблюдений в моменты времени mt, иначе наша информация определяется величинами $u(m)$, $u(2m)$ и т. д. Например, мы можем проводить наблюдения ежемесячно и в этом случае m в среднем равно примерно 30. Допустим теперь, что интервал между наблюдениями в некоторых пределах меняется так, что на самом деле мы располагаем величинами $u(m+\varepsilon_1)$, $u(2m+\varepsilon_2)$ и т. д. Если u имеет нулевое среднее и единичную дисперсию, что можно предполагать, не ограничивая общности, то автокорреляции первоначального ряда

$$\rho(km) = \mathbf{M}[u(tm)\,u\{(t+k)m\}], \qquad k=0,\ 1,\ \ldots \tag{49.65}$$

Тогда автокорреляции второго ряда

$$\rho^*(km) = \mathbf{M}\frac{1}{n}\sum_{p=1}^{n} u(pm+\varepsilon_p)\,u\{(p+k)m+\varepsilon_{p+k}\} = \\ = \frac{1}{n}\sum_{p=1}^{n}\rho(km-\varepsilon_p+\varepsilon_{p+k}). \tag{49.66}$$

Разлагая ρ в ряд Тейлора, или, что эквивалентно, в ряд разностей, получаем, что с точностью до второго порядка по ε

$$\rho^*(km) = \rho(km) + \frac{1}{n}\sum(\varepsilon_{p+k}-\varepsilon_p)\,\rho'(km) + \\ + \frac{1}{2n}\sum(\varepsilon_p-\varepsilon_{p+k})^2\,\rho''(km). \tag{49.67}$$

Если ε имеет нулевое среднее, второй член мал, а в пределе стремится к нулю. Обозначая символом σ^2 дисперсию ε и τ_k — k-ю автокорреляцию ε, получаем

$$\rho^*(km) \doteqdot \rho(km) + \sigma^2(1-\tau_k)\,\rho''(km). \tag{49.68}$$

Поэтому, если σ^2 мало, автокорреляции в среднем искажаются мало.

Рассмотрим, например, наблюдения, проводимые каждое первое число месяца. Средняя продолжительность месяца в високосный год равна 30,437, соответствующая дисперсия $= 0{,}7$. Первая автокорреляция τ_1 равна $-0{,}42$. Нормируя, получаем, что $\sigma^2 = 0{,}7/(30{,}437)^2 = 0{,}0^3 76$. Из (49.68) получаем, что автокорреляции по сравнению со случаем равных временных интервалов меняются мало, если ρ'' или вторая разность ρ не очень велики, а это на самом деле так.

То же самое касается и спектра. Низкие частоты слегка меняются, но общий эффект незначителен.

49.23 Дело обстоит не так, если мы рассматриваем агрегированные величины, например осадки. Здесь эффект, связанный с

различием длин интервалов, может быть значительным. Это становится очевидным, если вспомнить, что мы можем сравнивать суммы, основанные как на 28, так и на 31 наблюдении (в случае среднемесячных данных). По нашему мнению, в таких случаях нужно стандартизировать данные, переходя к равным интервалам. Это, в частности, верно для таких данных, как выпуск продукции за рабочую неделю, или потребление энергии за рабочий месяц.

Грэнджер (1963) детально обсуждал этот вопрос с точки зрения спектральной теории. См. также работу Кенуя (1958).

УПРАЖНЕНИЯ

49.1 $P(t)$ — полином степени k $(0 \leqslant t \leqslant T)$. Показать, что ордината периодограммы, соответствующая частоте α, равна

$$\frac{4P^2(T)}{\alpha^2 T^2} + O\left(T^{2k-3}\right).$$

49.2 Члены ряда принимают значения e^{ct}, $c > 0$, $0 \leqslant t \leqslant T$. Показать, что ордината периодограммы, соответствующая частоте α, равна

$$\frac{16 \exp(2cT)}{T^2} \sin^2 \frac{1}{2}\alpha T.$$

49.3 Известно, что

$$x = \frac{4}{\pi}\left(\sin x - \frac{\sin 3x}{3^2} + \frac{\sin 5x}{5^2} - \ldots\right), \qquad -\frac{1}{2}\pi \leqslant x \leqslant \frac{1}{2}\pi.$$

Начертить график ряда, член которого равен x на интервале $(0, 4\pi)$. Сравнить с 49.10 и прокомментировать появляющийся спектр.

49.4 Получить (49.47).

49.5 Положив в (49.54)

$$\lambda_k = 1 - 2a + 2a\cos(\pi k/q),$$

показать, что h равно сумме двух членов типа

$$\frac{1}{\pi}\left[(1-2a)\frac{\sin\left(q+\frac{1}{2}\right)\gamma}{\sin\frac{1}{2}\gamma} + a\left\{\frac{\sin\left\{\left(q+\frac{1}{2}\right)\gamma+\pi/q\right\}}{\sin\left(\frac{1}{2}\gamma+\pi/q\right)} + \frac{\sin\left\{\left(q+\frac{1}{2}\right)\gamma-\pi/q\right\}}{\sin\left(\frac{1}{2}\gamma-\pi/q\right)}\right\}\right],$$

где в одном случае $\gamma = \alpha - \beta$, а в другом $\gamma = \alpha + \beta$.

(Тьюки, см. Блэкмэн и Тьюки, 1958. Предлагались значения $a = 0{,}25$ или $a = 0{,}23$.)

49.6 Положив в предыдущем примере

$$\lambda_k = 1 - k^2/q^2,$$

получить выражение

$$\frac{1}{\pi q^2}\left\{\frac{\frac{1}{2}\sin q\gamma\cos\frac{1}{2}\gamma}{\sin^3\frac{1}{2}\gamma} - \frac{q\cos q\gamma}{\sin^2\frac{1}{2}\gamma}\right\}.$$

(Парзен, 1961.)

49.7 Положив в предыдущем примере

$$\lambda_k = \begin{cases} 1 - 6\left(\frac{k}{q}\right)^2 + 6\left(\frac{k}{q}\right)^3, & 0 \leqslant k \leqslant \frac{q}{2}, \\ 2\left(1 - \frac{k}{q}\right)^3, & \frac{q}{2} \leqslant k \leqslant q, \end{cases}$$

получить выражение

$$\frac{3}{4\pi q^3}\left\{\frac{\sin\frac{1}{4}q\gamma}{\sin\frac{1}{4}\gamma}\right\}^4.$$

(Парзен, 1961.)

49.8 u_t — стационарный нормальный ряд, $\gamma_1 = \mu_3/\mu_2^{3/2}$, $\gamma_2 = \left(\mu_4/\mu_2^2\right) - 3$. Показать, что асимптотически γ_1 распределено как $N(0, R_1)$ с

$$R_1^2 = 6\sum_{-\infty}^{\infty}\rho_j^3,$$

и γ_2 распределено как $N(0, R_2)$ с

$$R_2^2 = 24\sum_{-\infty}^{\infty}\rho_j^4.$$

Показать, как использовать это для проверки нормальности стационарного процесса.

(Ломницки, 1961.)

49.9 Ряд из $\rho\mu$ членов расположен в таблицу

u_1	u_2	$\ldots$	u_μ
$u_{\mu+1}$	$u_{\mu+2}$	$\ldots$	$u_{2\mu}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$
$u_{(\rho-1)\mu+1}$	$u_{(\rho-1)\mu+2}$	$\ldots$	$u_{\rho\mu}$
Суммы: m_1	m_2	$\ldots$	m_μ

Показать, что

$$A = \frac{2}{\rho\mu}\sum_{j=1}^{\mu} m_j \cos\frac{2\pi j}{\mu},$$

$$B = \frac{2}{\rho\mu}\sum_{j=1}^{\mu} m_j \sin\frac{2\pi j}{\mu},$$

где A и B — суммы, входящие в периодограмму.

49.10 Продолжая предыдущий пример, рассмотреть

$$\eta^2(\mu) = \mathrm{D}m/\mathrm{D}u$$

и показать, что если

$$u_j = a\sin(2\pi j/\lambda) + b_j,$$

где b_j и периодический член не коррелированы, то

$$\eta^2(\mu) = \left(\frac{a^2\mu^2\sin^2(n\pi/\lambda)}{2n^2\sin^2(\mu\pi/\lambda)} + \frac{\mu}{n}\mathrm{D}b\right)\Big/\left(\frac{1}{2}a^2 + \mathrm{D}b\right).$$

Вывести отсюда, что в окрестности λ график зависимости η от μ (периодограмма Уиттекера) имеет пик ширины $2\lambda^2/n$, окруженный меньшими пиками.

(Уиттекер, 1911.)

49.11 Значения ряда авторегрессии (47.74) типа Юла наблюдаются с ошибками, причем ошибки, соответствующие различным членам ряда, неза-

Спектральные плотности (таблица к упражнению 49.12)

Частота (число циклов за год)	Без сглаживания	$q=15$	$q=20$	$q=30$
0	0,0000	19,3065	15,5732	11,6349
0,0157	10,4916	20,0120	18,2743	15,8810
0,0313	39,5419	21,8959	21,9328	24,0449
0,0469	18,6496	24,3610	25,1102	28,3488
0,0625	40,5257	26,6535	27,2150	26,1922
0,0781	1,1554	28,0290	29,1099	25,9544
0,0938	34,9011	27,8621	30,4033	33,5149
0,1094	52,4760	25,8092	28,9764	35,8784
0,1250	36,6309	22,0280	23,7428	25,2362
0,1406	2,7163	17,2623	16,5376	13,2443
0,1563	3,7282	12,6006	10,4166	7,8444
0,1719	9,6669	8,9806	6,9051	5,9028
0,1875	2,5688	6,7729	5,5586	4,6738
0,2031	0,0968	5,7301	5,3329	4,8532
0,2188	16,3082	5,2859	5,4431	5,9667
0,2344	5,5837	4,9314	5,3382	6,0461
0,2500	7,7572	4,4191	4,7363	4,9161
0,2656	1,3151	3,7474	3,7738	3,5804
0,2813	0,4748	3,0386	2,8106	2,5929
0,2969	2,4491	2,4260	2,0975	1,8886
0,3125	0,9332	1,9872	1,7045	1,3929
0,3281	1,9702	1,7203	1,5886	1,3985
0,3438	0,7401	1,5573	1,6005	1,7491
0,3594	4,0405	1,4114	1,5402	1,8142
0,3750	0,3481	1,2300	1,3106	1,3800
0,3906	0,4831	1,0181	0,9851	0,8537
0,4063	0,5796	0,8206	0,7045	0,5774
0,4219	0,1827	0,6847	0,5519	0,4695
0,4375	0,4529	0,6311	0,5288	0,4463
0,4531	0,4774	0,6486	0,5974	0,5478
0,4688	0,1512	0,7041	0,7097	0,7256
0,4844	1,1829	0,7582	0,8112	0,8687
0,5000	1,8403	0,7801	0,8521	0,9190

висимы. Показать, что сериальные корреляции (за исключением r_0) уменьшаются в c раз, где c — постоянная. Пусть a' и b' — полученные из уравнений Юла — Уокера (47.66) оценки величин α и β в уравнении

$$u_{t+2} + \alpha u_{t+1} + \beta u_t = \varepsilon_{t+2}.$$

Показать, что для ряда, наблюдаемого с ошибками,

$$b'/a' > b/a,$$

где a и b — оценки для того же ряда, наблюдаемого без ошибок.

49.12 Нами приведены значения спектральной плотности ряда таблицы 47.4 и рис. 49.3, а также результаты сглаживания с помощью ядра Парзена (упражнение 49.7) и с использованием q сериальных корреляций (см. (49.52)—(49.54)). Начертить график спектра и отметить эффект возмущения при слишком больших q.

ГЛАВА 50

ВРЕМЕННЫЕ РЯДЫ: НЕКОТОРЫЕ ДОПОЛНИТЕЛЬНЫЕ АСПЕКТЫ

50.1 Теория временных рядов не достигла и, по-видимому, никогда не достигнет той стадии развития, когда становится возможным ясное, структурное описание предмета. До некоторой степени это обусловлено сложной природой изучаемого объекта: мы вынуждены описывать не только распределения, но и зависимость автокорреляции от времени. Это приводит к обескураживающему обилию параметров и большому количеству различных согласующихся с данными гипотез. Выбрать среди последних одну порой весьма затруднительно. В некоторых областях, особенно в экономике, накопленный опыт не столь велик, чтобы мы могли полагаться на наши модели с той же уверенностью, с какой это делается, например, в физике. Ряд данных за 50 лет считается достаточно большим для рядов экономических величин, но, будь он и длиннее, у нас нет гарантии, что за этот период сама система не претерпела важные структурные изменения.

50.2 Появление электронно-вычислительных машин избавило исследователей от утомительных вычислений, которые раньше служили серьезным препятствием к исследованию временных рядов. Однако остались трудности, связанные с формулировкой и проверкой гипотез, построением моделей изучаемых систем. По этой причине в практической статистике, для того чтобы строго поставить задачу и построить модель, очень часто прибегают к помощи посторонней нестатистической информации качественного характера. Мы не будем здесь касаться соображений и методов, которыми пользуются на этом этапе работы. Примем лишь во внимание, что они существуют, а в этой последней главе рассмотрим чисто статистические аспекты: оценивание и проверку гипотез, многомерные обобщения и несколько связанных между собой вопросов, касающихся идентификации и смешанных регрессионно-авторегрессионных систем.

Оценивание

50.3 Начнем с того, что подчеркнем некоторые общие моменты, которые в той или иной форме свойственны анализу временных рядов и являются камнем преткновения на пути получения точных результатов в оценивании и проверке гипотез.

Рассмотрим функцию правдоподобия авторегрессионного ряда. Ради наглядности ограничимся схемой Маркова, хотя наши рассуждения будут носить общий характер. Пусть $u_1, \ldots, u_n$ — последовательность наблюдений и

$$\begin{aligned} u_1 &= \rho u_0 + \varepsilon_1, \\ &\cdots\cdots\cdots \\ u_n &= \rho u_{n-1} + \varepsilon_n. \end{aligned} \tag{50.1}$$

Если бы распределение величин ε было известно, мы могли бы считать, что соотношения (50.1) определяют преобразование случайных переменных, приводящее к новым величинам $u_1, \ldots, u_n$. Но здесь мы сталкиваемся с той трудностью, что соотношения включают также u_0. В самом деле, имеется n величин ε и $n+1$ величин u. Добавим к (50.1) дополнительное равенство

$$u_0 = u_0. \tag{50.2}$$

Мы знаем, что u_0 зависит только от ε_0, ε_{-1} и т. д. и, следовательно, не зависит от $\varepsilon_1, \ldots, \varepsilon_n$. Если плотность u_0 равна $g(u_0)$, а совместная плотность $\varepsilon_1, \ldots, \varepsilon_n$ равна $f(\varepsilon_1, \ldots, \varepsilon_n)$, то величины $u_0, \varepsilon_1, \ldots, \varepsilon_n$ имеют совместное распределение

$$dF = f(\varepsilon_1, \ldots, \varepsilon_n)\, g(u_0)\, d\varepsilon_1 \ldots d\varepsilon_n\, du_0. \tag{50.3}$$

Осуществим теперь преобразование к величинам $u_0, u_1, \ldots, u_n$. Легко видеть, что якобиан преобразования равен единице, и, следовательно,

$$dF = f(u_1 - \rho u_0,\ u_2 - \rho u_1,\ \ldots,\ u_n - \rho u_{n-1})\, g(u_0)\, du_0\, du_1 \ldots du_n. \tag{50.4}$$

Использовать это выражение при построении оценок или критериев нам мешает то, что элемент u_0 неизвестен. (Если бы он был известен, мы бы начали с него строить ряд наблюдений.) Есть несколько способов преодоления указанной трудности, но все они сопряжены с определенными ограничениями в окончательных выводах, а именно:

(а) с ограничением, что u_0 известно, при этом последующие результаты поставлены в зависимость от его значения;

(б) с ограничением, что выборка циклична, т. е. что $u_n = u_0$;

(в) с предположением, что можно использовать асимптотическую пренебрежимость величины u_0.

50.4 Выберем третий путь. Предположим, например, что все ε имеют нормальное распределение с единичной дисперсией. Тогда u_0 нормально с дисперсией $1/(1-\rho^2)$ и с точностью до

постоянной

$$\log L = \frac{1}{2}\log(1-\rho^2) - \frac{1}{2}\sum_{j=1}^{n}(u_j - \rho u_{j-1})^2 - \frac{1}{2}(1-\rho^2)u_0^2. \qquad (50.5)$$

Член, включающий u_0, равен $-\frac{1}{2}u_0^2 + \rho u_0 u_1$. Интегрируя, получаем, что

$$\log L = \text{const} + \frac{1}{2}\log(1-\rho^2) - \frac{1}{2}\sum_{j=2}^{n}(u_j - \rho u_{j-1})^2 - \frac{1}{2}(1-\rho^2)u_1^2. \qquad (50.6)$$

При больших n сумма в (50.6) доминирует и асимптотически

$$\log L \sim -\frac{1}{2}\sum_{j=2}^{n}(u_j - \rho u_{j-1})^2. \qquad (50.7)$$

Мы можем оценить ρ, максимизируя функцию правдоподобия, что эквивалентно минимизации суммы квадратов $n-1$ членов.

Без учета погрешности аппроксимации результаты здесь такие же, какие мы получили бы, рассматривая авторегрессию как обычную регрессию. Поэтому оценка МП равна

$$\hat{\rho} = \sum_{2}^{n} u_j u_{j-1} \Big/ \sum_{2}^{n} u_{j-1}^2.$$

50.5 Тот же результат можно получить иным способом. Дисперсия u_i равна $(1-\rho^2)^{-1}$, а корреляция между u_i и u_j есть $\rho^{|i-j|}$. Поэтому матрица рассеяния $u_1, \ldots, u_n$ равна

$$V = \frac{1}{1-\rho^2}\begin{pmatrix} 1 & \rho & \rho^2 & \ldots & \rho^{n-1} \\ \rho & 1 & \rho & \ldots & \rho^{n-2} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \ldots & 1 \end{pmatrix}. \qquad (50.8)$$

Как нетрудно проверить, определитель равен $(1-\rho^2)^{-1}$, а обратная матрица равна

$$V^{-1} = \begin{pmatrix} 1 & -\rho & 0 & 0 & \ldots & 0 \\ -\rho & 1+\rho^2 & -\rho & 0 & \ldots & 0 \\ 0 & -\rho & 1+\rho^2 & -\rho & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \ldots & 1 \end{pmatrix}. \qquad (50.9)$$

Следовательно, в нормальном случае, с точностью до постоянных, логарифм функции правдоподобия равен

$$\frac{1}{2}\log(1-\rho^2) - \frac{1}{2}\left\{u_1^2 + \sum_{j=2}^{n-1}(1+\rho^2)u_j^2 + u_n^2 - 2\rho\sum_{j=2}^{n}u_j u_{j-1}\right\} =$$

$$= \frac{1}{2}\log(1-\rho^2) - \frac{1}{2}\left\{\sum_{j=2}^{n}(u_j - \rho u_{j-1})^2 + (1-\rho^2)u_1^2\right\}, \qquad (50.10)$$

что снова приводит нас к (50.6).

50.6 Второе обстоятельство, которое следует отметить, касается связи между авторегрессией и скользящими средними. Мы уже отмечали в **47.18**, что схема авторегрессии эквивалентна скользящему усреднению с бесконечным отрезком усреднения. Поэтому может показаться, что провести оценивание в случае скользящих средних довольно просто, если отрезок усреднения в действительности конечен. Оказывается, что это не так. Проиллюстрируем это на схеме

$$u_t = \varepsilon_t + \beta\varepsilon_{t-1}. \qquad (50.11)$$

Если дисперсия ε равна единице, матрица рассеяния ряда равна

$$\begin{pmatrix} 1+\beta^2 & \beta & 0 & 0 & \ldots & 0 \\ \beta & 1+\beta^2 & \beta & 0 & \ldots & 0 \\ 0 & \beta & 1+\beta^2 & \beta & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \ldots & 1+\beta^2 \end{pmatrix}. \qquad (50.12)$$

При $\beta = -\rho$ (50.12) почти совпадает с (50.9), но на самом деле различие существенно, и (50.12) не так легко обратить. Рассмотрим тем не менее (50.8) с $\beta = -\rho$, т. е. матрицу

$$\frac{1}{1-\beta^2}\begin{pmatrix} 1 & -\beta & \beta^2 & \ldots & (-\beta)^{n-1} \\ -\beta & 1 & -\beta & \ldots & (-\beta)^{n-2} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ (-\beta)^{n-1} & (-\beta)^{n-2} & (-\beta)^{n-3} & \ldots & 1 \end{pmatrix} \qquad (50.13)$$

Слегка видоизменим модель (50.11) так, чтобы ε_0 равнялось нулю. Тогда $\mathrm{D}u_1 = 1$. Точно также, если $\mathrm{D}\varepsilon_n = 1-\beta^2$, то $\mathrm{D}u_n = 1$. Другие значения не затрагиваются, и, следовательно, (50.13) является обращением матрицы рассеяния модифицированной схемы. Ясно, что асимптотически последняя совпадает со схемой (50.11), поскольку были изменены лишь два крайних члена.

Итак, с указанной степенью точности логарифм функции правдоподобия равен

$$\log L = \text{const} - \frac{1}{2}\log(1-\beta^2) - \\ - \frac{1}{2(1-\beta^2)}\left\{\sum_{j=1}^{n} u_j^2 - 2\beta\sum_{1}^{n-1} u_j u_{j+1} + 2\beta^2\sum_{1}^{n-2} u_j u_{j+2} - \ldots\right\}. \quad (50.14)$$

Это выражение включает все сериальные корреляции, тогда как в схеме авторегрессии мы нуждались в стольких сериальных корреляциях, сколько постоянных нам надо было оценить. Даже если пренебречь невыписанными в (50.14) членами, выражение все еще остается громоздким, и, в частности, уравнения МП окажутся практически не поддающимися решению.

Пример 50.1

Можно предположить, что указанное препятствие преодолимо, если использовать какую-нибудь иную оценку. Например, первая автокорреляция схемы (50.11) равна

$$\rho_1 = \beta/(1+\beta^2). \quad (50.15)$$

Тогда внушает доверие оценка b, являющаяся решением уравнения

$$b/(1+b^2) = r_1. \quad (50.16)$$

К сожалению, как показал Уиттл (1953а), это очень неэффективная оценка.

Действительно, для больших выборок (см. (48.9))

$$n\,\mathsf{D}r_1 = \sum_{i=-\infty}^{\infty}\{\rho_i^2 + \rho_{i-1}\rho_{i+1} - 4\rho_1\rho_i\rho_{i+1} + 2\rho_i^2\rho_1^2\},$$

что в нашем случае сводится к равенству

$$n\,\mathsf{D}r_1 = 1 - 3\left(\frac{\beta}{1+\beta^2}\right)^2 + 4\left(\frac{\beta}{1+\beta^2}\right)^4. \quad (50.17)$$

Из (50.16) получаем

$$dr_1 = \frac{1-b^2}{(1+b^2)^2}\,db, \quad (50.18)$$

и, следовательно, асимптотически

$$n\,\mathsf{D}b = \frac{1}{(1-\beta^2)^2}\{(1+\beta^2)^4 - 3\beta^2(1+\beta^2) + 4\beta^4\}. \quad (50.19)$$

Например, при $\beta = 0{,}5$

$$n\,\mathsf{D}b = 389/144. \quad (50.20)$$

Рассмотрим $\log L$ в форме

$$-\frac{1}{2(1-\beta^2)}\left\{\sum u_j^2 - 2\beta\sum u_j u_{j+1} + 2\beta^2\sum u_j u_{j+2} - \ldots\right\} = \\ = -\frac{1}{2(1-\beta^2)}A.$$

Имеем

$$\mathsf{M}A = 1-\beta^2,$$

$$\mathsf{M}\left(\frac{\partial A}{\partial\beta}\right) = -2\beta,$$

$$\mathsf{M}\left(\frac{\partial^2 A}{\partial\beta^2}\right) = 0,$$

$$\frac{\partial^2}{\partial\beta^2}\log L = -\frac{1}{2}\left[\left\{\frac{2}{(1-\beta^2)^2} + \frac{8\beta^2}{(1-\beta^2)^3}\right\}A + \\ + \frac{4\beta}{(1-\beta^2)^2}\frac{\partial A}{\partial\beta} + \frac{1}{1-\beta^2}\frac{\partial^2 A}{\partial\beta^2}\right],$$

откуда $\mathsf{M}\left(\frac{\partial^2}{\partial\beta^2}\log L\right) = -\frac{1}{1-\beta^2}$.

Таким образом, дисперсия оценки МП (ср. (18.60))

$$\mathsf{D}\hat{\beta} = \frac{1-\beta^2}{n}. \quad (50.21)$$

При $\beta = 1/2$ это равно $3/4n$, и сравнение с (50.20) показывает, что дисперсия оценки b из (50.16) в 3,6 раза больше оптимального значения.

Это неожиданно, но легко понять. Оценка, определяемая из (50.16), использует лишь первую сериальную корреляцию и игнорирует информацию, содержащуюся в остальных корреляциях, хотя, как мы видели, все они входят в функцию правдоподобия.

Оценивание в теории авторегрессионных рядов

50.7 Рассмотрим общую схему линейной авторегрессии:

$$\sum_{j=0}^{k}\alpha_j u_{t-j} = \varepsilon_t. \quad (50.22)$$

Рассуждения, аналогичные рассуждениям **50.4**, приводят нас к тому, что если пренебречь краевыми эффектами, в нормальном случае оценки МП можно получить, минимизируя

$$\sum_{t=k+1}^{n}\left\{\sum_{j=0}^{k}\alpha_j u_{t-j}\right\}^2.$$

Последнее приводит нас к уравнениям Юла — Уокера (47.66),

Иначе говоря, решать задачу оцениваиия можно так же, как и в случае обычной регрессии.

Основная теорема здесь принадлежит Манну и Вальду (1943), которые строго доказали, что асимптотические оценки $\hat{a}$ величин α по методу наименьших квадратов имеют те же выборочные свойства, что и регрессионные оценки по методу наименьших квадратов для многомерного нормального случая.

Этого полезного результата достаточно для большинства практических целей. Экспериментальное изучение рядов умеренной длины ($n \approx 60$) при равномерном распределении величин ε показало, что уравнения Юла — Уокера в этом случае довольно надежны, хотя лучше выправлять смещение оценок автокорреляций методом Кенуя (см. **48.4**).

Таблица 50.1

Остаточные значения ряда таблицы 45.4 (поголовье овец), полученные после исключения тренда простым скользящим усреднением по девяти точкам

Год	Остаток (10 000)	Год	Остаток (10 000)	Год	Остаток (10 000)	Год	Остаток (10 000)	Год	Остаток (10 000)
1871	−176	1884	+38	1897	−23	1910	+119	1923	−109
72	−112	85	+97	98	+17	11	+66	24	−23
73	+50	86	+8	99	+71	12	−52	25	+60
74	+141	87	−5	1900	+35	13	−117	26	+121
75	+60	88	−105	01	+16	14	−61	27	+94
76	−20	89	−99	02	−27	15	+19	28	−25
77	+12	90	+35	03	−32	16	+128	29	−90
78	+82	91	+159	04	−49	17	+97	30	−75
79	+130	92	+167	05	−61	18	+69	31	+72
80	−14	93	+34	06	−52	19	−29	32	+152
81	−166	94	−103	07	−24	20	−174	33	+112
82	−179	95	−104	08	+68	21	−107	34	−64
83	−84	96	−15	09	+141	22	−142	35	−87

50.8 Наиболее часто оказывается желательным проверить гипотезы, касающиеся порядка авторегрессионных схем. Иначе говоря, наиболее часто «авторегрессионность» ряда предполагают заранее и проверяют, насколько далеко простирается регрессия. Из вышесказанного ясно, что мы можем увеличивать порядок авторегрессии (вводя дополнительные члены) до тех пор, пока уменьшение суммы квадратов перестанет быть ощутимым. Фактически исследование в этом направлении авторегрессии проще, чем обычной регрессии. Объясняется это тем, что здесь мы не сталкиваемся с обычными трудностями, связанными с исключением «незначимых» переменных в том случае, когда независимые переменные в уравнении регрессии представляют смешанный тип.

Пример 50.2

В таблице 45.4 были приведены данные о поголовье овец в Англии и Уэльсе с 1867 по 1939 гг. Рис. 49.4 указывает на приблизительно линейный отрицательный тренд. В таблице 50.1 мы приводим «остатки», полученные после исключения тренда простым скользящим усреднением по девяти точкам. Нам следует рассмотреть, в какой мере остаточный ряд может быть представлен в виде

$$u_t = f(u_{t-1}, u_{t-2}, \ldots) + \varepsilon_t. \tag{50.23}$$

Первые десять сериальных корреляций таковы:

Порядок корреляции k	r_k	Порядок корреляции k	r_k
1	0,595	6	0,144
2	−0,151	7	0,203
3	−0,601	8	0,118
4	−0,537	9	0,006
5	−0,138	10	−0,078

Прежде всего рассмотрим, какой порядок имела бы схема линейной авторегрессии. Наиболее легко это сделать, используя частные корреляции между u_t и u_{t-k} при исключении всех промежуточных наблюдений и множественные корреляции, определенные в (27.61). Получаем следующие данные:

Запаздывание (лаг) k	Значение частной корреляции r с запаздыванием k	$\prod_1^k (1-r^2) = 1-R^2$
1	0,595	0,6460
2	−0,782	0,2509
3	+0,097	0,2485
4	−0,183	0,2402
5	0,031	0,2400
6	0,014	0,2400

Очевидно, что мы не получим ощутимого выигрыша, выбрав порядок линейной авторегрессии больший чем два. Отметим, что большим значениям $|r_3|$ и $|r_4|$ соответствуют малые значения частных корреляций, в то время как малому значению $|r_2|$ соответствует большое значение частной корреляции.

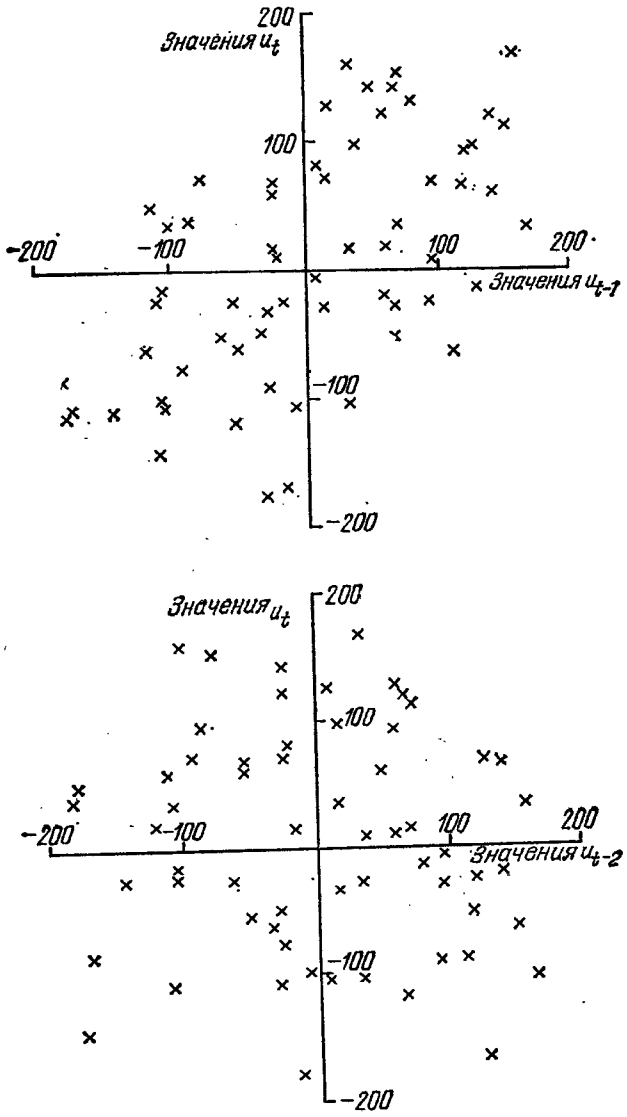

Рис. 50.1. Диаграммы рассеяния u_t и u_{t-1} (вверху) и u_t и u_{t-2} (внизу) для данных примера 50.2 (поголовье овец).

Теперь было бы желательно проверить, насколько улучшит согласие нашей модели с реальными данными введение в модель «криволинейных» членов (при существенном улучшении согласия мы примирились бы даже с нестационарностью модели). Наиболее наглядно на этот вопрос можно ответить с помощью диаграмм рассеяния u_t относительно u_{t-1} и u_t относительно u_{t-2}. Эти диаграммы приведены на рис. 50.1. Никакой криволинейности (по крайней мере, на глаз) не видно, и мы приходим к выводу, что если предположение об авторегрессии правомерно, то мы имеем дело со схемой Юла

$$u_t = -\alpha_1 u_{t-1} - \alpha_2 u_{t-2} + \varepsilon_t, \tag{50.24}$$

где дисперсия ε составляет около 25 процентов (значение $1 - R^2$ выше — см. (27.56)) от дисперсии u.

Постоянные α_1 и α_2 легко оценить, используя (47.77—8):

$$-\alpha_1 = \frac{r_1(1-r_2)}{1-r_1^2} = 1{,}060, \quad -\alpha_2 = \frac{r_2-1}{1-r_1^2} + 1 = -0{,}782.$$

Окончательно уравнение авторегрессии имеет вид

$$u_t = 1{,}060 u_{t-1} - 0{,}782 u_{t-2} + \varepsilon_t. \tag{50.25}$$

Критерий согласия для схем авторегрессии

50.9 Оказывается, что в схеме авторегрессии частные корреляции можно получить непосредственно. Этот факт использовал Кенуй (1947b) при построении одного остроумного критерия согласия (адекватности).

Аналогично (50.22) определим величины η_t равенством

$$\eta_t = \sum_0^k \alpha_j u_{t+j}, \tag{50.26}$$

где, образно говоря, величины u «идут во времени» не назад, а вперед.

Имеем

$$\operatorname{cov}(u_t \eta_{t+l}) = \mathbf{M}\left(\sum \alpha_i u_{t+i}\right)\left(\sum \alpha_j u_{t+j+l}\right) =$$
$$= \sum_{i,\,j=0}^{k} \alpha_i \alpha_j \gamma_{|l+j-i|} = \sum_{j=0}^{k} \alpha_j \sum_{i=0}^{k} \alpha_i \gamma_{|j+l-i|}, \tag{50.27}$$

где γ_p — p-я автоковариация u_t.

При $l > 0$ вторая сумма справа в силу уравнений Юла — Уокера (47.66) равна нулю. То же верно и при $l < 0$. При $l = 0$ получаем

$$\mathbf{D}\eta_t = \mathbf{D}\varepsilon_t. \tag{50.28}$$

Таким образом, если ε — нормальная случайная величина, то η — также нормальна и имеет ту же дисперсию. η_t зависит от ε_{t+k} и предыдущих величин ε и, следовательно, не зависит от ε_{t+k+l} при $l > 0$.

Рассмотрим теперь величины

$$q_j = \frac{1}{n}\sum_{t=1}^{n}(\varepsilon_{t+j}\eta_t). \tag{50.29}$$

Имеем

$$\mathsf{M}q_j = 0, \qquad j > k, \tag{50.30}$$

$$\mathsf{D}q_j = M\varepsilon_{t+j}^2 M\eta_t^2/n = (\mathsf{D}\varepsilon)^2/n, \qquad j > k, \tag{50.31}$$

$$\operatorname{cov}(q_j,\ q_{j+l}) = 0, \qquad l \neq 0. \tag{45.32}$$

Пусть величины

$$\omega_j = q_j/\mathsf{D}u = \frac{1}{n\,\mathsf{D}u}\sum_{t=1}^{n}\varepsilon_{t+j}(\alpha_0 u_t + \ldots + \alpha_k u_{t+k}), \qquad j > k. \tag{50.33}$$

Все ω имеют нулевое среднее, дисперсию, равную $(\mathsf{D}\varepsilon/\mathsf{D}u)^2/n$ и не коррелированы между собой.

Мы видим, что ε_{t+j} совпадает с u_{t+j}, если «исключить из регрессии» члены $u_{t+j-1}, \ldots, u_{t+j-k}$. С другой стороны, η_t совпадает с u_t, если исключить $u_{t+1}, \ldots, u_{t+k}$. Таким образом, при $j > k$ корреляция между ε_{t+j} и η_t, т. е. ω_j, равна частной корреляции членов ряда, отстоящих друг от друга на расстоянии j.

Для больших выборок, используя (50.22) и (50.33) и выборочные значения для сериальных корреляций, получаем

$$\omega_j = \frac{1}{n\,\mathsf{D}u}\sum_{t=1}^{n}(\alpha_0 u_{t+j} + \ldots + \alpha_k u_{t+j-k})(\alpha_0 u_t + \ldots + \alpha_k u_{t+k}) =$$

$$= A_0 r_j + A_1 r_{j-1} + \ldots + A_{2k} r_{j-2k}, \tag{50.34}$$

где

$$A_i = \sum_{j=0}^{i}\alpha_j\alpha_{i-j}. \tag{50.35}$$

Пример 50.3

Для схемы Юла (47.74) в силу (50.35) получаем, что

$$A_0 = 1, \quad A_1 = 2\alpha_1, \quad A_2 = \alpha_1^2 + 2\alpha_2, \quad A_3 = 2\alpha_1\alpha_2, \quad A_4 = \alpha_2^2.$$

Следовательно, асимптотически

$$\omega_j = r_j + 2\alpha_1 r_{j-1} + (\alpha_1^2 + 2\alpha_2)\,r_{j-2} + 2\alpha_1\alpha_2 r_{j-3} + \alpha_2^2 r_{j-4}, \qquad j > 2. \tag{50.36}$$

Дисперсия ω_j равна

$$\frac{1}{n}\left[\frac{1-\alpha_2}{1+\alpha_2}\{(1+\alpha_2)^2 - \alpha_1^2\}\right]. \tag{50.37}$$

В примере 50.2 для ряда из 65 остаточных членов мы получим, что

$$\alpha_1 = -1{,}060; \quad \alpha_2 = 0{,}782.$$

Подставляя в (50.36), имеем

$$\omega_j = r_j - 2{,}12 r_{j-1} + 2{,}688 r_{j-2} - 1{,}658 r_{j-3} + 0{,}612 r_{j-4}.$$

с дисперсией $9{,}69 \times 10^{-4}$.

Как оценки частных корреляций для рядов умеренной длины эти величины довольно безразличны к влиянию выборочных флуктуаций или случайных ошибок. Конечно, оценивание здесь основано на предположении, что мы можем использовать выборочные оценки величин α. Имеем

$$\omega_3 = 0{,}025, \quad \omega_4 = -0{,}043, \quad \omega_5 = -0{,}001$$

при стандартной ошибке, равной 0,031. Как и в примере 50.2, приходим к выводу, что схема второго порядка хорошо согласуется с реальными данными.

Процессы скользящих средних

50.10 Мы уже отмечали трудности, возникающие при оценивании параметров чистого процесса скользящих средних (50.11). Очевидно, что эти трудности возрастают с увеличением длины отрезка усреднения. В этом случае также можно получить асимптотические выражения для функции правдоподобия, но уравнения МП становятся очень громоздкими. Мы опишем метод, принадлежащий Дербину (1959b) и сводящий в конечном счете нашу задачу к задаче авторегрессии.

Рассмотрим простую модель (50.11). Модель эквивалентна бесконечной авторегрессии —

$$u_t - \beta u_{t-1} + \beta^2 u_{t-2} - \ldots = \varepsilon_t. \tag{50.38}$$

Сравним последнюю с конечной схемой авторегрессии порядка k —

$$u_t + \alpha_1 u_{t-1} + \alpha_2 u_{t-2} + \ldots + \alpha_k u_{t-k} = \varepsilon_t, \tag{50.39}$$

где

$$\alpha_k = (-\beta)^k.$$

Схемы отличаются следующими после k-го члена членами авторегрессии:

$$(-\beta)^{k+1} u_{t-k-1} + (-\beta)^{k+2} u_{t-k-2} + \ldots = \\ = (-\beta)^{k+1} \{u_{t-k-1} - \beta u_{t-k-2} + \ldots\} = (-\beta)^{k+1} \varepsilon_{t-k-1}. \quad (50.40)$$

Дисперсия последней величины равна $\beta^{2k+2}\mathrm{D}\varepsilon$ и при $|\beta| < 1$ быстро стремится к нулю при увеличении k. Следовательно, представление (50.39) можно сделать сколь угодно близким к представлению (50.38), если выбирать k достаточно большим (но малым по сравнению с n).

Пусть $a_1, \ldots, a_k$ — оценки величин $\alpha_1, \ldots, \alpha_k$ в (50.39), полученные методом наименьших квадратов. Из теоремы Манна — Вальда (см. **50.7** и (19.16)) следует, что все разности $(a - \alpha)$ асимптотически нормальны с нулевым средним и матрицей рассеяния $\boldsymbol{V}_k^{-1}/n$, где $\boldsymbol{V}_k\,\mathrm{D}\varepsilon$ — матрица рассеяния независимых переменных в уравнении регрессии, т. е. $u_{t-1}, \ldots, u_{t-k}$. Последняя матрица дается (50.12). Следовательно, асимптотически $a_1, \ldots, a_k$ имеют распределение

$$dF = \frac{n^{1/2}\,|V_k|^{1/2}}{(2\pi)^{k/2}} \exp\left[-\frac{n}{2}\left\{(1+\beta^2)\sum_{j=1}^{k}(a_j-\alpha_j)^2 + \right.\right. \\ \left.\left. + 2\beta \sum_{j=1}^{k-1}(a_j-\alpha_j)(a_{j+1}-\alpha_{j+1})\right\}\right] da_1 \ldots da_k. \quad (50.41)$$

Выражение в фигурных скобках, обозначим его Q, — существенная часть функции правдоподобия, так как $|V_k| = (1-\beta^{2k+2})(1-\beta^2)$ (см. упражнение 50.1). Выражение Q можно несколько упростить. Рассмотрим уравнения Юла — Уокера (47.66) в форме

$$\left.\begin{aligned} \alpha_1\gamma_0 + \alpha_2\gamma_1 + \ldots + \alpha_k\gamma_{k-1} &= -\gamma_1, \\ \alpha_1\gamma_1 + \alpha_2\gamma_2 + \ldots + \alpha_k\gamma_k &= -\gamma_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \alpha_1\gamma_{k-1} + \alpha_2\gamma_k + \ldots + \alpha_k\gamma_0 &= -\gamma_k. \end{aligned}\right\} \quad (50.42)$$

Положим $\gamma_0 = (1+\beta^2)\sigma^2$; $\sigma^2 = \mathrm{D}\varepsilon$; $\gamma_1 = \beta\sigma^2$. Тогда

$$\left.\begin{aligned} (1+\beta^2)\alpha_1 + \beta\alpha_2 &= -\beta, \\ \beta\alpha_{r-1} + (1+\beta^2)\alpha_r + \beta\alpha_{r-1} &= 0, \qquad r = 2, \ldots, k-1, \\ \beta\alpha_{k-1} + (1+\beta^2)\alpha_k &= 0. \end{aligned}\right\} \quad (50.43)$$

Умножая j-е уравнение на $-2a_j + \alpha_j$ $(j = 1, \ldots, k)$ и складывая, получаем

$$Q = (1+\beta^2)\sum_{j=1}^{k} a_j^2 + 2\beta\sum_{j=1}^{k-1} a_j a_{j+1} + 2\beta a_1 - \beta\alpha_1. \quad (50.44)$$

Так как при больших k α_1 приблизительно равно $-\beta$, положив $a_0 = 1$, получаем

$$Q = (1+\beta^2)\sum_{j=0}^{k} a_j^2 + 2\beta\sum_{j=0}^{k-1} a_j a_{j+1} - 1. \quad (50.45)$$

Оценку величины β получаем обычным образом, дифференцируя Q и приравнивая производную нулю. Окончательно,

$$b = -\sum_{j=0}^{k-1} a_j a_{j+1} \Big/ \sum_{j=0}^{k} a_j^2. \quad (50.46)$$

50.11 Значение оценки легко вычислить, исходя из величин a; последние в свою очередь легко вычисляются с помощью обычных авторегрессионных приемов. Дербин (1959b), кроме того, указал, что с точностью до первого порядка по n^{-1} (см. упражнение 50.6)

$$\mathrm{D}b = \frac{2}{n\mathsf{M}\,(\partial^2 Q/\partial\beta^2)}. \quad (50.47)$$

В нашем случае *)

$$\mathsf{M}\left(\frac{\partial^2 Q}{\partial\beta^2}\right) = 2\sum_{j=0}^{k}\alpha_j^2,$$

что при больших k стремится к $2\sum_{j=0}^{\infty}(-\beta)^{2j} = 2/(1-\beta^2)$. Итак, при достаточно больших k

$$\mathrm{D}b = (1-\beta^2)/n, \quad (50.48)$$

и оценка b асимптотически эффективна.

50.12 Аналогично можно получить приемлемые результаты и для процессов более высокого порядка, но выражения становятся здесь сложнее. Приведем без доказательства основной результат.

Пусть

$$u_t = \sum_{0}^{n} \beta_j \varepsilon_{t-j}, \quad (50.49)$$

где все ε — независимы и одинаково распределены (не обязательно нормально). Асимптотически оценки a величин α, полученные методом наименьших квадратов, имеют распределение

$$dF = \frac{n^{1/2}\,|B|^{1/2}}{(2\pi)^{k/2}} \exp\left(-\frac{1}{2}nQ\right) da, \quad (50.50)$$

*) С точностью до $o(1/n)$. (*Прим. ред.*)

где $\sigma^2 \boldsymbol{B}$ — матрица рассеяния величин $u_{t-1}, \ldots, u_{t-k}$, а

$$Q = (\boldsymbol{a} - \boldsymbol{\alpha})' \boldsymbol{B} (\boldsymbol{a} - \boldsymbol{\alpha}) = \boldsymbol{a}'\boldsymbol{B}\boldsymbol{a} - 2\boldsymbol{a}'\boldsymbol{B}\boldsymbol{\alpha} + \boldsymbol{a}'\boldsymbol{B}\boldsymbol{\alpha}. \tag{50.51}$$

Последнее упрощается до вида

$$Q = \boldsymbol{a}'\boldsymbol{B}\boldsymbol{a} + 2\boldsymbol{a}'\boldsymbol{c} - \boldsymbol{\alpha}'\boldsymbol{c}, \tag{50.52}$$

где $\boldsymbol{c} = (c_1, \ldots, c_k)$, а $c_j = \gamma_j/\sigma^2$, и снова до вида

$$Q = \boldsymbol{a}'\boldsymbol{B}\boldsymbol{a} + 2\boldsymbol{a}'\boldsymbol{c} + \sum_{j=1}^{h} \beta_j^2. \tag{50.53}$$

Оценки b величин β получаются из соотношения

$$\begin{bmatrix} \sum_0^k a_j^2 & \sum_0^{k-1} a_j a_{j+1} & \cdots & \sum_0^{k-h+1} a_j a_{j+h-1} \\ \sum_0^{k-1} a_j a_{j+1} & \sum_0^k a_j^2 & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \sum_0^{k-h+1} a_j a_{j+h-1} & \sum_0^{k-h+2} a_j a_{j+h-2} & \cdots & \sum_0^k a_j^2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \\ b_h \end{bmatrix} =$$

$$= -\begin{bmatrix} \sum_0^{k-1} a_j a_{j+1} \\ \sum_0^{k-2} a_j a_{j+2} \\ \cdots \\ \sum_0^{k-h} a_j a_{j+h} \end{bmatrix}. \tag{50.54}$$

При больших n матрица рассеяния величин b приблизительно равна

$$\left(\frac{1}{n}\left[\mathsf{M}\,\frac{\partial^2 Q}{\partial\beta_i\,\partial\beta_j}\right]^{-1}\right) \tag{50.55}$$

Можно показать, что последняя равна $\boldsymbol{U}/n$, где

$$\boldsymbol{U} = \begin{pmatrix} 1-\beta_h^2 & \beta_1 - \beta_{h-1}\beta_h & \cdots & \beta_{h-1} - \beta_1\beta_h \\ \beta_1 - \beta_{h-1}\beta_h & 1+\beta_1^2-\beta_{h-1}^2-\beta_h^2 & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \beta_{h-1} - \beta_1\beta_h & \cdots & \cdots & 1-\beta_h^2 \end{pmatrix}. \tag{50.56}$$

50.13 Полученные результаты позволяют в случае больших выборок построить критерии для проверки гипотез. Например, чтобы проверить гипотезу о том, что в (50.11) $\beta = \beta_0$, надо вычислить b, используя (50.46), и проверить, согласуется ли с данными, что

$$z = \sqrt{n}\,(b - \beta_0)\,(1 - \beta_0^2)^{-1/2}, \tag{50.57}$$

имеет стандартное нормальное распределение.

Аналогично в общей схеме (50.49) надо проверить, согласуется ли с данными, что

$$v = n \sum_{i=1}^{h} \sum_{j=1}^{h} u_{ij}^{-1} (b_i - \beta_i)(b_j - \beta_j), \tag{50.58}$$

имеет χ^2-распределение с h ст. св. (Здесь (u_{ij}^{-1}) — матрица, обратная (50.56).)

50.14 Мы снова можем построить критерий согласия для всей модели. Рассматривая простую модель (50.11), заметим, что nQ, где Q дается (50.45), распределено приблизительно, как χ^2 с k ст. св. Подстановка из (50.46) в (50.45) показывает, что Q можно представить еще в виде

$$Q = (1 - b^2) \sum_0^k a_j^2 - 1 + (b - \beta)^2 \sum_0^k a_j^2. \tag{50.59}$$

Асимптотически $n(b-\beta)^2 \sum a_j^2$ эквивалентно «сумме квадратов» в линейной модели регрессии, оставшаяся часть является «остаточной суммой квадратов». Таким образом, адекватность модели можно проверить, исследуя согласие (с реальными данными) того факта, что величина

$$n\left\{(1 - b^2) \sum_0^k a_j^2 - 1\right\} \tag{50.60}$$

распределена по закону χ^2 с $k - 1$ ст. св.

В более общей модели (50.49) минимальное значение nQ равно

$$n\left\{\sum_0^k a_j^2 + \sum_{j=1}^{h} b_j \sum_{i=0}^{k-j} a_i a_{i+j} - 1\right\} \tag{50.61}$$

и распределено, как χ^2 с $k - h$ ст. св., при условии, что модель адекватна.

Подробности и некоторые численные результаты смотрите у Дербина (1959b). Ранее Волд (1949) предложил более сложный критерий. Дербин для случая второго порядка доказал, а для общего случая высказал гипотезу о том, что предельный определитель рассеяния в схеме авторегрессии $\sum_0^h \alpha_j u_{t-j} = \varepsilon_t$ совпадает с предельным определителем рассеяния в схеме

скользящих средних $u_t = \sum_0^h \alpha_j \varepsilon_{t-j}$. Это было строго доказано Финчем (1960) и А. Уокером (1961). Если и существует простое объяснение этого замечательного факта двойственности, то оно до сих пор неизвестно.

Схема авторегрессии со скользящими средними в качестве ошибок

50.15 Рассмотрим смешанную схему

$$\sum_{j=0}^{k} \alpha_j u_{t-j} = \sum_{j=0}^{h} \beta_j \varepsilon_{t-j}. \tag{50.62}$$

Задача получения эффективных оценок величин α и β еще полностью не исследована. Наиболее обещающий метод, принадлежащий Дербину (1960b), носит итерационный характер и состоит в следующем.

Допустим, что нам дана некоторая последовательность $\boldsymbol{a}$ значений α. Авторегрессионным преобразованием сведем величины u к величинам

$$z_t = \sum_{j=0}^{k} a_j u_{t-j} \tag{50.63}$$

и оценим постоянные β в модели

$$z_t = \sum \beta_j \varepsilon_{t-j}. \tag{50.64}$$

Получив оценки $\boldsymbol{b}$ величин β, перейдем от уравнений

$$\sum a_j u_{t-j} = \sum b_j \varepsilon_{t-j}$$

к уравнениям авторегрессии вида

$$\sum \alpha'_j u_{t-j} = \varepsilon_t, \tag{50.65}$$

где все α' — линейные комбинации величин α. Мы можем теперь получить оценки величин α', а следовательно, и α. При неограниченном повторении процедуры значения оценок величин α и β сходятся к желаемым значениям.

50.16 Применение описанной процедуры связано с решением двух задач. Первая состоит в обеспечении достаточно быстрой сходимости, вторая — в выборе «хорошей» последовательности начальных значений. Первая задача полностью еще не исследована, — в анализе временных рядов хорошая сходимость — это не то, на что можно положиться без достаточной проверки на практике. Что касается второй задачи, то здесь следует помнить, что хорошим приближением для схемы (50.62) служит авторегрессионная схема высокого порядка. Выберем какую-либо схему такого типа и обозначим «остатки» символами e_t. Заменим в (50.62) ε_t на e_t. Тогда предварительные оценки величин α и β можно получить, минимизируя

$$\left\{ \sum_{j=0}^{k} \alpha_j u_{t-j} - \sum_{j=1}^{h} \beta_j e_{t-j} \right\}^2. \tag{50.66}$$

Пример 50.4

Рассмотрим модель

$$u_t + \alpha u_{t-1} = \varepsilon_t + \beta \varepsilon_{t-1}. \tag{50.67}$$

Приблизим ее схемой порядка k с «остатками» e_t:

$$e_t = u_t + a_1 u_{t-1} + \ldots + a_k u_{t-k}. \tag{50.68}$$

Минимизируем

$$\sum (u_t + \alpha u_{t-1} - \beta e_{t-1})^2$$

и для оценок величин α и β получим

$$\sum u_t u_{t-1} + \hat{\alpha} \sum u_{t-1}^2 - \hat{\beta} \sum e_{t-1} u_{t-1} = 0, \tag{50.69}$$

$$\sum u_t e_{t-1} - \hat{\alpha} \sum u_{t-1} e_{t-1} - \hat{\beta} \sum e_{t-1}^2 = 0. \tag{50.70}$$

Подставляя значения e из (50.68) и заменяя $\sum u_t u_{t-p}$ на $\left(\sum u_t^2\right) r_p$, получим асимптотические выражения

$$\hat{\alpha} = -\frac{a_1 r_2 + a_2 r_3 + \ldots + a_k r_{k+1}}{a_1 r_1 + a_2 r_2 + \ldots + a_k r_k}, \tag{50.71}$$

$$\hat{\beta} = \hat{\alpha} + \frac{r_1 + a_1 r_2 + \ldots + a_k r_{k+1}}{1 + a_1 r_1 + \ldots + a_k r_k}. \tag{50.72}$$

С этих оценок можно начать итерационный процесс.

50.17 Производящая функция автоковариаций для смешанной схемы (50.62) со случайными значениями ε равна (см. **47.15** и **47.18**)

$$G(z) = \frac{\left(\sum \beta_j z^j\right)\left(\sum \beta_j z^{-j}\right)}{\left(\sum \alpha_j z^j\right)\left(\sum \alpha_j z^{-j}\right)}. \tag{50.73}$$

Соответствующая спектральная плотность (см. **47.14**) равна

$$w(\alpha) = \frac{\left(\sum \beta_j e^{i\alpha j}\right)\left(\sum \beta_j e^{-i\alpha j}\right)}{\left(\sum \alpha_j e^{i\alpha j}\right)\left(\sum \alpha_j e^{-i\alpha j}\right)}. \tag{50.74}$$

Основываясь на оценках величин α и β, можно оценить и спектральную плотность. Вероятно, более уместно ставить вопрос об

оценивании величин α и β с помощью наблюдений спектра. Этот вопрос изучался Дербином (1961).

50.18 Уиттл в серии своих статей (в частности, (1951), (1953a) и (1953b)) решал более общие задачи проверки гипотез и оценивания. Методы этих работ, основанные большей частью на максимизации функций правдоподобия, и сами результаты относятся к стационарным временным рядам с непрерывным спектром и носят весьма общий характер. Однако они не слишком удобны для вычислений.

Некоторые многомерные обобщения

50.19 Допустим, что мы располагаем p рядами u_i $(i = 1, \ldots, p)$, наблюдаемыми в n временных точках, или в непрерывном случае определенными на некотором отрезке времени. Значение u_i во временной точке t обозначим u_{it}, а совокупность $(p \times n)$ значений u матрицей $\boldsymbol{u}$. Как и в одномерном случае, будем рассматривать u как реализацию процесса и его параметры.

В общем случае любой вектор-строка из $\boldsymbol{u}$, рассматриваемый как один из рядов, может содержать тренд, сезонность и осцилляции. Однако попытка *одновременно* разложить все векторы на их составляющие приводит к практически непреодолимым трудностям. Предположим, что тренд и сезонные изменения уже выделены, и мы остались с многомерным стационарным комплексом $\boldsymbol{u}$. Будем следовать Кеную (1957).

50.20 Обозначим $\gamma_{(ij)s}$ и $\rho_{(ij)s}$ соответственно ковариацию и корреляцию u_{it} и $u_{j,t-s}$. Аналогичные выборочные величины обозначим $c_{(ij)s}$ и $r_{(ij)s}$.

При любом s совокупности из $p(p+1)/2$ такнх величин будем представлять в виде квадратных матриц $\boldsymbol{\gamma}_s$, $\boldsymbol{\rho}_s$, $\boldsymbol{c}_s$ и $\boldsymbol{r}_s$. В одномерном случае $\rho_s = \rho_{-s}$, но ясно, что $\mathbf{M}(u_{it}u_{j,t+s})$ не равно $\mathbf{M}(u_{it}u_{j,t-s})$, но равно $\mathbf{M}(u_{i,t-s}u_{jt})$. Следовательно,

$$\boldsymbol{\gamma}_s = \boldsymbol{\gamma}'_{-s}. \tag{50.75}$$

Как и свойственно многомерному случаю, число параметров и оценок резко вырастает с ростом p. Будем называть $\gamma_{(ij)s}$ *взаимной ковариацией* u_i и u_j с *запаздыванием* s. Аналогично $\rho_{(ij)s}$ назовем *взаимной корреляцией*. Мы будем оговаривать те случаи, когда необходимо различать теоретические и выборочные величины, хотя чаще всего это можно не делать, если символы сами по себе указывают, с чем мы имеем дело.

50.21 Как и в одномерном случае рассмотрим три типа моделей: авторегрессию, скользящие средние и смешанную схему

авторегрессии и скользящих средних. Пусть ε_{it} — ряд независимых случайных элементов. Аналогично одномерному случаю

$$u_t = \sum_{s=0}^{l} \beta_s \varepsilon_{t-s};$$

теперь мы будем опираться на соотношения

$$u_{it} = \sum_{s=0}^{l} \sum_{j=1}^{p} \beta_{ijs} \varepsilon_{j,t-s}. \tag{50.76}$$

Для соответствующей схемы авторегрессии имеем

$$\sum_{s=0}^{k} \sum_{j=1}^{p} \alpha_{ijs} u_{j,t-s} = \varepsilon_{it}. \tag{50.77}$$

Пусть D — оператор сдвига такой, что $Du_t = u_{t-1}$. Выразим (50.76) в форме

$$u_{it} = \sum_{s=0}^{l} \sum_{j=1}^{p} \beta_{ijs} D^s \varepsilon_{jt}, \tag{50.78}$$

или в матричной форме

$$\boldsymbol{u}_t = \sum_{s=0}^{l} \boldsymbol{B}_s D^s \boldsymbol{\varepsilon}_t = \tag{50.79}$$

$$= \boldsymbol{B}(D)\, \boldsymbol{\varepsilon}_t, \tag{50.80}$$

где

$$\boldsymbol{B}(D) = \sum_{s=0}^{l} \boldsymbol{B}_s D^s. \tag{50.81}$$

Точно так же в схеме авторегрессии (50.77)

$$\boldsymbol{\varepsilon}_t = \sum_{s=0}^{k} \boldsymbol{A}_s D^s \boldsymbol{u}_t = \boldsymbol{A}(D)\, \boldsymbol{u}_t, \tag{50.82}$$

где

$$\boldsymbol{A}(D) = \sum_{s=0}^{k} \boldsymbol{A}_s D^s. \tag{50.83}$$

Мы можем выписать решение (50.82):

$$\boldsymbol{u}_t = \boldsymbol{A}^{-1}(D)\, \boldsymbol{\varepsilon}_t.$$

Элементы матриц A и B — полиномы от D. Определим также величину

$$\boldsymbol{\gamma} = \sum_{s=-\infty}^{\infty} \boldsymbol{\gamma}_s D^s. \tag{50.84}$$

50.22 Не теряя общности, мы можем изменить масштаб так, чтобы ε_i имели нулевые средние, одну и ту же дисперсию, скажем σ^2, и были не коррелированы. Тогда

$$\gamma_s = \mathsf{M}(u_t u'_{t-s}).$$

Подставляя u_t из (50.79), получим, что

$$\gamma_s = \mathsf{M}\left(\sum_{j=0}^{l} B_j D^j \varepsilon_t\right)\left(\sum_{m=0}^{l} B_m D^m \varepsilon_{t-s}\right)' =$$
$$= \sum_{j=0}\sum_{m=0} B_j B'_m \mathsf{M}(\varepsilon_{t-j}\varepsilon'_{t-s-m}) = \sigma^2 \sum_{s,m} B_{s+m} B'_m =$$
$$= \text{коэфф. при } D^s \text{ в } \sigma^2\left(\sum B_j D^j\right)\left(\sum B'_j D^{-j}\right). \qquad (50.85)$$

Следовательно, можно записать, что

$$\gamma = B(D)\, B'(D^{-1}). \qquad (50.86)$$

Аналогично для уравнения авторегрессии

$$\gamma = A^{-1}(D)\left(A'(D^{-1})\right)^{-1}. \qquad (50.87)$$

Легко показать, что для смешанной схемы $Au_t = B\varepsilon_t$

$$\gamma = A^{-1}(D)\, B(D)\, B'(D^{-1})\left(A'(D^{-1})\right)^{-1}. \qquad (50.88)$$

Это многомерные аналоги (50.73). В них все D можно рассматривать как вспомогательные переменные аналогично тому, что мы раньше записывали как z. Фактически (50.86)—(50.88) дают нам производящие функции ковариаций.

50.23 Получим для схемы авторегрессии естественное обобщение уравнений Юла — Уокера (47.66).

Умножая (50.82) справа на u'_{t-q} и беря математическое ожидание, получаем

$$\mathsf{M}\sum_{s=0}^{k} A_s D^s u_t u'_{t-q} = \sum_{s=0}^{k} A_s \mathsf{M}(u_{t-s} u'_{t-q}) =$$
$$= \sum_{s=0}^{k} A_s \gamma_{q-s} = 0, \qquad q > 0. \qquad (50.89)$$

Однако решить эти уравнения — дело нелегкое.

Вырождение

50.24 К обычным трудностям анализа временных рядов, с которыми мы сталкивались и в одномерном случае, в многомерной ситуации добавляются еще две.

Прежде всего между переменными могут существовать линейные связи, и в этом случае матрицы A и B могут оказаться вырожденными. Тогда следует предпринять шаги к исключению некоторых переменных.

Пример 50.5 (Кенуй, 1957)

Рассмотрим схему

$$\left.\begin{aligned} u_{1t} &= \varepsilon_{1t} + \varepsilon_{2,t-1},\\ u_{2t} &= \varepsilon_{1t} + \varepsilon_{2t},\\ u_{3t} &= \varepsilon_{2t} + \varepsilon_{2,t-1}. \end{aligned}\right\} \qquad (50.90)$$

Имеем

$$B = \begin{pmatrix} 1 & D \\ 1 & 1 \\ 0 & 1+D \end{pmatrix}. \qquad (50.91)$$

В силу (50.86)

$$\gamma = B(D)\, B'(D^{-1}) = \begin{pmatrix} 2 & 1+D & 1+D \\ 1+D^{-1} & 2 & 1+D^{-1} \\ 1+D^{-1} & 1+D & 2+D+D^{-1} \end{pmatrix}. \qquad (50.92)$$

Отсюда получаем, что $|\gamma| = 0$ и ранг матрицы равен 2. Тогда между величинами должна существовать линейная связь. В этом случае ее можно увидеть почти сразу, но формально мы должны найти собственный вектор, соответствующий нулевому собственному числу γ. Этот вектор пропорционален $(1+D, -(1+D), 1-D)$, и, таким образом, связь такова:

$$(1+D)\,u_{1t} - (1+D)\,u_{2t} + (1-D)\,u_{3t} = 0,$$

или

$$u_{1t} - u_{2t} + u_{3t} = -u_{1,t-1} + u_{2,t-1} + u_{3,t-1}. \qquad (50.93)$$

50.25 На практике вырожденность встречается скорее как исключение, чем как правило, и если она появляется, не составляет труда действовать в духе примера 50.5. Бóльшие трудности приносит то важное обстоятельство, что, какими бы хорошими ни были оценки величин γ, уравнение (50.88) не определяет A и B однозначно.

Будем временно писать F вместо $A^{-1}B$. Тогда (50.88) принимает вид

$$\gamma = F(D)\, F'(D^{-1}) \qquad (50.94)$$

Пусть Φ — любая диагональная матрица, i-й диагональный элемент которой равен $\varphi_i(D)/\varphi_i(D^{-1})$, Ψ — любая диагональная

матрица с диагональными элементами типа $\psi_i(D)/\psi_i(D^{-1})$, а $\boldsymbol{J}$ — любая матрица, такая что $\boldsymbol{JJ}' = \boldsymbol{I}$. Легко видеть, что тогда в (50.94) $\boldsymbol{F}(D)$ можно заменить на $\boldsymbol{F}(D)\boldsymbol{\Phi}(D)\boldsymbol{J}\boldsymbol{\Psi}(D)$ и, таким образом, одной ковариационной матрице соответствует много различных схем.

Пример 50.6 (Кенуй, 1957)

Рассмотрим матрицы

$$\boldsymbol{F}_1 = \begin{pmatrix} 2+D & D \\ 1 & 6+D \end{pmatrix}, \qquad \boldsymbol{F}_2 = \begin{pmatrix} 1+2D & -D \\ 5 & 3+2D \end{pmatrix},$$

$$\boldsymbol{F}_3 = \frac{1}{5}\begin{pmatrix} 2+7D & 4+9D \\ -11-6D & 28+3D \end{pmatrix}, \qquad \boldsymbol{F}_4 = \frac{1}{17}\begin{pmatrix} 14+37D & -5-12D \\ 55+30D & 1+84D \end{pmatrix}.$$

Легко проверить, что

$$|F_1| = (4+D)(3+D), \qquad |F_2| = (3+D)(1+4D),$$
$$|F_3| = (4+D)(1+3D), \qquad |F_4| = (1+4D)(1+3D),$$

и что для каждой F

$$\boldsymbol{\gamma} = \begin{pmatrix} 6+2D+2D^{-1} & 3+7D \\ 3+7D^{-1} & 38+6D+6D^{-1} \end{pmatrix}.$$

Более того, если мы умножим $\boldsymbol{F}_1$, $\boldsymbol{F}_2$, $\boldsymbol{F}_3$, $\boldsymbol{F}_4$ соответственно на ортогональные матрицы $\boldsymbol{J}_1$, $\boldsymbol{J}_2$, $\boldsymbol{J}_3$, $\boldsymbol{J}_4$, $\boldsymbol{\gamma}$ не изменится.

50.26 Нам осталось понять, все ли решения (50.88) приемлемы; например, все ли они обеспечивают стационарность рядов. Насколько известно, должны выполняться довольно жесткие условия, для того чтобы решение было единственным. Следующий подход принадлежит Филлипсу (1959).

Рассмотрим смешанную схему с независимыми остатками и предположим, что: (1) $\boldsymbol{A}$ не вырождена (этого можно добиться, поступая так, как это делалось в примере 50.5) и (2) $|A|$, как полином от D, имеет *различные* корни $\lambda_1, \ldots, \lambda_m$. Пусть $\boldsymbol{\alpha}(D)$ — присоединенная матрица для $A(D)$ *). Тогда (50.88) можно переписать в виде

$$\boldsymbol{\gamma} = \frac{\boldsymbol{\alpha}(D)\boldsymbol{B}(D)\boldsymbol{B}'(D^{-1})\boldsymbol{\alpha}'(D^{-1})}{|A(D)|\,|A'(D^{-1})|}. \tag{50.95}$$

Представляя $|A|$ в виде произведения $\prod_{r=1}^{m}(D-\lambda_r)$, мы видим, что правая часть (50.95) допускает разложение на элементарные

*) Термин «присоединенная» употребляется разными авторами в разных смыслах. Здесь под «присоединенной для A» матрицей имеется в виду матрица $\boldsymbol{A}^{-1}|A|$. (*Прим. перев.*)

дроби —

$$\sum_{r=1}^{m}\frac{\boldsymbol{K}_r}{D-\lambda_r} + \sum_{r=1}^{m}\frac{\boldsymbol{K}_r'}{D^{-1}-\lambda_r}, \tag{50.96}$$

где $\boldsymbol{K}_r$ — $(p \times p)$-матрица, определяемая, как это делается, в обычной теории элементарных дробей, формулой

$$\boldsymbol{K}_r = \left[\frac{(D-\lambda_r)\,\boldsymbol{\alpha}(D)\boldsymbol{B}(D)\boldsymbol{B}'(D^{-1})\boldsymbol{\alpha}'(D^{-1})}{|A(D)|\,|A'(D^{-1})|}\right]_{D=\lambda_r}. \tag{50.97}$$

Не желая, чтобы (50.96) содержало положительные степени D, мы считаем, что степень $|B|$ меньше, чем $|A|$.

Так как λ_r — простой корень, $\boldsymbol{A}(\lambda_r)$ — однократно вырождена, а ранг присоединенной матрицы $\boldsymbol{\alpha}(\lambda_r)$ равен единице (известный результат из теории матриц). Тогда

$$\boldsymbol{\alpha}(\lambda_r) = \boldsymbol{k}_r\boldsymbol{\varkappa}_r, \tag{50.98}$$

где $\boldsymbol{k}_r$ — $(p \times 1)$-вектор-столбец и $\boldsymbol{\varkappa}_r$ — $(1 \times p)$-вектор-строка такие, что

$$\boldsymbol{A}(\lambda_r)\boldsymbol{k}_r = \boldsymbol{0}, \tag{50.99}$$

$$\boldsymbol{\varkappa}_r\boldsymbol{A}(\lambda_r) = \boldsymbol{0}. \tag{50.100}$$

В действительности ранг $\boldsymbol{K}_r$ тоже равен единице, так как, если $(1 \times p)$-вектор-строка

$$\boldsymbol{l}_r = \left[\frac{(D-\lambda_r)\,\boldsymbol{\varkappa}_r\boldsymbol{B}(D)\boldsymbol{B}'(D^{-1})\boldsymbol{\alpha}'(D^{-1})}{|A(D)|\,|A'(D^{-1})|}\right]_{D=\lambda_r}, \tag{50.101}$$

то, подставляя (50.98) в (50.97), получим

$$\boldsymbol{K}_r = \boldsymbol{k}_r\boldsymbol{l}_r. \tag{50.102}$$

Далее из (50.99) следует, что

$$\boldsymbol{A}(\lambda_r)\boldsymbol{K}_r = \boldsymbol{0}. \tag{50.103}$$

Тогда, если матрица $\boldsymbol{\gamma}$ задана, мы можем разложить ее на элементарные дроби и определить $\boldsymbol{K}_r$ и λ_r. Таким образом, мы можем получить систему уравнений (50.103). Однако не ясно, достаточно ли этих уравнений, чтобы определить A единственным образом.

50.27 Рассмотрим прежде всего тот случай, когда все скалярные уравнения в $\boldsymbol{A}\boldsymbol{u}_t = \boldsymbol{B}\boldsymbol{\varepsilon}_t$ имеют один порядок v, и порядок не может быть понижен. Наложим еще два ограничения: (а) все элементы главной диагонали $\boldsymbol{A}$ имеют степень v, а степень недиагональных элементов не превосходит $v-1$ (это означает, кроме всего, что $|A|$ не равен нулю); (б) элементы соответствующей строки в $\boldsymbol{B}$ имеют степень, меньшую чем v (это означает, кроме всего, что λ не присутствует в числителях (50.96)).

Не теряя общности, можно предположить, что коэффициент при D^v в каждом диагональном элементе $\boldsymbol{A}$ равен единице. $|A|$ имеет степень pv, и, следовательно, $pv = m$. В каждой строке $\boldsymbol{A}$ нужно определить pv коэффициентов. Для последних, положив последовательно в (50.103) $r = 1, 2, \ldots, m$, получаем неоднородную систему независимых уравнений. Таким образом, элементы матрицы $\boldsymbol{A}$ определяются однозначно (т. е. матрица $\boldsymbol{A}$ *идентифицируема*).

После того как $\boldsymbol{A}$ найдено, из (50.95) и (50.96) получаем формулу для $\boldsymbol{B}(D)\boldsymbol{B}'(D^{-1})$

$$\boldsymbol{B}(D)\boldsymbol{B}'(D^{-1}) = \sum_{r=1}^{m} \frac{\boldsymbol{A}(D)\boldsymbol{K}_r\boldsymbol{A}'(D^{-1})}{D-\lambda_r} + \sum_{r=1}^{m} \frac{\boldsymbol{A}(D)\boldsymbol{K}_r'\boldsymbol{A}'(D^{-1})}{D^{-1}-\lambda_r}. \quad (50.104)$$

Сама матрица $\boldsymbol{B}$ остается неоднозначно определенной. Однако из этого положения можно выйти, лишь используя дополнительную информацию.

50.28 Если уравнения системы имеют разный порядок, для идентификации $\boldsymbol{A}$ требуется еще одно предположение. Пусть уравнения в $\boldsymbol{A}\boldsymbol{u}_t = \boldsymbol{B}\boldsymbol{\varepsilon}_t$ устроены так, что первое имеет наименьший порядок, а каждое следующее имеет порядок не меньший, чем предыдущее. Если $\boldsymbol{A}$ и $\boldsymbol{B}$ удовлетворяют (50.88), то это же верно и для $\boldsymbol{\mu}\boldsymbol{A}$ и $\boldsymbol{\mu}\boldsymbol{B}$, если $\boldsymbol{\mu}$ — произвольная матрица постоянных. Мы можем добавить любую строку матрицы $\boldsymbol{A}$ к любой последующей строке, не нарушая условия, состоящего в том, что степень недиагональных элементов меньше, чем у диагональных. Но мы не можем прибавлять строку к предшествующей. Тогда $\boldsymbol{\mu}$ может быть только треугольной матрицей с нулями выше диагонали.

Мы можем сделать систему идентифицируемой (однозначно определенной), если предположим, что элементы $\boldsymbol{\varepsilon}$ не коррелированы от одного уравнения к другому, иначе говоря, что $\boldsymbol{B}$ диагональна. Тогда $\boldsymbol{B}(D)\boldsymbol{B}'(D^{-1})$ диагональна, и, следовательно, недиагональные элементы матрицы $\boldsymbol{\mu}\boldsymbol{B}\boldsymbol{B}'\boldsymbol{\mu}'$ равны нулю. Если b_{ii} — диагональные элементы $\boldsymbol{B}\boldsymbol{B}'$, то недиагональные элементы $\boldsymbol{\mu}\boldsymbol{B}(D)\boldsymbol{B}'(D^{-1})\boldsymbol{\mu}'$, находящиеся выше диагонали, равны

$$\begin{array}{lll} \mu_{11}b_{11}\mu_{21} & \mu_{11}b_{11}\mu_{31} & \mu_{11}b_{11}\mu_{41}\ \ldots \\ & \mu_{21}b_{11}\mu_{31} + \mu_{22}b_{22}\mu_{32} & \mu_{21}b_{11}\mu_{41} + \mu_{22}b_{22}\mu_{42}\ \ldots \\ & \cdot & \cdot \\ & \cdot & \cdot \\ & \cdot & \cdot \end{array}$$

Так как μ_{11} не может равняться нулю, μ_{21}, μ_{31} и т. д. равны нулю, а следовательно, равны нулю μ_{32}, μ_{42} и т. д. Итак, $\boldsymbol{\mu}$ — диагональна, и уравнения являются идентифицируемыми.

50.29 Достаточно ясно, что задачи, связанные с идентификацией очень трудны. Мы можем, по крайней мере исходя из эвристических соображений, оценить матрицу $\boldsymbol{\gamma}$, а следовательно, и произведение $\boldsymbol{A}^{-1}(D)\boldsymbol{B}(D)\boldsymbol{B}'(D^{-1})[\boldsymbol{A}'(D^{-1})]^{-1}$. Условия, которые обеспечивают переход к отдельным элементам в $\boldsymbol{A}$ и $\boldsymbol{B}$, не всегда легко проверить, и во всяком случае, чтобы достигнуть определенности, часто приходится прибегать к посторонней информации. Задача построения однозначно определенных моделей является одной из самых серьезных в теории многомерных временных систем, но она выходит за рамки выборочного анализа.

Взаимные спектральные функции

50.30 Если мы могли рассматривать взаимные корреляции рядов и получать то, что можно назвать взаимной коррелограммой, то мы можем перейти и к обобщению на многомерный случай спектрального анализа.

Любой паре рядов, скажем u_1 и u_2, соответствует последовательность взаимных корреляций $\rho_{(12)s}$, $s = -\infty, \ldots, \infty$. В качестве обобщения понятия спектра одного ряда (см. (47.21)), определим спектральную плотность формулой

$$w_{12}(\alpha) = \sum_{-\infty}^{\infty} \rho_{(12)s} \exp(is\alpha). \quad (50.105)$$

Соответствующая спектральная функция $W(\alpha)$ определена на интервале от 0 до π. Обратно, как и в (47.23),

$$\rho_{(12)s} = \frac{1}{\pi} \sum w_{12}(\alpha) \exp(-is\alpha). \quad (50.106)$$

В одномерных формулах, благодаря симметрии типа $\rho_s = \rho_{-s}$, в выражениях, связывающих спектральную плотность с ковариациями или корреляциями, члены с синусами взаимно уничтожались. В многомерном случае $\rho_{(12)s}$ не совпадает с $\rho_{(12),(-s)}$. Из (50.105), получаем, что

$$w_{12}(\alpha) = 1 + \sum_{1}^{\infty} \{\rho_{(12)s} \cos s\alpha + \rho_{(12)(-s)} \cos s\alpha\} +$$
$$+ i \sum_{1}^{\infty} \{\rho_{(12)s} \sin s\alpha - \rho_{(12)(-s)} \sin s\alpha\} = c(\alpha) + iq(\alpha). \quad (50.107)$$

50.31 $c(\alpha)$ называют *ко-спектром* или *ко-спектральной* плотностью, $q(\alpha)$ — *квадратурным спектром* или *квадратурной спектральной плотностью*. Иногда чертят заранее графики

зависимости обеих этих функций от α. Сумму квадратов $c^2 + q^2$ называют *амплитудой* спектра. Нормированное выражение

$$C(\alpha) = \frac{c^2(\alpha) + q^2(\alpha)}{w_1(\alpha)\, w_2(\alpha)}, \tag{50.108}$$

где w_1 и w_2 — спектральные плотности u_1 и u_2, называют *когерентностью*.

Фазовые связи между процессами изучаются с помощью диаграмм трех типов: *фазовая диаграмма* дает зависимость $\psi(\alpha)$ от α, где

$$\psi(\alpha) = \operatorname{arctg}\left\{\frac{q(\alpha)}{c(\alpha)}\right\}; \tag{50.109}$$

в *диаграмме Арганда* по оси абсцисс откладываются значения $c(\alpha)/w_1(\alpha)$, а по оси ординат — значения $q(\alpha)/w_1(\alpha)$; *диаграмма усиления* дает зависимость $R_{12}^2(\alpha)$ от α, где

$$R_{12}^2(\alpha) = \frac{w_1(\alpha)\, C(\alpha)}{w_2(\alpha)}. \tag{50.110}$$

Провести вычисления и построить графики можно с помощью вычислительной машины. (50.108) и (50.110) являются аналогами коэффициентов корреляции и регрессии. Дальнейшие подробности можно найти в работе Грэнджера и Хатанаки (1964).

50.32 Для многомерных рядов типа скользящих средних или авторегрессии существует непосредственная связь между производящей функцией ковариаций и спектральной плотностью. Действительно,

$$w(\alpha) = \boldsymbol{A}^{-1}(e^{i\alpha})\, \boldsymbol{B}(e^{i\alpha})\, \boldsymbol{B}'(e^{-i\alpha})\, [\boldsymbol{A}'(e^{i\alpha})]^{-1}. \tag{50.111}$$

Однако практически это не очень полезная формула.

Бриллинджер (1965) обсуждал переход от простого и взаимного спектра к полиспектру k-мерного временного ряда.

Пример 50.7

Чтобы показать, как выглядят взаимные корреляции и взаимный спектр, рассмотрим искусственный ряд, построенный Кенуем (1957) и приведенный в таблице 50.2. Ряд удовлетворяет соотношениям

$$\begin{aligned} u_{1t} &= u_{1,t-1} - 0{,}1u_{2,t-1} + \varepsilon_{1t}, \\ u_{2t} &= 0{,}2u_{1,t-1} + u_{2,t-1} - 0{,}3u_{3,t-1} + \varepsilon_{2t}, \\ u_{3t} &= 0{,}9u_{2,t-1} + \varepsilon_{3t}. \end{aligned} \tag{50.112}$$

Таблица 50.2

Искусственный ряд (см. текст)

t	u_{1t}	u_{2t}	u_{3t}	t	u_{1t}	u_{2t}	u_{3t}	t	u_{1t}	u_{2t}	u_{3t}
1	118	155	150	35	102	19	35	69	—45	—137	—88
2	102	128	114	36	139	61	—15	70	—74	—125	—132
3	131	148	122	37	128	58	27	71	—30	—92	—73
4	129	99	132	38	101	64	85	72	—46	—70	—43
5	72	97	114	39	132	91	85	73	—48	—93	—72
6	78	63	123	40	95	89	95	74	—88	—66	—78
7	117	33	91	41	43	94	73	75	—70	—48	—15
8	134	44	67	42	43	41	91	76	—82	—16	—11
9	174	73	12	43	62	—6	0	77	—74	—34	—1
10	146	63	104	44	108	—11	16	78	—117	—22	13
11	94	54	38	45	88	—19	2	79	—97	—10	—13
12	62	27	42	46	134	36	—2	80	—51	—64	4
13	56	2	—17	47	139	51	—8	81	—42	—101	—34
14	39	37	2	48	153	69	52	82	—29	—90	—131
15	37	0	70	49	178	108	82	83	—1	—88	—105
16	8	—42	45	50	181	127	129	84	—28	—29	—71
17	—25	—79	—54	51	199	162	84	85	—1	—46	—48
18	17	—104	—103	52	137	139	142	86	26	17	—39
19	38	—65	—109	53	126	100	142	87	62	70	—29
20	42	2	—18	54	73	130	70	88	81	107	33
21	49	6	—10	55	76	124	121	89	77	149	103
22	52	21	38	56	46	152	124	90	57	172	126
23	51	—20	29	57	31	173	135	91	52	156	169
24	11	—16	—19	58	—20	165	160	92	54	72	172
25	2	—29	—26	59	—54	109	162	93	82	11	19
26	—18	11	—18	60	—29	60	52	94	45	45	2
27	7	—15	—3	61	—32	68	64	95	40	53	54
28	20	28	—58	62	—82	72	95	96	33	11	21
29	—29	6	27	63	—110	3	55	97	46	46	—28
30	—20	39	53	64	—76	—34	—23	98	—7	108	48
31	11	—12	68	65	—69	—83	—15	99	4	94	94
32	57	7	—20	66	—92	—51	—35	100	37	61	48
33	65	—3	—37	67	—90	—100	—46				
34	98	9	43	68	—71	—122	—166				

Величины ε равномерно распределены на отрезке от —49 до 49. В таблице 50.3 приведены значения теоретических ковариаций γ_s и выборочных ковариаций c_s для $s = 0, 1, 2, 3, 4, 5$.

Сериальные корреляции вплоть до 25-го порядка приведены в таблице 50.4.

На рис. 50.2 и 50.3 приведены графики логарифмов спектральных плотностей для трех рядов и логарифмов амплитуд взаимных спектров. Трехмерный ряд является марковским. Взаимные спектры его компонент достаточно хорошо отражают модель. Вообще говоря, при изучении используются и моменты порядка, большего чем 2.

50.33 Вряд ли стоит говорить, что задачи оценивания и проверки гипотез, сами по себе далеко не простые, значительно усложняются при переходе от одномерного случая к многомерному.

Коррелограммы отдельных рядов и спектры обычно рассматривают для того, чтобы выяснить, достаточно ли нам марковской схемы для объяснения эволюции данных или требуется более сложная модель. Решение таких вопросов основывается на характере сериальных корреляций и ковариаций. Поскольку в многомерной теории роль дисперсии играет определитель матрицы рассеяния, можно ожидать, что для многомерных временных рядов соответствующие характеристики будут основаны на автоковариационных и автокорреляционных определителях. Точная выборочная теория таких величин еще не развита, и на сегодняшнем уровне знаний мы вынуждены довольствоваться недостаточно обоснованными, хотя и интуитивно разумными, процедурами.

50.34 Рассмотрим сначала ковариационный определитель $|\gamma_s|$. В схеме скользящих средних такие определители равны нулю, начиная с некоторого значения s, скажем l. В схеме авторегрессии между последовательными матрицами γ существует связь. Так, в марковском случае

$$|\gamma_s| = \beta^s |\gamma_0|, \tag{50.113}$$

(ср. с (47.22)), где

$$\beta = -|A_1|/|A_0|; \tag{50.114}$$

а в схеме Юла (см. пример 47.8)

$$\begin{vmatrix} \gamma_{s+1} & \gamma_s \\ \gamma_s & \gamma_{s-1} \end{vmatrix} = \beta^s \begin{vmatrix} \gamma_1 & \gamma_0 \\ \gamma_0 & \gamma_1' \end{vmatrix}, \tag{50.115}$$

где

$$\beta = -|A_2|/|A_0|. \tag{50.116}$$

К сожалению, с ростом размерности растут выборочные ошибки, искажающие истинный характер связей.

Таблица 50.3

Значения γ_s и c_s для ряда таблицы 50.2. (Подчеркнуты значения определителей соответствующих матриц)

s	γ_s	c_s
0	$\begin{bmatrix} 7\,878{,}14 & 4\,392{,}92 & 3\,396{,}37 \\ 4\,392{,}92 & 6\,191{,}72 & 4\,010{,}45 \\ 3\,396{,}37 & 5\,010{,}45 & 5\,831{,}96 \end{bmatrix}$ $\underline{5{,}2245 \times 10^{10}}$	$\begin{bmatrix} 6\,896{,}05 & 4\,413{,}37 & 3\,511{,}74 \\ 4\,413{,}37 & 6\,625{,}07 & 5\,573{,}84 \\ 3\,511{,}74 & 5\,573{,}84 & 6\,161{,}75 \end{bmatrix}$ $\underline{3{,}8320 \times 10^{10}}$
1	$\begin{bmatrix} 7\,438{,}85 & 3\,773{,}75 & 2\,895{,}33 \\ 4\,949{,}64 & 5\,567{,}17 & 3\,940{,}14 \\ 3\,935{,}63 & 5\,572{,}55 & 4{,}509{,}41 \end{bmatrix}$ $\underline{1{,}4106 \times 10^{10}}$	$\begin{bmatrix} 6\,487{,}04 & 3\,877{,}69 & 3\,097{,}75 \\ 4\,822{,}82 & 5\,984{,}09 & 4\,624{,}31 \\ 3\,982{,}18 & 5\,944{,}23 & 4\,979{,}27 \end{bmatrix}$ $\underline{8{,}2402 \times 10^{9}}$
2	$\begin{bmatrix} 6\,943{,}89 & 3\,217{,}03 & 2\,501{,}31 \\ 5\,251{,}32 & 4\,650{,}15 & 3\,166{,}38 \\ 4\,454{,}67 & 5\,010{,}45 & 3\,546{,}13 \end{bmatrix}$ $\underline{3{,}8087 \times 10^{9}}$	$\begin{bmatrix} 6\,053{,}67 & 3\,384{,}15 & 2\,784{,}75 \\ 5\,008{,}42 & 5\,175{,}63 & 3\,681{,}88 \\ 4\,421{,}45 & 5\,315{,}40 & 3\,925{,}26 \end{bmatrix}$ $\underline{3{,}4807 \times 10^{9}}$
3	$\begin{bmatrix} 6\,418{,}76 & 2\,752{,}01 & 2\,184{,}68 \\ 5\,303{,}69 & 3\,790{,}42 & 2\,602{,}81 \\ 4\,726{,}19 & 4\,185{,}14 & 2\,849{,}74 \end{bmatrix}$ $\underline{1{,}0283 \times 10^{9}}$	$\begin{bmatrix} 5\,589{,}50 & 2\,865{,}88 & 2\,292{,}22 \\ 4\,877{,}85 & 4\,224{,}71 & 2\,932{,}37 \\ 4\,510{,}29 & 4\,549{,}42 & 3\,332{,}35 \end{bmatrix}$ $\underline{2{,}6326 \times 10^{9}}$
4	$\begin{bmatrix} 5\,888{,}39 & 2\,372{,}97 & 1\,924{,}39 \\ 5\,169{,}59 & 3\,085{,}29 & 2\,184{,}82 \\ 4\,773{,}33 & 3\,411{,}38 & 2\,342{,}53 \end{bmatrix}$ $\underline{2{,}7765 \times 10^{8}}$	$\begin{bmatrix} 5\,128{,}65 & 2\,388{,}92 & 1\,928{,}00 \\ 4\,698{,}35 & 3\,511{,}09 & 2\,359{,}65 \\ 4\,376{,}84 & 3\,819{,}96 & 2\,824{,}63 \end{bmatrix}$ $\underline{2{,}5781 \times 10^{9}}$
5	$\begin{bmatrix} 5\,371{,}43 & 2\,064{,}44 & 1\,705{,}91 \\ 4\,915{,}27 & 2\,536{,}47 & 1{,}866{,}94 \\ 4\,652{,}63 & 2\,776{,}76 & 1\,966{,}34 \end{bmatrix}$ $\underline{7{,}4966 \times 10^{7}}$	$\begin{bmatrix} 4\,602{,}27 & 2\,096{,}25 & 1\,825{,}86 \\ 4\,549{,}36 & 2\,770{,}88 & 1\,835{,}51 \\ 4\,211{,}30 & 3\,140{,}85 & 2\,009{,}68 \end{bmatrix}$ $\underline{9{,}1749 \times 10^{8}}$

Таблица 50.4

Сериальные и взаимные корреляции ряда таблицы 50.2

Порядок корреляций	Корреляции (значения дробной части)								
	Ряд 1 авто	Ряд 2 авто	Ряд 3 авто	2 впереди 1	3 впереди 1	1 впереди 2	3 впереди 2	1 впереди 3	2 впереди 3
0	1000	1000	1000	598	459	598	855	459	855
1	933	898	787	677	551	512	929	389	693
2	860	769	600	714	635	434	825	341	534
3	782	617	496	697	655	354	697	258	408
4	707	500	404	671	638	278	574	194	307
5	617	377	257	650	614	229	457	177	218
6	525	276	169	603	612	190	309	127	161
7	436	201	113	544	557	146	213	080	113
8	353	159	080	445	476	102	152	072	059
9	281	093	054	351	392	072	126	031	023
10	227	053	016	287	304	033	066	—018	—025
11	201	—014	—026	233	266	—012	034	—058	—116
12	174	—090	—099	155	219	—070	001	—085	—179
13	117	—177	—116	086	144	—116	—046	—143	—232
14	068	—258	—110	015	083	—172	—091	—216	—283
15	005	—316	—162	—036	031	—252	—159	—271	—350
16	—045	—361	—260	—096	—006	—318	—218	—305	—374
17	—110	—404	—276	—120	—059	—362	—253	—357	—388
18	—179	—419	—271	—137	—063	—416	—270	—425	—397
19	—259	—438	—234	—192	—049	—471	—276	—474	—412
20	—323	—462	—330	—279	—094	—516	—291	—493	—429
21	—381	—459	—382	—345	—159	—535	—358	—501	—482
22	—434	—535	—414	—391	—214	—545	—411	—497	—493
23	—470	—548	—401	—431	—259	—538	—448	—507	—485
24	—467	—531	—406	—485	—321	—517	—459	—465	—444
25	—473	—476	—397	—501	—345	—476	—449	—457	—407

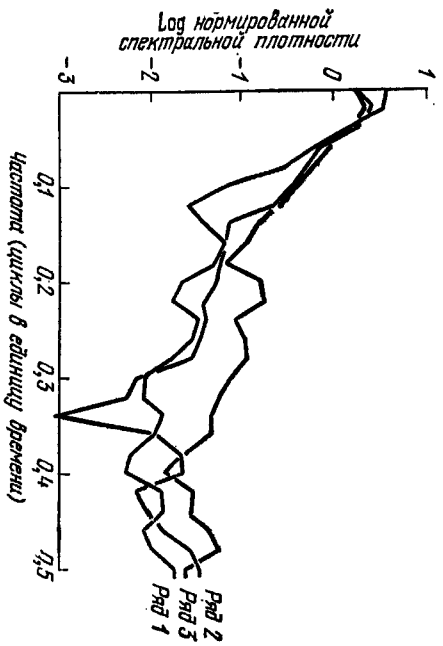

Рис. 50.2. Спектральные функции трех рядов из таблицы 50.2.
По ординате отложен десятичный логарифм спектральной плотности, деленной на дисперсию ряда.

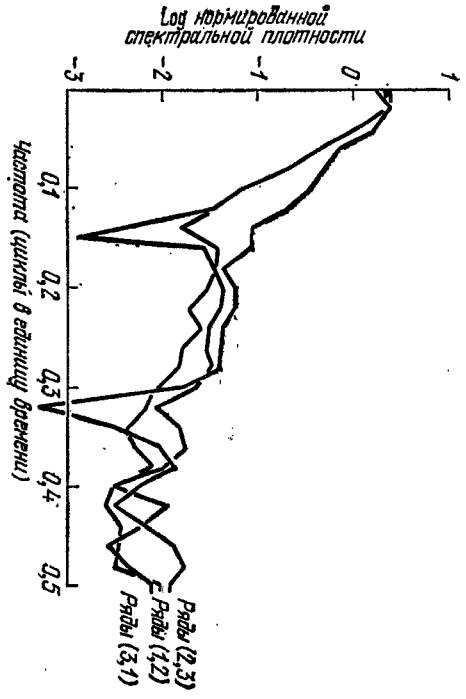

Рис. 50.3. Амплитуда взаимного спектра трех рядов из таблицы 50.2.
По ординате отложен десятичный логарифм взаимной спектральной плотности, деленной на квадратный корень из произведения дисперсий соответствующих рядов.

Рассмотрим, например, значения $|c_s|$ и $|\gamma_s|$ при $s = 0, \ldots, 5$ для ряда примера 50.7:

s	$\lvert\gamma_s\rvert$	$\lvert c_s\rvert$
0	$5{,}225 \times 10^{10}$	$3{,}832 \times 10^{10}$
1	$1{,}411 \times 10^{10}$	$8{,}240 \times 10^{9}$
2	$3{,}809 \times 10^{9}$	$3{,}481 \times 10^{9}$
3	$1{,}028 \times 10^{9}$	$2{,}633 \times 10^{9}$
4	$2{,}777 \times 10^{8}$	$2{,}578 \times 10^{9}$
5	$7{,}497 \times 10^{7}$	$9{,}175 \times 10^{8}$

Флуктуации слишком велики, чтобы получить хорошее представление о связях.

50.35 Можно также рассматривать отношения $\gamma_s\gamma_{s-1}^{-1}$. Например, согласно (50.113), мы должны были бы ожидать, что для схемы Маркова выборочные значения определителей таких матриц стремятся к постоянной. Однако и здесь флуктуации значительны.

Еще ряд результатов в этой области читатель может найти у Бартлетта и Раджалакшмана (1953) и в монографии Кенуя, где критерий, рассмотренный в **50.9**, обобщается на многомерный случай.

Системы уравнений

50.36 Обычно при построении математической модели системы мы вводим ряд зависимостей различного типа между характеристиками системы. Простоты ради предположим, что эти зависимости даются уравнениями (а не, например, неравенствами) и линейны. На практике это условие не столь ограничительно, как это может показаться, — иногда мы можем избавиться от нелинейности, преобразуя переменные, иногда нелинейные зависимости можно заменить линейными аналогично тому, как кривую можно приблизить кусочно линейной функцией.

50.37 За исключением физических систем, точные детерминированные зависимости на практике встречаются редко. В типичной ситуации в линейном соотношении между величинами присутствует член, характеризующий «ошибку». Рассмотрим, например, случай, когда между наблюдаемыми величинами y и x предполагается простая зависимость

$$y = \beta x. \tag{50.117}$$

Соотношение (50.117) может быть неточным по крайней мере по трем причинам: 1) зависимость между y и x на самом деле нелинейна, 2) наблюдаемые величины подвержены ошибкам наблюдения, и в этом случае истинная зависимость относится к ненаблюдаемым величинам η и ξ, 3) зависимость сама по себе точна, но существуют другие величины, также влияющие на y, так что правильно было бы писать

$$y = \beta x + \varepsilon, \tag{50.118}$$

где ε пока означает нечто неизвестное или, во всяком случае, то, что мы не можем определить точно.

50.38 Уравнение (50.118) описывает структурную, не обязательно стохастическую связь между величинами. Это *не* уравнение регрессии. Однако, сталкиваясь с такими зависимостями на практике, не так уж неразумно предполагать заранее, что ε случайно, и отказываться от такого предположения лишь после того, как будет накоплен материал об истинном характере ε. Мы, кроме того, будем предполагать, что величины y и x не включают в себя ошибок наблюдения.

Таким образом, мы пришли к системам линейных уравнений, которые не включают в себя ошибок наблюдений, но содержат случайные элементы. Наша цель — на основе наблюдений оценить постоянные в этих уравнениях и дисперсии случайных членов. Мы уже рассматривали некоторые системы типа (а) регрессии с независимыми ошибками, (б) авторегрессии с независимыми ошибками, (в) авторегрессии со скользящими средними в качестве ошибок. Перейдем теперь к беглому рассмотрению двух других типов: (г) регрессии с автокоррелированными ошибками, (д) смешанных регрессионно-авторегрессионных систем.

Регрессия с автокоррелированными ошибками

50.39 Этот случай, по-видимому, впервые подробно обсуждался Кокрэйном и Оркуттом (1949), которые указали, что оценки по методу наименьших квадратов не свободны от смещения, если члены, соответствующие ошибкам, коррелированы. Критерий для проверки наличия ненулевых корреляций был построен Дербином и Ватсоном (1950—1). Точные результаты в этом случае получить трудно, однако Дербину и Ватсону удалось построить статистику, ограниченную снизу и сверху статистиками, подчиняющимися распределению Р. Андерсона (**48.8**). См. также работы Ватсона (1955) и Ватсона и Хеннана (1956).

50.40 Рассмотрим регрессию величин y по величинам x:

$$y_t = \beta_1 x_{1t} + \ldots + \beta_q x_{qt} + u_t, \qquad t = 1, \ldots, n. \tag{50.119}$$

Мы пишем здесь u_t вместо обычных ε_t для того, чтобы подчеркнуть, что последние автокоррелированы. Часто, не

совершая серьезной ошибки, можно полагать, что корреляции величин u имеют авторегрессионную структуру:

$$\sum_{0}^{k} \alpha_j u_{t-j} = \varepsilon_t. \tag{50.120}$$

Если бы величины α были известны, мы могли бы привести (50.119) к виду

$$\sum_{j=0}^{k} \alpha_j y_{t-j} = \sum_{l=0}^{q} \beta_l \sum_{j=0}^{k} \alpha_j x_{l,\,t-j} + \sum_{j=0}^{k} \alpha_j u_{t-j}, \tag{50.121}$$

или к виду

$$y'_t = \sum_{1}^{q} \beta_l x'_{lt} + \varepsilon_t, \tag{50.122}$$

где

$$y'_t = \sum_{j=0}^{k} \alpha_j y_{t-j}, \tag{50.123}$$

$$x'_{lt} = \sum_{j=0}^{k} \alpha_j x_{l,\,t-j}. \tag{50.124}$$

Уравнение (50.122) — обычное уравнение регрессии. Кокрэйн и Оркутт (1949), которым и принадлежит это так называемое «авторегрессионное» преобразование, предложили, угадывая каким-то образом значения величин α, оценивать β с помощью (50.122) и, если это необходимо, повторять процесс, перевычисляя остатки и получая дальнейшие приближения к значениям величин α.

Дербин (1960b) предложил иную процедуру, приводящую к асимптотически эффективным оценкам. Положив $\gamma_{lj} = \beta_l \alpha_j$, приводим (50.121) к виду

$$y_t + \sum_{j=1}^{k} \alpha_j y_{t-j} = \sum_{l,\,j} \gamma_{lj} x_{l,\,t-j} + \varepsilon_t. \tag{50.125}$$

Если бы все γ были независимы, мы могли бы, как будет показано ниже, рассматривать это асимптотически, как регрессию y_t по остальным y и x, и получать оценки α и γ по методу наименьших квадратов. Если a, b, c — соответствующие оценки α, β, γ, в силу (50.119) получаем, что

$$y_t + \sum_{j=1}^{k} a_j y_{t-j} - \sum c_{lj} x_{l,\,t-j} = \\ = y_t + \sum a_j u_{t-j} - \sum (c_{lj} - a_j \beta_l) x_{l,\,t-j}. \tag{50.126}$$

Следовательно, все a и $(c - a\beta)$ — полученные методом наименьших квадратов коэффициенты регрессии по u_{t-j} и $x_{l,\,t-j}$.

Следовательно, $a_j - \alpha_j$ и $c_{lj} - \alpha_j \beta_l$ асимптотически нормальны с нулевым средним и матрицей рассеяния, которую можно определить. Поэтому мы можем выписать их функцию правдоподобия и, максимизируя последнюю, получить оценки всех α и β.

В некоторых случаях асимптотически эффективными являются оценки метода наименьших квадратов, полученные просто из (50.119) (см. Гренандер и Розенблатт (1957) и Р. Андерсон и Т. Андерсон (1950)). Однако в этом случае становятся непригодными критерии для проверки гипотез.

Смешанная регрессионно-авторегрессионная схема

50.41 Рассмотрим теперь случай регрессии авторегрессионно связанных y_k по фиксированным x:

$$\sum_{j=0}^{k} \alpha_j y_{t-j} = \sum_{l=1}^{q} \beta_l x_{lt} + \varepsilon_t. \tag{50.127}$$

Последнее можно привести к виду, сходному с (50.126):

$$y_t = -\sum_{j=1}^{k} \alpha_j y_{t-j} + \sum_{l=1}^{q} \beta_l x_{lt} + \varepsilon_t. \tag{50.128}$$

Это уже не регрессия с фиксированными величинами справа, поскольку в схеме появились «запаздывающие» y. Дербин (1960a) показал, что независимо от того, распределены остаточные члены нормально или нет, появление «запаздывающих величин» не влияет на асимптотические свойства оценок, полученных методом наименьших квадратов. Это естественное обобщение теоремы Манна — Вальда, уже упоминавшейся в **50.7**.

50.42 Мы не имеем здесь возможности развивать дальше теорию оценивания и проверки гипотез в статистических моделях, наиболее важное направление которой к тому же изобилует многочисленными подводными камнями. Однако некоторые общие замечания сделать необходимо.

(a) Важно помнить, какие величины мы считаем фиксированными, а какие — в силу ли их собственной природы или из модельных соображений — случайными. Это, в частности, важно, если мы начинаем преобразовывать наши уравнения. Например, если мы будем обозначать случайные величины строчными буквами, а фиксированные — заглавными, то регрессия

$$y = \beta X + \varepsilon \tag{50.129}$$

отнюдь не совпадает с регрессией

$$x = \frac{1}{\beta} Y - \frac{1}{\beta} \varepsilon. \tag{50.130}$$

(б) Интересно, что трактовка величины u_t зависит от того, в какой форме она появляется в схеме. В уравнении

$$u_t = \rho u_{t-1} + \varepsilon_t \tag{50.131}$$

мы обычно рассматриваем и u_t и u_{t-1} как случайные величины. Однако относительно момента времени t момент времени $(t-1)$ уже «прошел» и u_{t-1} известно. Таким образом, в определенном смысле u_{t-1} не случайно. Например, если уравнение (50.131) играет роль «предсказывающего», нас интересует условная величина $u_t | u_{t-1}$, а не совместное распределение величин u_t и u_{t-1}.

(в) Ясно, и к этому усиленно привлекал внимание Хаавелмо (1943), что оценивание постоянных лишь в некоторых уравнениях, а не во всей системе, приводит к смещению. Таким образом, имеется еще один источник ошибок, о котором не следует забывать: мы можем упустить из внимания часть модели.

(г) Иногда характер имеющихся в распоряжении исследователя данных может спровоцировать на построение некорректных моделей.

Так, спрос на товары определяет их цену, цена определяет предложение. Но записать соотношения типа

$$C_t = f(Ц_t), \qquad Ц_t = f(П_t) \tag{50.132}$$

— значит проглядеть основное свойство системы, состоящее в том, что существует разрыв во времени между изменением одной величины и изменением другой. Запаздывание может быть столь небольшим, что оно не окажет влияния на статистический материал, который мы в состоянии собрать, но игнорировать указанный факт — значит, уменьшить полезность модели.

50.43 Скандинавская школа во главе с Волдом (1964) настаивает на том, что при построении экономических моделей следует ограничиваться так называемыми моделями типа «причинных цепочек». Такой подход, уместный и при решении общих задач анализа динамических систем, связан с явлениями, порождаемыми некоей цепочкой причинных связей. Поведение величины, которую мы изучаем (наблюдаемая величина), подчинено причинным влияниям, которые характеризуются некоторым числом независимых переменных, характер и значения которых нам известны. Наблюдаемые величины могут зависеть от других наблюдаемых нами величин, но характер зависимости нам также должен быть известен. Теоретически уравнения, выражающие указанные зависимости, должны, по-видимому, отражать запаздывание во времени. Если это сделать невозможно, уравнения должны рассматриваться как несимметричные и читаться слева направо. Например, простые линейные уравнения, указывающие на зависимость цены от спроса —

$$Ц = \alpha C, \tag{50.133}$$

— нельзя обращать, получая, что

$$C = \frac{1}{\alpha} Ц. \tag{50.134}$$

Литература по построению моделей разрозненна, разноречива и неполна. Со статистическим анализом моделей дело обстоит еще хуже. Монография Фиска (1967) дает полезный обзор задач, связанных с системами уравнений. О методе причинных цепочек можно прочесть в сборнике статей Волда (1964). См. также сборник статей C—E—I—R (1968).

Прогнозирование

50.44 Одна из основных целей анализа временных рядов состоит в прогнозировании поведения изучаемой системы на некоторый исследуемый период времени; по крайней мере, мы должны уметь определять, возможно ли предсказание при принятом нами уровне ошибок. Известны два подхода к задаче. Первый носит чисто статистический характер: изучается прошлое поведение ряда, и в предположении, что система не меняется во времени, предпринимается попытка экстраполяции ряда на будущее без детального изучения самой системы. Так, например, оценив постоянные авторегрессионного ряда, мы можем написать, что

$$u_t = -\hat{\alpha}_1 u_{t-1} - \hat{\alpha}_2 u_{t-2} + \varepsilon_t, \tag{50.135}$$

и оценить u_t, во-первых, подставляя в это уравнение известные значения u_{t-1} и u_{t-2}; во-вторых, предполагая, что наилучшая оценка возмущающего члена ε_t, состоит в приравнивании его к нулю. Если на основании прошлого опыта мы оценили дисперсии ε и оценок величин α_1 и α_2, мы можем построить доверительные интервалы для u_t.

50.45 Этот откровенно эмпирический подход основан на следующих предположениях: (а) наша система такова, что схема авторегрессии (или какая-то другая выбранная схема) хорошо отражает истинный механизм явления, (б) этот механизм не меняется или, во всяком случае, не меняется столь быстро, чтобы сделать неверным предположение о том, что уравнение, основанные на прошлом опыте, будут отражать и поведение системы в будущем. Если мы, однако, хотим глубже понять природу явления, мы должны построить модель самой системы, т. е. мы должны постараться определить в специфической форме связи, обусловливающие движение системы. Это более сложная

задача, требующая, с одной стороны, бо́льшего проникновения в причинные механизмы явления, с другой — больших усилий при оценивании различных параметров. Статистики обычно предпочитают более простой подход и «экстраполируют будущее» на основе прошлого, не предпринимая попыток построить модель самого явления. Это приводит к хорошим результатам при прогнозировании на короткий срок. Но такой подход не дает возможности предсказать, что произойдет, если характер системы изменится.

50.46 Если бы обнаружилось, что схема авторегрессии или регрессионная схема хорошо соответствуют прошлому опыту, осуществить прогноз было бы легко. Мы просто использовали бы аутентичные связи, чтобы предсказать будущее значения. Это можно сделать для временных рядов любого типа. Если ряд разлагается на такие компоненты, как тренд, сезонные изменения и осцилляции, мы можем предсказывать будущее каждой компоненты и, воссоединяя их, предсказывать будущее всего ряда. Как мы уже раньше отмечали, здесь основным является предположение о взаимной независимости компонент.

50.47 На практике схемы второго порядка часто оказываются вполне удовлетворительными, т. е. мало что достигается добавлением дополнительных членов. Очень большое внимание уделялось случаю схем первого порядка, т. е. схеме Маркова. В этом случае «предсказывающие» уравнения очень просты, хотя, возможно, и слишком просты. Холт (1957) предложил эвристический подход, имеющий некоторые привлекательные особенности.

Рассмотрим схему авторегрессии типа

$$u_{t+1} = \alpha u_t + \alpha(1-\alpha) u_{t-1} + \alpha(1-\alpha)^2 u_{t-2} + \ldots$$
$$\ldots + \alpha(1-\alpha)^k u_{t-k} + \varepsilon_{t+1}. \qquad (50.136)$$

Интуитивно кажется, что при $|1-\alpha| < 1$ эта схема весьма удобна для прогноза, так как вклад прошлых членов в формирование u_{t+1} уменьшается при переходе к более давним значениям времени. Если мы оценим u_{t+1} систематической компонентой $\hat{u}_{t+1}$ из правой части (50.136), т. е. будем игнорировать ε_{t+1}, получим, что

$$\hat{u}_{t+1} - \hat{u}_t = \alpha \sum_0^k (1-\alpha)^j u_{t-j} - \alpha \sum_0^k (1-\alpha)^j u_{t-j-1} =$$
$$= \alpha(u_t - \hat{u}_t) - (1-\alpha)^{k+1} u_{t-k-1}. \qquad (50.137)$$

При значениях $|1-\alpha|$, не слишком близких к единице, и умеренно больших k, можно написать, что

$$\hat{u}_{t+1} - \hat{u}_t = \alpha(u_t - \hat{u}_t) = \alpha\varepsilon_t. \qquad (50.138)$$

Допустим, что α известно. В момент $t+1$ мы знаем ε_t. Это последнее было ошибкой оценки в момент t. Тогда мы можем оценить u_{t+1}, взяв оценку для t и добавив $\alpha\varepsilon_t$.

50.48 Оценить параметр α не так просто. Наиболее непосредственный и требующий больших вычислений подход состоит в попытке оценить «размах» значений α и вычислении сумм

$$\sum \{u_{t+1} - \alpha u_t - \ldots - \alpha(1-\alpha)^k u_{t-k}\}^2$$

для различных k. При этом α и k выбираются так, чтобы минимизировать последнее выражение.

Такая процедура не предусматривает большой точности при определении оптимального α.

Системы типа (50.136) называют, по понятным причинам, предсказывающими (или упреждающими) системами, использующими скользящие средние с экспоненциальными весами. Более детально они изучались Брауном (1959), Барнардом Г. (1959), Коксом (1961), Боксом и Дженкинсом (1962) и Уордом (1963). Уинтерс (1960) обобщил результаты на случай, когда имеются сезонные изменения.

50.49 На этом мы должны кончить. Мы сознаем, что некоторые вопросы можно было бы обсуждать и дольше, а многие еще требуют дальнейшего изучения. Мы надеемся, что читатель, который смог выдержать до конца, сможет простить нам и недостатки нашей работы. Совсем не легко придать стройность и законченность изложению предмета, который претерпевает столь бурное развитие. И об этом не стоит жалеть — появляются новые результаты, и их появление обусловлено жизнеспособностью предмета, дальнейшего триумфа которого мы, несомненно, будем свидетелями. Мы в большом долгу перед авторами, чьи работы мы столь вольно излагали, мы просим их простить нас за возможные ошибки и пропуски. И мы пишем эти последние строки с чувством глубокого облегчения.

Библиографические замечания

Теперь стала доступной исчерпывающая «Библиография статистической литературы» М. Д. Кендалла и Алисона Г. Дойга (Оливер и Бойд, Эдинбург) *), включающая 30 000 наименований с 16 века по 1958 г. 1-й том (1962) относится к 1950—58 гг., 2-й том (1965) — к 1940—49 гг., 3-й том (1966) — ко всем годам вплоть до 1939. Ссылки на работы, вышедшие после 1958 года можно найти в Journal of Statistical Abstracts, периодически издаваемом Международным статистическим институтом.

Существует также некоторое число специализированных библиографий. Отметим, в частности, превосходную библиографию, изданную под редакцией Г. Волда (1969).

*) «Bibliography Statistical Literature», M. G. Kendall and Alison G. Doig (Oliver and Boyd, Edinburg).

УПРАЖНЕНИЯ

50.1 Показать, что если V_n — определитель матрицы (50.12), то

$$V_n = (1+\beta^2)\,V_{n-1} - \beta^2 V_{n-2},$$

и, следовательно, $V_n = (1-\beta^{2n+2})/(1-\beta^2)$.

50.2 Показать, что для схемы Маркова (обозначения 50.9)

$$\omega_j = r_j + 2\rho r_{j-1} + \rho^2 r_{j-2}$$

и что, следовательно, ряд примера 50.2 этой схеме не удовлетворяет.

50.3 Показать, что если $G(z)$ — производящая функция автоковариаций величин u, то производящая функция автоковариаций величин η из (50.26) равна

$$\left(\sum_{j=-\infty}^{\infty} \alpha_{k-j} z^j\right)\left(\sum_{j=-\infty}^{\infty} \alpha_{k-j} z^{-j}\right) G(z),$$

и, что, следовательно, ε и η имеют одну и ту же производящую функцию автоковариаций.

50.4 Докажите (50.32)

50.5 (Последовательное решение уравнений Юла — Уокера.)

Схема линейной авторегрессии имеет порядок k, коэффициенты α вычислены в предположении, что порядок схемы равен $s \leqslant k$, что привело к значениям $\alpha_{s1}, \ldots, \alpha_{ss}$, и $\alpha_{11} = -\rho_1$. Показать, что

$$\alpha_{st} = \alpha_{s-1,t} + \alpha_{ss}\alpha_{s-1,s-t}, \qquad t = 1, \ldots, s-1,$$

$$\alpha_{ss} = -\frac{\rho_s + \alpha_{s-1,1}\rho_{s-1} + \alpha_{s-2,2}\rho_{s-2} + \ldots + \alpha_{s-1,s-1}\rho_1}{1 + \alpha_{s-1,1}\rho_1 + \ldots + \alpha_{s-1,s-1}\rho_{s-1}} \qquad s = 1, \ldots k.$$

(Дербин, 1960b.)

50.6 Используя обозначения 50.10, показать, что если функцию правдоподобия записать в виде $(1-\beta^2)^{\frac{1}{2}} f(Q)\,da$ так, что

$$\int f(Q)\,da = (1-\beta^2)^{1/2},$$

то

$$\mathsf{M}\left(\frac{\partial Q}{\partial \beta}\right) = O(n^{-1}), \qquad \mathsf{M}\left(\frac{\partial^2 Q}{\partial \beta^2}\right) = \frac{n}{2}\,\mathsf{M}\left(\frac{\partial Q}{\partial \beta}\right)^2 + O(n^{-1}).$$

Вывести отсюда, что приближенно $\mathsf{D}b = \mathsf{M}\left(\frac{\partial Q}{\beta}\right)^2 \Big/ \mathsf{M}\left(\frac{\partial^2 Q}{\partial \beta^2}\right)$, а затем получить (50.48).

(Дербин, 1959b.)

50.7 Доказать (50.88).

50.8 Показать, что если два ряда типа авторегрессии или скользящих средних порождены одним случайным рядом, то для всех k

$$\sum_{i=-\infty}^{\infty} \rho_{(11)\,i}\rho_{(22)\,k-i} = \sum_{i=-\infty}^{\infty} \rho_{(12)\,i}\rho_{(21)\,k-i}.$$

(Кенуй, 1957.)

50.9 Показать, что, если в примере 50.6 матрицы $\boldsymbol{F}$ являются $\boldsymbol{A}$-матрицами схемы авторегрессии, то лишь одна из них определяет стационарный процесс.

50.10 Пусть матрица $\boldsymbol{B}$ из 50.26 — единичная матрица. Обсудить условия, при которых схема авторегрессии даст однозначно определенное стационарное решение.

ПОСЛЕСЛОВИЕ К ТРЕТЬЕМУ ТОМУ

«Я полагаю разумным, перед тем как подойдет к концу Ваше пребывание в Парламенте, высказать Вам некоторые свои мысли. Это уместно сделать именно сейчас, ибо, уезжая, Вы сможете унести сказанное с собой...

Вы знаете сами, что я никогда не рассматривала ни одного дела, не испросив совета, одинаково чтя и правду, и справедливость. Как должно всем королям, желающим знать истину, я всегда готова была выслушать доводы и pro и contra *). Но, внимая всем спорам, я сама всегда оставалась свободной от предрассудков и предубежденности».

Елизавета I своему последнему Парламенту

*) pro et contra (лат.) — за и против.

ПРИЛОЖЕНИЕ

Таблица 1

Функция плотности нормального распределения

$y = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$ с первыми и вторыми разностями

x	y	Δ^1 (−)	Δ^2	x	y	Δ^1 (−)	Δ^2
0,0	0,39894	199	−392	2,5	0,01753	395	+79
0,1	0,39695	591	−374	2,6	0,01358	316	+66
0,2	0,39104	965	−347	2,7	0,01042	250	+53
0,3	0,38139	1312	−308	2,8	0,00792	197	+45
0,4	0,36827	1620	−265	2,9	0,00595	152	+36
0,5	0,35207	1885	−212	3,0	0,00443	116	+27
0,6	0,33322	2097	−159	3,1	0,00327	89	+23
0,7	0,31225	2256	−104	3,2	0,00238	66	+17
0,8	0,28969	2360	−52	3,3	0,00172	49	+13
0,9	0,26609	2412	−0	3,4	0,00123	36	+10
1,0	0,24197	2412	+46	3,5	0,00087	26	+7
1,1	0,21785	2366	+84	3,6	0,00061	19	+6
1,2	0,19419	2282	+118	3,7	0,00042	13	+4
1,3	0,17137	2164	+143	3,8	0,00029	9	+2
1,4	0,14973	2021	+161	3,9	0,00020	7	+3
1,5	0,12952	1860	+173	4,0	0,00013	4	—
1,6	0,11092	1687	+177	4,1	0,00009	3	—
1,7	0,09405	1510	+177	4,2	0,00006	2	—
1,8	0,07895	1333	+170	4,3	0,00004	2	—
1,9	0,06562	1163	+162	4,4	0,00002	—	—
2,0	0,05399	1001	+150	4,5	0,00002	—	—
2,1	0,04398	851	+137	4,6	0,00001	—	—
2,2	0,03547	714	+120	4,7	0,00001	—	—
2,3	0,02833	594	+108	4,8	0,00000	—	—
2,4	0,02239	486	+91				

Таблица 2

Функция нормального распределения

В таблице даны значения $F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}x^2} dx$, т. е. значения площади под кривой $y = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$, лежащей левее точки x; например, для $x = 1{,}86\ (= 1{,}5 + 0{,}36)$ $F(1{,}86) = 0{,}9686$

x	0,0+	0,5+	1,0+	1,5+	2,0+	2,5+	3,0+	3,5+
0,00	5000	6915	8413	9332	9772	$9^{2}379$	$9^{2}865$	$9^{3}77$
0,01	5040	6950	8438	9345	9778	$9^{2}396$	$9^{2}869$	$9^{3}78$

Продолжение

x	0,0+	0,5+	1,0+	1,5+	2,0+	2,5+	3,0+	3,5+
0,02	5080	6985	8461	9357	9783	$9^{2}413$	$9^{2}874$	$9^{3}78$
0,03	5120	7019	8485	9370	9788	$9^{2}430$	$9^{2}878$	$9^{3}79$
0,04	5160	7054	8508	9382	9793	$9^{2}446$	$9^{2}882$	$9^{2}80$
0,05	5199	7088	8531	9394	9798	$9^{2}461$	$9^{2}886$	$9^{3}81$
0,06	5239	7123	8554	9406	9803	$9^{2}477$	$9^{2}889$	$9^{3}81$
0,07	5279	7157	8577	9418	9808	$9^{2}492$	$9^{2}893$	$9^{3}82$
0,08	5319	7190	8599	9429	9812	$9^{2}506$	$9^{2}897$	$9^{3}83$
0,09	5359	7224	8621	9441	9817	$9^{2}520$	$9^{2}900$	$9^{3}83$
0,10	5398	7257	8643	9452	9821	$9^{2}534$	$9^{3}03$	$9^{3}84$
0,11	5438	7291	8665	9463	9826	$9^{2}547$	$9^{3}06$	$9^{3}85$
0,12	5478	7324	8686	9474	9830	$9^{2}560$	$9^{3}10$	$9^{3}85$
0,13	5517	7357	8708	9484	9834	$9^{2}573$	$9^{3}13$	$9^{3}86$
0,14	5557	7389	8729	9495	9838	$9^{2}585$	$9^{3}16$	$9^{3}86$
0,15	5596	7422	8749	9505	9842	$9^{2}598$	$9^{3}18$	$9^{3}87$
0,16	5636	7454	8770	9515	9846	$9^{2}609$	$9^{3}21$	$9^{3}87$
0,17	5675	7486	8790	9525	9850	$9^{2}621$	$9^{3}24$	$9^{3}88$
0,18	5714	7517	8810	9535	9854	$9^{2}632$	$9^{3}26$	$9^{3}88$
0,19	5753	7549	8830	9545	9857	$9^{2}643$	$9^{3}29$	$9^{3}89$
0,20	5793	7580	8849	9554	9861	$9^{2}653$	$9^{3}31$	$9^{3}89$
0,21	5832	7611	8869	9564	9864	$9^{2}664$	$9^{3}34$	$9^{3}90$
0,22	5871	7642	8888	9573	9868	$9^{2}674$	$9^{3}36$	$9^{3}90$
0,23	5910	7673	8907	9582	9871	$9^{2}683$	$9^{3}38$	$9^{4}04$
0,24	5948	7704	8925	9591	9875	$9^{2}693$	$9^{3}40$	$9^{4}08$
0,25	5987	7738	8944	9599	9878	$9^{2}702$	$9^{3}42$	$9^{4}12$
0,26	6026	7764	8962	9608	9881	$9^{2}711$	$9^{3}44$	$9^{4}15$
0,27	6064	7794	8980	9616	9884	$9^{2}720$	$9^{3}46$	$9^{4}18$
0,28	6103	7823	8997	9625	9887	$9^{2}728$	$9^{3}48$	$9^{4}22$
0,29	6141	7852	9015	9633	9890	$9^{2}736$	$9^{3}50$	$9^{4}25$
0,30	6179	7881	9032	9641	9893	$9^{2}744$	$9^{3}52$	$9^{4}28$
0,31	6217	7910	9049	9649	9896	$9^{2}752$	$9^{3}53$	$9^{4}31$
0,32	6255	7939	9066	9656	9898	$9^{2}760$	$9^{3}55$	$9^{4}33$
0,33	6293	7967	9082	9664	9901	$9^{2}767$	$9^{3}57$	$9^{4}36$
0,34	6331	7995	9099	9671	9904	$9^{2}774$	$9^{3}58$	$9^{4}39$
0,35	6368	8023	9115	9678	9906	$9^{2}781$	$9^{3}60$	$9^{4}41$
0,36	6406	8051	9131	9686	9909	$9^{2}788$	$9^{3}61$	$9^{4}43$
0,37	6443	8078	9147	9693	9911	$9^{2}795$	$9^{3}62$	$9^{4}46$
0,38	6480	8106	9162	9699	9913	$9^{2}801$	$9^{3}64$	$9^{4}48$
0,39	6517	8133	9177	9706	9916	$9^{2}807$	$9^{3}65$	$9^{4}50$
0,40	6554	8159	9192	9713	9918	$9^{2}813$	$9^{3}66$	$9^{4}52$
0,41	6591	8186	9207	9719	9920	$9^{2}819$	$9^{3}68$	$9^{4}54$
0,42	6628	8212	9222	9726	9922	$9^{2}825$	$9^{3}69$	$9^{4}56$
0,43	6664	8238	9236	9732	9925	$9^{2}831$	$9^{3}70$	$9^{4}58$
0,44	6700	8264	9251	9738	9927	$9^{2}836$	$9^{3}71$	$9^{4}59$
0,45	6736	8289	9265	9744	9929	$9^{2}841$	$9^{3}72$	$9^{4}61$
0,46	6772	8315	9279	9750	9931	$9^{2}846$	$9^{3}73$	$9^{4}63$
0,47	6808	8340	9292	9756	9932	$9^{2}851$	$9^{3}74$	$9^{4}64$
0,48	6844	8365	9306	9761	9934	$9^{2}856$	$9^{3}75$	$9^{4}66$
0,49	6879	8389	9319	9767	9936	$9^{2}861$	$9^{3}76$	$9^{4}67$

Примечание. Знаки целой части (т. е. 0) в таблице опущены. Повторяющиеся девятки заменены степенями; например, $9^{3}71$ означает 0.99971].

Таблица 3

Квантили χ^2-распределения

(Заимствованы из таблицы III книги сэра Рональда Фишера *Statistical Methods for Research Workers*, Oliver and Boyd Ltd., Edinburgh.)

$P=1-F$	0,99	0,98	0,95	0,90	0,80	0,70	0,50	0,30	0,20	0,10	0,05	0,02	0,01
$\nu=1$	$0{,}0^3157$	$0{,}0^3628$	$0{,}0^2393$	0,0158	0,0642	0,148	0,455	1,074	1,642	2,706	3,841	5,412	6,635
2	0,0201	0,0404	0,103	0,211	0,446	0,713	1,386	2,408	3,219	4,605	5,991	7,824	9,210
3	0,115	0,185	0,352	0,584	1,005	1,424	2,366	3,665	4,642	6,251	7,815	9,837	11,345
4	0,297	0,429	0,711	1,064	1,649	2,195	3,357	4,878	5,989	7,779	9,488	11,668	13,277
5	0,554	0,752	1,145	1,160	2,343	3,000	4,351	6,064	7,289	9,236	11,070	13,388	15,086
6	0,872	1,134	1,635	2,204	3,070	3,828	5,348	7,231	8,558	10,645	12,592	15,033	16,812
7	1,239	1,564	2,167	2,833	3,822	4,671	6,346	8,383	9,803	12,017	14,067	16,622	18,475
8	1,646	2,032	2,733	3,490	4,594	5,527	7,344	9,524	11,030	13,362	15,507	18,168	20,090
9	2,088	2,532	3,325	4,168	5,380	6,393	8,343	10,656	12,242	14,684	16,919	19,679	21,666
10	2,358	3,059	3,940	4,865	6,179	7,267	9,342	11,781	13,442	15,987	18,307	21,161	23,209
11	3,053	3,609	4,575	5,578	6,989	8,148	10,341	12,899	14,631	17,275	19,675	22,618	24,725
12	3,571	4,178	5,226	6,304	7,807	9,034	11,340	14,011	15,821	18,549	21,026	24,054	26,217
13	4,107	4,765	5,862	7,042	8,634	9,926	12,340	15,119	16,985	19,812	22,362	25,472	27,688
14	4,660	5,368	6,571	7,790	9,467	10,821	13,339	16,222	18,151	21,064	23,685	26,873	29,141
15	5,229	5,985	7,261	8,547	10,307	11,721	14,339	17,322	19,311	22,307	24,996	28,259	30,578
16	5,812	6,614	7,962	9,312	11,152	12,624	15,338	18,418	20,465	23,542	26,296	29,633	32,000
17	6,408	7,255	8,672	10,085	12,002	13,531	16,338	19,511	21,615	24,769	27,587	30,995	33,409
18	7,015	7,906	9,390	10,865	12,857	14,440	17,338	20,601	22,760	25,989	28,869	32,346	34,805
19	7,633	8,567	10,117	11,651	13,716	15,352	18,338	21,689	23,900	27,204	30,144	33,687	36,191
20	8,260	9,237	10,851	12,443	14,578	16,266	19,337	22,775	25,038	28,412	31,410	35,020	37,566
21	8,897	9,915	11,591	13,240	15,445	17,182	20,337	23,858	26,171	29,615	32,671	36,343	38,932
22	9,542	10,600	12,338	14,041	16,314	18,101	21,337	24,939	27,301	30,813	33,924	37,659	40,289
23	10,196	11,293	13,091	14,848	17,187	19,021	22,337	26,018	25,429	32,007	35,172	38,968	41,638
24	10,856	11,992	13,848	15,659	18,062	19,943	23,337	27,096	29,553	33,196	36,415	40,270	42,980
25	11,524	12,697	14,611	16,473	18,940	20,867	24,337	28,172	30,675	34,382	37,652	41,566	44,314
26	12,198	13,409	15,379	17,292	19,820	21,792	25,336	29,246	31,795	35,563	38,885	42,856	45,642
27	12,879	14,125	16,151	18,114	20,703	22,719	26,336	30,319	32,912	36,741	39,113	44,140	46,963
28	13,565	14,847	16,928	18,939	21,588	23,647	27,336	31,391	34,027	37,916	41,337	45,419	48,278
29	14,256	15,574	17,708	19,768	22,475	24,577	28,336	32,461	35,139	39,087	42,557	46,693	49,588
30	14,953	16,306	18,493	20,599	23,364	25,508	29,336	33,520	36,250	40,256	43,773	47,962	50,892

Примечание. Для значений ν, бо́льших 30, величину $\sqrt{2\chi}$ можно считать распределенной нормально со средним $\sqrt{2\nu-1}$ и единичной дисперсией.

Таблица 4а

Функция χ^2-распределения с одной степенью свободы

для $0 \leqslant \chi^2 \leqslant 1$ с шагом 0,01

χ^2	P	Λ	χ^2	P	Λ
0	1,00000	7966	0,50	0,47950	436
0,01	0,92034	3280	0,51	0,47514	430
0,02	0,88754	2505	0,52	0,47084	423
0,03	0,86249	2101	0,53	0,46661	418
0,04	0,84148	1842	0,54	0,46243	411
0,05	0,82306	1656	0,55	0,45832	406
0,06	0,80650	1516	0,56	0,45426	400
0,07	0,79134	1404	0,57	0,45026	395
0,08	0,77730	1312	0,58	0,44631	389
0,09	0,76418	1235	0,59	0,44242	384
0,10	0,75183	1169	0,60	0,43858	379
0,11	0,74014	1111	0,61	0,43479	374
0,12	0,72903	1060	0,62	0,43105	369
0,13	0,71843	1015	0,63	0,42736	365
0,14	0,70828	974	0,64	0,42371	360
0,15	0,69854	938	0,65	0,42011	355
0,16	0,68916	905	0,66	0,41656	351
0,17	0,68011	874	0,67	0,41305	346
0,18	0,67137	845	0,68	0,40959	343
0,19	0,66292	820	0,69	0,40616	338
0,20	0,65472	795	0,70	0,40278	334
0,21	0,64677	773	0,71	0,39944	330
0,22	0,63904	752	0,72	0,39614	326
0,23	0,63152	731	0,73	0,39288	322
0,24	0,62421	733	0,74	0,38966	318
0,25	0,61708	696	0,75	0,38648	315
0,26	0,61012	679	0,76	0,38333	311
0,27	0,60333	663	0,77	0,38022	308
0,28	0,59670	648	0,78	0,37714	304
0,29	0,59022	634	0,79	0,37410	301
0,30	0,58388	620	0,80	0,37109	297
0,31	0,57768	607	0,81	0,36812	294
0,32	0,57161	595	0,82	0,36518	291
0,33	0,56566	583	0,83	0,36227	287
0,34	0,55983	572	0,84	0,35940	285
0,35	0,55411	560	0,85	0,35655	281
0,36	0,54851	551	0,86	0,35374	278
0,37	0,54300	540	0,87	0,35096	276
0,38	0,53760	530	0,88	0,34820	272
0,39	0,53230	521	0,89	0,34548	270
0,40	0,52709	512	0,90	0,34278	267
0,41	0,52197	503	0,91	0,34011	264
0,42	0,51694	495	0,92	0,33747	261
0,43	0,51199	487	0,93	0,33486	258
0,44	0,50712	479	0,94	0,33228	256
0,45	0,50233	471	0,95	0,32972	253
0,46	0,49762	463	0,96	0,32719	251
0,47	0,49299	457	0,97	0,32468	248
0,48	0,48842	449	0,98	0,32220	246
0,49	0,48393	443	0,99	0,31974	243
0,50	0,47950	436	1,00	0,31731	241

Таблица 4б

Функция χ^2-распределения с одной степенью свободы
для $1 \leqslant \chi^2 \leqslant 10$ с шагом 0,1

χ^2	P	Δ	χ^2	P	Δ
1,0	0,31731	2304	5,5	0,01902	106
1,1	0,29427	2095	5,6	0,01796	99
1,2	0,27332	1911	5,7	0,01697	94
1,3	0,25421	1749	5,8	0,01603	89
1,4	0,23672	1605	5,9	0,01514	83
1,5	0,22067	1477	6,0	0,01431	79
1,6	0,20590	1361	6,1	0,01352	74
1,7	0,19229	1258	6,2	0,01278	71
1,8	0,17971	1163	6,3	0,01207	66
1,9	0,16808	1078	6,4	0,01141	62
2,0	0,15730	1000	6,5	0,01079	59
2,1	0,14730	929	6,6	0,01020	56
2,2	0,13801	864	6,7	0,00964	52
2,3	0,12937	803	6,8	0,00912	50
2,4	0,12134	749	6,9	0,00862	47
2,5	0,11385	699	7,0	0,00815	44
2,6	0,10686	651	7,1	0,00771	42
2,7	0,10035	609	7,2	0,00729	39
2,8	0,09426	568	7,3	0,00690	38
2,9	0,08858	532	7,4	0,00652	35
3,0	0,08326	497	7,5	0,00617	33
3,1	0,07829	465	7,6	0,00584	32
3,2	0,07364	436	7,7	0,00552	30
3,3	0,06928	408	7,8	0,00522	28
3,4	0,06520	383	7,9	0,00494	26
3,5	0,06137	359	8,0	0,00468	25
3,6	0,05778	337	8,1	0,00443	24
3,7	0,05441	316	8,2	0,00419	23
3,8	0,05125	296	8,3	0,00396	21
3,9	0,04829	279	8,4	0,00375	20
4,0	0,04550	262	8,5	0,00355	19
4,1	0,04288	246	8,6	0,00336	18
4,2	0,04042	231	8,7	0,00318	17
4,3	0,03811	217	8,8	0,00301	16
4,4	0,03594	205	8,9	0,00285	15
4,5	0,03389	192	9,0	0,00270	14
4,6	0,03197	181	9,1	0,00256	14
4,7	0,03016	170	9,2	0,00242	13
4,8	0,02846	160	9,3	0,00229	12
4,9	0,02686	151	9,4	0,00217	12
5,0	0,02535	142	9,5	0,00205	10
5,1	0,02393	134	9,6	0,00195	11
5,2	0,02259	126	9,7	0,00184	10
5,3	0,02133	119	9,8	0,00174	9
5,4	0,02014	112	9,9	0,00165	8
5,5	0,01902	106	10,0	0,00157	8

Таблица 5

Квантили t-распределения

Заимствованы из таблиц сэра Рональда Фишера и Ф. Изйтса: *Statistical Tables for Biological, Medical and Agricultural Research*, Oliver and Boyd Ltd., Edinburgh.)

$P=2(1-F)$	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,05	0,02	0,01	0,001
$\nu=1$	0,158	0,325	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	31,599
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	12,924
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,260	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,129	0,260	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,128	0,259	0,395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,128	0,259	0,394	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,128	0,258	0,393	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,128	0,258	0,393	0,536	0,691	0,866	0,074	1,341	1,753	2,131	2,602	2,947	4,073

Продолжение

$P=2(1-F)$	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,05	0,02	0,01	0,001
$\nu=16$	0,128	0,258	0,392	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,128	0,257	0,392	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,127	0,257	0,392	0,534	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,127	0,257	0,391	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,127	0,257	0,391	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,127	0,257	0,391	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,127	0,256	0,390	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,127	0,256	0,390	0,532	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807	3,767
24	0,127	0,256	0,390	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,127	0,256	0,389	0,531	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771	3,690
28	0,127	0,256	0,389	0,530	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756	3,659
30	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,646
40	0,126	0,255	0,388	0,529	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704	3,551
60	0,126	0,254	0,387	0,527	0,679	0,848	1,046	1,296	1,671	2,000	2,390	2,660	3,460
120	0,126	0,254	0,386	0,526	0,677	0,845	1,041	1,289	1,658	1,980	2,358	2,617	3,373
∞	0,126	0,253	0,385	0,524	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576	3,291

Таблица 6

5%-ные точки z-распределении

(точки, в которых ф. р. = 0,95). (Заимствованы из таблицы VI книги сэра Рональда Фишера: *Statistical Methods for Research Workers*, Oliver and Boyd Ltd., Edinburgh.)

Значения ν_2	Значения ν_1: 1	2	3	4	5	6	8	12	24	∞
1	2,5421	2,6479	2,6870	2,7071	2,7194	2,7276	2,7380	2,7484	2,7588	2,7693
2	1,4592	1,4722	1,4765	1,4787	1,4800	1,4808	1,4819	1,4830	1,4840	1,4851
3	1,1577	1,1284	1,1137	1,1051	1,0994	1,0953	1,0899	1,0842	1,0781	1,0716
4	1,0212	0,9690	0,9429	0,9272	0,9168	0,9093	0,8993	0,8885	0,8767	0,8639
5	0,9441	0,8777	0,8441	0,8236	0,8097	0,7997	0,7862	0,7714	0,7550	0,7368
6	0,8948	0,8188	0,7798	0,7558	0,7394	0,7274	0,7112	0,6931	0,6729	0,6499
7	0,8606	0,7777	0,7347	0,7080	0,6896	0,6761	0,6576	0,6369	0,6134	0,5862
8	0,8355	0,7475	0,7014	0,6725	0,6525	0,6378	0,6175	0,5945	0,5682	0,5371
9	0,8163	0,7242	0,6757	0,6450	0,6238	0,6080	0,5862	0,5613	0,5324	0,4979
10	0,8012	0,7058	0,6553	0,6232	0,6009	0,5843	0,5611	0,5346	0,5035	0,4657
11	0,7889	0,6909	0,6387	0,6055	0,5822	0,5648	0,5406	0,5126	0,4795	0,4387
12	0,7788	0,6786	0,6250	0,5907	0,5666	0,5487	0,5234	0,4941	0,4592	0,4156
13	0,7703	0,6682	0,6134	0,5783	0,5535	0,5350	0,5089	0,4785	0,4419	0,3957
14	0,7630	0,6594	0,6036	0,5677	0,5423	0,5233	0,4964	0,4649	0,4269	0,3782
15	0,7568	0,6518	0,5950	0,5585	0,5326	0,5131	0,4855	0,4532	0,4138	0,3628

Продолжение

Значения ν_2	Значения ν_1: 1	2	3	4	5	6	8	12	24	∞
16	0,7514	0,6451	0,5876	0,5505	0,5241	0,5042	0,4760	0,4428	0,4022	0,3490
17	0,7466	0,6393	0,5811	0,5434	0,5166	0,4964	0,4676	0,4337	0,3919	0,3366
18	0,7424	0,6341	0,5753	0,5371	0,5099	0,4894	0,4602	0,4255	0,3827	0,3253
19	0,7386	0,6295	0,5701	0,5315	0,5040	0,4832	0,4535	0,4182	0,3743	0,3151
20	0,7352	0,6254	0,5654	0,5265	0,4986	0,4776	0,4474	0,4116	0,3668	0,3057
21	0,7322	0,6216	0,5612	0,5219	0,4938	0,4725	0,4420	0,4055	0,3599	0,2971
22	0,7294	0,6182	0,5574	0,5178	0,4894	0,4679	0,4370	0,4001	0,3536	0,2892
23	0,7269	0,6151	0,5540	0,5140	0,4854	0,4636	0,4325	0,3950	0,3478	0,2818
24	0,7246	0,6123	0,5508	0,5106	0,4817	0,4598	0,4283	0,3904	0,3425	0,2749
25	0,7225	0,6097	0,5478	0,5074	0,4783	0,4562	0,4244	0,3862	0,3376	0,2685
26	0,7205	0,6073	0,5451	0,5045	0,4752	0,4529	0,4209	0,3823	0,3330	0,2625
27	0,7187	0,6051	0,5427	0,5017	0,4723	0,4499	0,4176	0,3786	0,3287	0,2569
28	0,7171	0,6030	0,5403	0,4992	0,4696	0,4471	0,4146	0,3752	0,3248	0,2516
29	0,7155	0,6011	0,5382	0,4969	0,4671	0,4444	0,4117	0,3720	0,3211	0,2466
30	0,7141	0,5994	0,5362	0,4947	0,4648	0,4420	0,4090	0,3691	0,3176	0,2419
60	0,6933	0,5738	0,5073	0,4632	0,4311	0,4064	0,3702	0,3255	0,2654	0,1644
∞	0,6729	0,5486	0,4787	0,4319	0,3974	0,3706	0,3309	0,2894	0,2085	0

Таблица 7

5%-ные точки *F*-распределения (точки, в которых ф. р. = 0,95)
(Заимствованы из таблиц сэра Рональда Фишера и Ф. Иэйтса: *Statistical Tables for Biological, Medical and Agricultural Research*, Oliver and Boyd Ltd., Edinburgh.)

ν_2 \ ν_1	1	2	3	4	5	6	8	12	24	∞
1	161,40	199,50	215,70	224,60	230,20	234,00	238,90	243,90	249,00	254,30
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,45	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,64	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,77	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	3,84	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,12	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,07	2,90	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,91	2,74	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,79	2,61	2,40
12	4,75	3,88	3,49	3,26	3,11	3,00	2,85	2,69	2,50	2,30
13	4,67	3,80	3,41	3,18	3,02	2,92	2,77	2,60	2,42	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,53	2,35	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,48	2,29	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,42	2,24	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,38	2,19	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,34	2,15	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,31	2,11	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,28	2,08	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,42	2,25	2,05	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2,23	2,03	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,38	2,20	2,00	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,36	2,18	1,98	1,73
25	4,24	3,38	2,99	2,76	2,60	2,49	2,34	2,16	1,96	1,71
26	4,22	3,37	2,98	2,74	2,59	2,47	2,32	2,15	1,95	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,30	2,13	1,93	1,67
28	4,20	3,34	2,95	2,71	2,56	2,44	2,29	2,12	1,91	1,65
29	4,18	3,33	2,93	2,70	2,54	2,43	2,28	2,10	1,90	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,09	1,89	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,00	1,79	1,51
60	4,00	3,15	2,76	2,52	2,37	2,25	2,10	1,92	1,70	1,39
120	3,92	3,07	2,68	2,45	2,29	2,17	2,02	1,83	1,61	1,25
∞	3,84	2,99	2,60	2,37	2,21	2,09	1,94	1,75	1,52	1,00

Примечание. Нижняя 5%-ная точка представляет собой обратную величину того табличного значения, которое получится, если ν_1 и ν_2 поменять местами, т. е. в качестве ν_1 всегда следует выбирать количество степеней свободы, соответствующее наибольшей выборочной дисперсии.

Таблица 8

1%-ные точки z-распределения (точки, в которых ф. р. = 0,99)

(Заимствованы из таблицы VI книги сэра Рональда Фишера: *Statistical Methods for Research Workers*, Oliver and Boyd Ltd., Edinburgh.)

Значения ν_2	Значения ν_1: 1	2	3	4	5	6	8	12	24	∞
1	4,1535	4,2585	4,2974	4,3175	4,3297	4,3379	4,3482	4,3585	4,3689	4,3794
2	2,2950	2,2976	2,2984	2,2988	2,2991	2,2992	2,2994	2,2997	2,2999	2,3001
3	1,7649	1,7140	1,6915	1,6786	1,6703	1,6645	1,6569	1,6489	1,6404	1,6314
4	1,5270	1,4452	1,4075	1,3856	1,3711	1,3609	1,3473	1,3327	1,3170	1,3000
5	1,3943	1,2929	1,2449	1,2164	1,1974	1,1838	1,1656	1,1457	1,1239	1,0997
6	1,3103	1,1955	1,1401	1,1068	1,0843	1,0680	1,0460	1,0218	0,9948	0,9643
7	1,2526	1,1281	1,0672	1,0300	1,0048	0,9864	0,9614	0,9335	0,9020	0,8658
8	1,2106	1,0787	1,0135	0,9734	0,9459	0,9259	0,8983	0,8673	0,8319	0,7904
9	1,1786	1,0411	0,9724	0,9299	0,9006	0,8791	0,8494	0,8157	0,7769	0,7305
10	1,1535	1,0114	0,9399	0,8954	0,8646	0,8419	0,8104	0,7744	0,7324	0,6816
11	1,1333	0,9874	0,9136	0,8674	0,8354	0,8116	0,7785	0,7405	0,6958	0,6408
12	1,1166	0,9677	0,8919	0,8443	0,8111	0,7864	0,7520	0,7122	0,6649	0,6061
13	1,1027	0,9511	0,8737	0,8248	0,7907	0,7652	0,7295	0,6882	0,6386	0,5761
14	1,0909	0,9370	0,8581	0,8082	0,7732	0,7471	0,7103	0,6675	0,6159	0,5500
15	1,0807	0,9249	0,8448	0,7939	0,7582	0,7314	0,6937	0,6496	0,5961	0,5269
16	1,0719	0,9144	0,8331	0,7814	0,7450	0,7177	0,6791	0,6339	0,5786	0,5064
17	1,0641	0,9051	0,8229	0,7705	0,7335	0,7057	0,6663	0,6199	0,5630	0,4879
18	1,0572	0,8970	0,8138	0,7607	0,7232	0,6950	0,6549	0,6075	0,5491	0,4712
19	1,0511	0,8897	0,8057	0,7521	0,7140	0,6854	0,6447	0,5964	0,5366	0,4560
20	1,0457	0,8831	0,7985	0,7443	0,7058	0,6768	0,6355	0,5864	0,5253	0,4421
21	1,0408	0,8772	0,7920	0,7372	0,6984	0,6690	0,6272	0,5773	0,5150	0,4294
22	1,0363	0,8719	0,7860	0,7309	0,6916	0,6620	0,6196	0,5691	0,5056	0,4176
23	1,0322	0,8670	0,7806	0,7251	0,6855	0,6555	0,6127	0,5615	0,4969	0,4068
24	1,0285	0,8626	0,7757	0,7197	0,6799	0,6496	0,6064	0,5545	0,4890	0,3967
25	1,0251	0,8585	0,7712	0,7148	0,6747	0,6442	0,6006	0,5481	0,4816	0,3872
26	1,0220	0,8548	0,7670	0,7103	0,6699	0,6392	0,5952	0,5422	0,4748	0,3784
27	1,0191	0,8513	0,7631	0,7062	0,6655	0,6346	0,5902	0,5367	0,4685	0,3701
28	1,0164	0,8481	0,7595	0,7023	0,6614	0,6303	0,5856	0,5316	0,4626	0,3624
29	1,0139	0,8451	0,7562	0,6987	0,6576	0,6263	0,5813	0,5269	0,4570	0,3550
30	1,0116	0,8423	0,7531	0,6954	0,6540	0,6226	0,5773	0,5224	0,4519	0,3481
60	0,9784	0,8025	0,7086	0,6472	0,6028	0,5687	0,5189	0,4574	0,3746	0,2352
∞	0,9462	0,7636	0,6651	0,5999	0,5522	0,5152	0,4604	0,3908	0,2913	0

Таблица 9

1%-ные точки *F*-распределения (точки, в которых ф. р. = 0,99)
(Заимствованы из книги сэра Рональда Фишера и Ф. Йэйтса: *Statistical Tables for Biological, Medical and Agricultural Research.* Oliver and Boyd Ltd., Edinburgh.)

ν_2 \ ν_1	1	2	3	4	5	6	8	12	24	∞
1	4052	4999	5403	5625	5764	5859	5981	6106	6234	6366
2	98,49	99,00	99,17	99,25	99,30	99,33	99,36	99,42	99,46	99,50
3	34,12	30,81	29,46	28,71	28,24	27,91	27,49	27,05	26,60	26,12
4	21,20	18,00	16,69	15,98	15,52	15,21	14,80	14,37	13,93	13,46
5	16,26	13,27	12,06	11,39	10,97	10,67	10,27	9,89	9,47	9,02
6	13,74	10,92	9,78	9,15	8,75	8,47	8,10	7,72	7,31	6,88
7	12,25	9,55	8,45	7,85	7,46	7,19	6,84	6,47	6,07	5,65
8	11,26	8,65	7,59	7,01	6,63	6,37	6,03	5,67	5,28	4,86
9	10,56	8,02	6,99	6,42	6,06	5,80	5,47	5,11	4,73	4,31
10	10,04	7,56	6,55	5,99	5,64	5,39	5,06	4,71	4,33	3,91
11	9,65	7,20	6,22	5,67	5,32	5,07	4,74	4,40	4,02	3,60
12	9,33	6,93	5,95	5,41	5,06	4,82	4,50	4,16	3,78	3,36
13	9,07	6,70	5,74	5,20	4,86	4,62	4,30	3,96	3,59	3,16
14	8,86	6,51	5,56	5,03	4,69	4,46	4,14	3,80	3,43	3,00
15	8,68	6,36	5,42	4,89	4,56	4,32	4,00	3,67	3,29	2,87
16	8,53	6,23	5,29	4,77	4,44	4,20	3,89	3,55	3,18	2,75
17	8,40	6,11	5,18	4,67	4,34	4,10	3,79	3,45	3,08	2,65
18	8,28	6,01	5,09	4,58	4,25	4,01	3,71	3,37	3,00	2,57
19	8,18	5,93	5,01	4,50	4,17	3,94	3,63	3,30	2,92	2,49
20	8,10	5,85	4,94	4,43	4,10	3,87	3,56	3,23	2,86	2,42
21	8,02	5,78	4,87	4,37	4,04	3,81	3,51	3,17	2,80	2,36
22	7,94	5,72	4,82	4,31	3,99	3,76	3,45	3,12	2,75	2,31
23	7,88	5,66	4,76	4,26	3,94	3,71	3,41	3,07	2,70	2,26
24	7,82	5,61	4,72	4,22	3,90	3,67	3,36	3,03	2,66	2,21
25	7,77	5,57	4,68	4,18	3,86	3,63	3,32	2,99	2,62	2,17
26	7,72	5,53	4,64	4,14	3,82	3,59	3,29	2,96	2,58	2,13
27	7,68	5,49	4,60	4,11	3,78	3,56	3,26	2,93	2,55	2,10
28	7,64	5,45	4,57	4,07	3,75	3,53	3,23	2,90	2,52	2,06
29	7,60	5,42	4,54	4,04	3,73	3,50	3,20	2,87	2,49	2,03
30	7,56	5,39	4,51	4,02	3,70	3,47	3,17	2,84	2,47	2,01
40	7,31	5,18	4,31	3,83	3,51	3,29	2,99	2,66	2,29	1,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,82	2,50	2,12	1,60
120	6,85	4,79	3,95	3,48	3,17	2,96	2,66	2,34	1,95	1,38
∞	6,64	4,60	3,78	3,32	3,02	2,80	2,51	2,18	1,79	1,00

Примечание. Нижняя 1%-ная точка представляет собой обратную величину того табличного значения, которое получится, если ν_1 и ν_2 поменять местами, т. е. в качестве ν_1 всегда следует выбирать количество степеней свободы, соответствующее наибольшей выборочной дисперсии.

Таблица 10

Симметрические функции. Формулы представлений расширенных симметрических функций в терминах сумм степеней и обратные формулы
(Заимствованы из таблиц Дэвид и Кендалла (1949).)

Вес 1

$(1) = [1]$

Вес 2

	[2]	[1²]
(2)	1	−1
(1)²	1	1

Вес 3

	[3]	[21]	[1³]
(3)	1	−1	2
(2) (1)	1	1	−3
(1)³	1	3	1

Вес 4

	[4]	[31]	[2²]	[21²]	[1⁴]
(4)	1	−1	−1	2	−6
(3) (1)	1	1	.	−2	8
(2)²	1	.	1	−1	3
(2)(1)²	1	2	1	1	−6
(1)⁴	1	4	3	6	1

Вес 5

	[5]	[41]	[32]	[31²]	[2²1]	[21³]	[1⁵]
(5)	1	−1	−1	2	2	−6	24
(4) (1)	1	1	.	−2	−1	6	−30
(3) (2)	1	.	1	−1	−2	5	−20
(3) (1)²	1	2	1	1	.	−3	20
(2)² (1)	1	1	2	.	1	−3	15
(2) (1)³	1	3	4	3	3	1	−10
(1)⁵	1	5	10	10	15	10	1

Вес 6

	[6]	[51]	[42]	[41²]	[3²]	[321]	[31³]	[2³]	[2²1²]	[21⁴]	[1⁶]
(6)	1	−1	−1	2	−1	2	−6	2	−6	24	−120
(5) (1)	1	1	.	−2	.	−1	6	.	4	−24	144
(4) (2)	1	.	1	−1	.	−1	3	−3	5	−18	90
(4) (1)²	1	2	1	1	.	.	−3	.	−1	12	−90
(3)²	1	.	.	.	1	−1	2	.	2	−8	40
(3) (2) (1)	1	1	1	.	1	1	−3	.	−4	20	−120
(3) (1)³	1	3	3	3	1	3	1	.	.	−4	40
(2)³	1	.	3	.	.	.	.	1	−1	3	−15
(2)² (1)²	1	2	3	1	2	4	.	1	1	−6	45
(2) (1)⁴	1	4	7	6	4	16	4	3	6	1	−15
(1)⁶	1	6	15	15	10	60	20	15	45	15	1

Примечание. — Для представления функций [ ] в терминах ( ) используются только числа, стоящие в соответствующем столбце таблицы над главной диагональю (включительно). Например, $[41^2] = 2(6) - 2(5)(1) - (4)(2) + (4)(1)^2$. Аналогично для представления ( ) в терминах функций [ ] следует использовать только числа, стоящие в соответствующей строке слева от главной диагонали (включительно). Например, $(4)(1)^2 = [6] + 2[51] + [42] + [41^2]$.

ЦИТИРОВАННАЯ ЛИТЕРАТУРА*)

Адичи (Adichie J. N.) (1967a), Asymptotic efficiency of a class of non-parametric tests for regression parameters, Ann. Math. Statist., 38, 884.

Адичи (Adichie J. N.) (1967b), Estimates of regression parameters based on rank tests, Ann. Math. Statist. 38, 894.

Андерсон О. (Anderson O.) (1914), Nochmals über «The elimination of spurious correlation due to position in time and space», Biometrika **10**, 269.

Андерсон Р. (Anderson R. L.) (1942), Distribution of the serial correlation coefficient, Ann. Math. Statist. **13**, 1.

Андерсон Р. и Андерсон Т. (Anderson R. L. and Anderson T. W) (1956), Distribution of the circular serial correlation coefficient for residuals from a fitted Fourier series, Ann. Math. Statist. **21**, 59.

Андерсон Р. и Банкрофт (Anderson R. L. and Bancroft T. A.) (1952), Statistical Theory in Research, McGraw-Hill, New York.

Андерсон Т. (Anderson T. W.) (1946), The non-central Wishart distribution and certain problems of multivariate statistics, Ann. Math. Statist. **17**, 409.

Андерсон Т. (Anderson T. W.) (1958), Introduction to multivariate statistical analysis, Wiley, New York. (Русский перевод: Т. Андерсон, Введение в многомерный статистический анализ, М., Физматгиз, 1963).

Андерсон Т. (Anderson T. W.) (1962), The choice of the degree of a polynomial regression as a multiple decision problem, Ann. Math. Statist. **33**, 255.

Андерсон Т. (Anderson T. W.) (1963a), Asymptotic theory for principal component analysis, Ann. Math. Statist. **34**, 122.

Андерсон Т. (Anderson T. W.) (1963b), A test for equality of means when covariance matrices are unequal, Ann. Math. Statist. **34**, 671.

Андерсон Т. и Дас Гупта (Anderson T. W. and Das Gupta S.) (1964a), Monotonicity of the power functions of some tests of independence between two sets of variates, Ann. Math. Statist. **35**, 206.

Андерсон Т. и Дас Гупта (Anderson T. W. and Das Gupta S.) (1964b), A monotonicity property of the power functions of some tests of the equality of two covariance matrices, Ann. Math. Statist. **35**, 1059.

Армитедж (Armitage P.) (1947), A comparison of stratified with unrestricted random sampling from a finite population, Biometrika **34**, 273.

Арнольд (Arnold H. J.) (1964), Permutation support for multivariate techniques, Biometrika **51**, 65.

Атикулла (Atiqullah M.) (1962), The estimation of residual variance in quadratically balanced least-squares problems and the robustness of the F-test, Biometrika **49**, 83.

*) Отмеченные звездочкой статьи Вальда, Неймана, Пирсона, Стьюдента, Уилкса и Фишера можно найти также в приведенных в конце списка собраниях статей.

Атикулла (Atiqullah M.) (1964), The robustness of the covariance analysis of a one-way classification, Biometrika **51**, 365.

Афифи и Елашов Р. (Afifi A. A. and Elashoff R. M.) (1966), Missing observations in multivariate statistics, I, Review of literature, J. Amer. Statist. Ass. **61**, 595.

Балмер (Bulmer M. G.) (1957), Approximate confidence limits for components of variance, Biometrika **44**, 159.

Банерджи (Banerjee K. S.) (1964), A note on idempotent matrices, Ann. Math. Statist. **35**, 880.

Барнард Г. (Barnard G. A.) (1959), Control charts and stochastic processes, J. R. Statist. Soc. **B**, **21**, 239.

Барнард М. (Barnard M. M.) (1935), The secular variations of skull characters in four series of Egyptian skulls, Ann. Eugen. **6**, 352.

Барнетт и Льюис (Barnett V. D. and Lewis T.) (1963), A study of the relations between G. C. E. and degree results, J. R. Statist. Soc. **A**, **126**, 187.

Барра (Barra J. R.) (1965), Carrés latins et eulériens, Rev. Int. Statist. Inst. **33**, 16.

Бартлетт (Bartlett M. S.) (1933), On the theory of statistical regression, Proc. Roy. Soc. Edin. **53**, 260.

Бартлетт (Bartlett M. S.) (1936), The square root transformation in analysis of variance, Suppl. J. R. Statist. Soc. **3**, 68.

Бартлетт (Bartlett M. S.) (1946), On the theoretical specification and sampling properties of autocorrelated time-series, Suppl. J. R. Statist. Soc. **8**, 27, 85 (список опечаток, 1948 : **10**).

Бартлетт (Bartlett M. S.) (1947a), The use of transformations, Biometrics **3**, 39.

Бартлетт (Bartlett M. S.) (1947b), The general canonical correlation distribution, Ann. Math. Statist. **18**, 1.

Бартлетт (Bartlett M. S.) (1947c), Multivariate analysis, Suppl. J. R. Statist. Soc. **9**, 176.

Бартлетт (Bartlett M. S.) (1950), Periodogram analysis and continuous spectra, Biometrika **37**, 1.

Бартлетт (Bartlett M. S.) (1951a), The effect of standardization on a χ^2 approximation in factor analysis, Biometrika **38**, 337.

Бартлетт (Bartlett M. S.) (1951b), The goodness of fit of a single hypothetical discriminant function in the case of several groups, Ann. Eugen. **16**, 199.

Бартлетт (Bartlett M. S.) (1954), A note on the multiplying factors for various χ^2 approximations, J. R. Statist. Soc. **B**, **16**, 296.

Бартлетт (Bartlett M. S.) (1955), An Introduction to Stochastic Processes, with Special Reference to Methods and Applications, Cambridge Univ. Press. (Русский перевод: М. Бартлетт, Введение в теорию случайных процессов, М., ИЛ, 1958).

Бартлетт и Кендалл Д. (Bartlett M. S. and Kendall D. G.) (1946), The statistical analysis of variance-heterogeneity and the logarithmic transformation, Suppl. J. R. Statist. Soc. **8**, 128.

Бартлетт и Плиз (Bartlett M. S. and Please N. W.) (1963), Discrimination in the case of zero mean differences, Biometrika **50**, 17.

Бартлетт и Раджалакшман (Bartlett M. S. and Rajalakshman D. V.) (1953), Goodness of fit tests for simultaneous autoregressive series, J. R. Statist. Soc. **B**, **15**, 107.

Бартоломью (Bartholomew D. J.) (1961), Ordered tests in the analysis of variance, Biometrika **48**, 325.

Басу (Basu D.) (1958), On sampling with and without replacement, Sankhyā **20**, 287.

Беверидж (Beveridge W. H.) (1921), Weather and harvest cycles, Econ J. **31**, 429.
Бенард и Ван Элтерн (Benard A. and Van Elteren P.) (1953), A generalization of the method of m rankings, Proc. Kon. Ned. Akad. Wetensch. **A**, **56** (Indag. Math. **15**, 358).
Беннетт (Bennett B. M.) (1951), Note on the solution of the generalized Behrens-Fisher problem, Ann. Inst. Statist. Math., Tokyo, **2**, 87.
Бёрмэн (Burman J. P.) (1965), Moving seasonal adjustment of economic time series, J. R. Statist. Soc. **A**, **128**, 534; см. также тот же журнал (1966), **129**, 274.
Биггерс (Biggers J. D.) (1959), The estimation of missing and mixed-up observations in several experimental designs, Biometrika **46**, 91.
Бикел (Bickel P. J.) (1965), On some robust estimates of location, Ann. Math. Statist. **36**, 847.
Бил, Кендалл М. и Манн (Beale E. M. L., Kendall M. G. and Mann D. W.) (1967), The discarding of variables in multivariate analysis, Biometrika **54**, 357.
Биркгоф (Birkhoff G. D.) (1931), Proof of the ergodic theorem, Proc. Nat. Acad. Sci. USA, **17**, 656.
Блом (Blom G.) (1954), Transformations of the binomial, negative binomial, Poisson and χ^2 distributions, Biometrika **41**, 302.
Блэкис (Blackith R. E.) (1960), A synthesis of multivariate techniques to distinguish patterns of growth in grasshoppers, Biometrics **16**, 28.
Блэкмэн и Тьюки (Blackman R. B. and Tukey J. W.) (1958), The Measurement of Power Spectra from the Point of View of Communications Engineering, Dover Co., New York.
Бозе (Bose R. C.) (1936), On the exact distribution and moment-coefficients of the D^2 statistic, Sankhyā **2**, 143.
Бозе (Bose R. C.) (1939), On the construction of balanced incomplete block designs, Ann. Eugen. **9**, 353.
Бозе (Bose R. C.) (1942), A note on the resolvability of balanced incomplete block designs, Sankhyā **6**, 105.
Бозе (Bose R. C.) (1956), Paired comparison designs for testing concordance between judges, Biometrika **43**, 113.
Бозе и Дрэйпер (Bose R. C. and Draper N. R.) (1959), Second order rotatable designs in three dimensions, Ann. Math. Statist. **30**, 1097.
Бозе и Картер (Bose R. C. and Carter R. L.) (1959), Complete representation in the construction of rotatable designs, Ann. Math. Statist. **30**, 771.
Бозе, Клатуэрти и Шриканде (Bose R. C., Clatwothy W. H. and Shrikande S. S.) (1954), Tables of partially balanced designs with two associate classes, North Carolina Agric. Exper. Sta. Tech. Bull. no. 107.
Бозе и Нэйр (Bose R. C. and Nair K. R.) (1939), Partially balanced incomplete block designs, Sankhyā **4**, 337.
Бозе и Рой С. (Bose R. C. and Roy S. N.) (1938), The distribution of the studentized D^2 statistic, Sankhyā **4**, 19.
Бозе и Шриканде (Bose R. C. and Shrikande S. S.) (1959), On the falsity of Euler's conjecture about the non-existence of two orthogonal latin squares of order $4t+2$, Proc. Nat. Acad. Sci. USA **45**, 734.
Бозе и Шриканде (Bose R. C. and Shrikande S. S.) (1960), On the construction of sets of mutually orthogonal latin squares and the falsity of a conjecture of Euler, Trans. Amer. Math. Soc. **95**, 191.
Бозе, Шриканде и Паркер (Bose R. C., Shrikande S. S. and Parker R.) (1960), Further results on the construction of mutually orthogonal latin squares and the falsity of Euler's conjecture, Canadian J. Math. **12**, 189.
Бозивич, Баикрофт и Хартли (Bozivich H., Bancroft T. A. and Hartley H. O.) (1956), Power of analysis of variance test procedures for

certain incompletely specified models, I, Ann. Math. Statist, **27**, 1017.
Бокс (Box G. E. P.) (1949), A general distribution theory for a class of likelihood criteria, Biometrika **36**, 317.
Бокс (Box G. E. P.) (1954), Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification, II. Effects of inequality of variance and of correlation between errors in the two-way classification, Ann. Math. Statist. **25**, 290, 484.
Бокс (Box G. E. P.) (1957), Evolutionary operation: a method for increasing industrial productivity, Appl. Statist. **6**, 81.
Бокс и Андерсен (Box G. E. P. and Andersen S. L.) (1955), Permutation theory in the derivation of robust criteria and the study of departures from assumption, J. R. Statist. Soc. **B**, **17**, 1.
Бокс и Бенкен (Box G. E. P. and Behnken D. W.) (1960), Simplex-sum designs: a class of second order rotatable designs derivable from those of first order, Ann. Math. Statist. **31**, 838.
Бокс и Ватсон (Box G. E. P. and Watson G. S.) (1962), Robustness to non-normality of regression tests, Biometrika **49**, 93.
Бокс и Дженкинс (Box G. E. P. and Jenkins G. M.) (1962), Some statistical aspects of adaptive optimization and control, J. R. Statist. Soc. **B**, **24**, 297.
Бокс и Кокс Д. (Box G. E. P. and Cox D. R.) (1964), An analysis of transformations, J. R. Statist. Soc. **B**, **26**, 211.
Бокс и Тидвелл (Box G. E. P. and Tidwell P. W.) (1962), Transformation of the independent variables, Technometrics **4**, 531.
Бокс и Уилсон К. (Box G. E. P. and Wilson K. B.) (1951), On the experimental attainment of optimum conditions, J. R. Statist. Soc. **B**, **13**, 1.
Бокс и Хантер (Box (G. E. P. and Hunter J. S.) (1957), Multifactor experimental designs for exploring response surfaces, Ann. Math. Statist. **28**, 195.
Борер (Bohrer R. (1967), On sharpening Scheffé bounds, J. R. Statist. Soc. **B**, **29**, 110.
Браду (Bradu D.) (1965), Main effect analysis of the general non-orthogonal layout with any number of factors, Ann. Math. Statist. **36**, 88.
Браун Г. и Муд (Brown G. W. and Mood A. M.) (1951), On median tests for linear hypotheses, Proc. 2nd Berkeley Symp. Math. Statist. and Prob., 159.
Браун Р. (Brown R. G.) (1959), Statistical Forecasting for Inventory Control, McGraw-Hill, New York.
Бриллинджер (Brillinger D. R.) (1965), An introduction to polyspectra, Ann. Math. Statist. **36**, 1351.
Брюэр (Brewer K. R. W.) (1967), A note on Fellegi's method of sampling without replacement with probability proportional to size, J. Amer. Statist. Ass. **62**, 79.
Буднэ (Budne T. A.) (1959), The application of random balance designs, Technometrics **1**, 139.
Бхапкар (Bhapkar V. P.) (1963), The asymptotic power and efficiency of Mood's test for two-way classification, J. Indian Statist. Ass. **1**, 24.
Бхаттачария (Bhattachharyya G. K.) (1967), Asymptotic efficiency of multivariate normal score test, Ann. Math. Statist. **38**, 1753.
Бхучонгкул и Пури (Bhuchongkul S. and Puri M. L.) (1965), On the estimation of contrasts in linear models, Ann. Math. Statist. **36**, 198.
Вайда (Vajda S.) (1967a), Patterns and Configurations in Finite Spaces, Griffin, London.
Вайда (Vajda S.) (1967b), The Mathematics of Experimental Design: Incomplete Block Designs and Latin Squares, Criffin, London.

Вальд (Wald A.) (1940)*, A note on the analysis of variance with unegual class frequencies, Ann. Math. Statist. **11**, 96.
Вальд (Wald A.) (1941)*, On the analysis of variance in case of multiple classifications with unequal class frequencies, Ann. Math. Statist. **12**, 346.
Вальд (Wald A.) (1944)*, On a statistical problem arising in the classification of an individual into one of two groups, Ann. Math. Statist. **15**, 145.
Вальд (Wald A.) (1947)*, A note on regression analysis, Ann. Math. Statist. **18**, 586.
Вальд и Волфовиц (Wald A. and Wolfowitz J.) (1943)*, An exact test for randomness in the non-parametric case based on serial correlation, Ann. Math. Statist. **14**, 378.
Ван Элтерн и Нетер (Van Elteren P. and Noether G. E.) (1959), The asymptotic efficiency of the χ^2-test for a balanced incomplete block design, Biometrika **46**, 475.
Ватсон (Watson G. S.) (1955), Serial correlation in regression analysis, I, Biometrika **42**, 327.
Ватсон (Watson G. S.) (1956), On the joint distribution of the circular serial correlation coefficients, Biometrika **43**, 161.
Ватсон и Хеннан (Watson G. S. and Hannan E. J.) (1956), Serial correlation in regression analysis, II, Biometrika **43**, 436.
Вейсман (Wijsman R. A.) (1957), Random orthogonal transformations and their use in some classical distribution problems in multivariate analysis, Ann. Math. Statist. **28**, 415.
Винер (Wiener N.) (1930), Generalized harmonic analysis, Acta Math. 55, 117.
Волд (Wold H. O. A.) (1938), A Study in the Analysis of Stationary Time Series, 2nd edn, 1954, Almqvist and Wiksell, Stockholm.
Волд (Wold H. O. A.) (1949), A large-sample test for moving averages, J. R. Statist. Soc. **B**, **11**, 297.
Волд (Wold H. O. A.) (ed.) (1964), Econometric Model Building: Essays on the Causal Chain Approach, North Holland Publishing Co., Amsterdam.
Волд (Wold H. O. A.) (1965), Bibliography on Time Series and Stochastic Processes: an International Team Project, Oliver and Boyd, Edinburgh.
Волфовиц (Wolfowitz J.) (1944), Asymptotic distribution of runs up and down, Ann. Math. Statist. **15**, 163.
Вос (Vos J. W. E.) (1964), Sampling in space and time, Rev. Int. Statist. Inst. **32**, 226.
Габриэл (Gabriel K. R.) (1963), Analysis of variance of proportions with unequal frequencies, J. Amer. Statist. Ass. **58**, 1133.
Габриэл (Gabriel K. R.) (1964), A procedure for testing the homogeneity of all sets of means in analysis of variance, Biometrics **20**, 459.
Габриэл (Gabriel K. R.) (1967), Simultaneous test procedures with applications, Ann. Math. Statist. 38.
Гаек (Hajek J.) (1960), Limiting distributions in simple random sampling from a finite population, Pub. Math. Inst. Hung. Acad. Sci. **B**, **5**, 361.
Гаек (Hajek J.) (1964), Asymptotic theory of rejective sampling with varying probabilities from a finite population, Ann. Math. Statist. **35**, 1491.
Гардинер, Грэндейдж и Хейдер (Gardiner D. A., Grandage A. H. E. and Hader R. J.) (1959), Third order rotatable designs for exploring response surfaces, Ann. Math. Statist. **30**, 1082.
Гасснер (Gassner B. J.) (1965), Equal-difference BIB designs, Proc. Amer. Math. Soc. **16**, 378.
Гаучи (Gautschi W.) (1959), Some remarks on Herbach's paper, «Optimum nature of the *F*-test for Model II in the balanced case», Ann. Math. Statist. **30**, 960.
Гейлор и Свини (Gaylor D. W. and Sweeny H. C.) (1965), Design for optimal prediction in simple linear regression, J. Amer. Statist. Ass. **60**, 205.

Гейссер (Geisser S.) (1964), Estimation in the uniform covariance case, J. R. Statist. Soc. **B**, **26**, 477.
Гербах (Herbach L. H.) (1959), Properties of Model II-type analysis of variance tests, A: optimum nature of the *F*-test for Model II in the balanced case, Ann. Math. Statist. **30**, 939.
Гесс, Сэти и Балакришнан (Hess I., Sethi V. K. and Balakrishnan T. R.) (1966), Stratification: a practical investigation, J. Amer. Statist. Ass. **61**, 74.
Гест (Guest P. G.) (1958), The spacing of observations in polynomial regression, Ann. Math. Statist. **29**, 294.
Гири (Geary R. C.) (1944), Extension of a theorem by Harald Cramér on the frequency distribution of the quotient of two variables, J. R. Statist. Soc. **107**, 56.
Гиршик (Girshick M. A.) (1939), On the sampling theory of the roots of determinantal equations, Ann. Math. Statist. **10**, 203.
Глезер (Gleser L. J.) (1966), A note on the sphericity test, Ann. Math. Statist. **37**, 464.
Глейзберг (Gleissberg W.) (1945), Eine Aufgabe der Kombinatorik und Wahrscheinlichkeitsrechnung, Univ. Istanbul Rev. Fac. Sci. **A**, **10**, 25.
Годэмб (Godambe V. P.) (1955), A unified theory of sampling from finite populations, J. R. Statist. Soc. **B**, **17**, 269.
Годэмб (Godambe V. P.) (1965), A review of the contributions towards a unifield theory of sampling from finite populations, Rev. Int. Statist. Inst. **33**, 242.
Гранди (Grundy P. M.) (1954), A method of sampling with probability exactly proportional to size, J. R. Statist. Soc. **B**, **16**, 236.
Грейбилл и Марсалья (Graybill F. A. and Marsaglia G.) (1957), Idempotent matrices and quadratic forms in the general linear hypothesis, Ann. Math. Statist. 28, 678.
Грейбилл и Сэшадри (Graybill F. A. and Seshadri V.) (1960), On the unbiasedness of Yates' method of estimation using interblock information, Ann. Math. Statist. **31**, 786.
Грейбилл и Уикс (Graybill F. A. and Weeks D. L.) (1959), Combining inter-block and intrablock information in balanced incomplete blocks, Ann. Math. Statist. **30**, 799.
Грейбилл и Халтквист (Graybill F. A. and Hultquist R. A.) (1961), Theorems concerning Eisenhart's Model II, Ann. Math. Statist. **32**, 261.
Гренандер и Розенблатт (Grenander U. and Rosenblatt M.) (1957), Statistical Analysis of Stationary Time Series, Wiley, New York.
Гринберг (Greenberg V. L.) (1966), Robust estimation in incomplete blocks designs, Ann. Math. Statist. **37**, 1331.
Грэнджер (Granger C. W. J.) (1963), The effect of varying month length on the analysis of economic time series, L'industrie **1**, 41.
Грэнджер и Хатанака (Granger C. W. J. and Hatanaka M.) (1964), Spectral Analysis of Economic Time Series, Princeton Univ. Press, Princeton.
Гуд (Good I. J.) (1950), The inversion of circulant matrices, Biometrika **37**, 185.
Гудмэн и Хартли (Goodman L. A. and Hartley H. O.) (1958), The precision of unbiased ratio-type estimators, J. Amer. Statist. Ass. **53**, 491.
Гуерин (Guérin R.) (1965), Vue d'ensemble sur les plans en blocs incomplets équilibrés et partiellement équilibrés, Rev. Int. Statist. Inst. **33**, 24.
Гхош Б. (Ghosh B. K.) (1964), Simultaneous tests by sequential methods in hierarchical classifications, Biometrika **51**, 439.
Гхош Б. (Ghosh B. K.) (1967), Sequential analysis of variance under random and mixed models, J. Amer. Statist. Ass. **62**, 1401.

Гхош М. и Шарма (Ghosh M. N. and Sharma D.) (1963), Power of Tukey's test for non-additivity, J. R. Statist. Soc. **B, 25,** 213.
Гхош С. (Ghosh S. P.) (1963a), Post cluster sampling, Ann. Math. Statist. **34,** 587.
Гхош С. (Ghosh S. P.) (1963b), Optimum stratification with two characters, Ann. Math. Statist. **34,** 866.
Далэниус (Dalenius T.) (1950), The problem of optimum stratification, Skand. Aktuartidskr. **33,** 203.
Далэниус (Dalenius T.) (1952), The problem of optimum stratification in a special type of design, Skand. Aktuartidskr. **35,** 61.
Далэниус (Dalenius T.) (1953), The economics of one-stage stratified sampling, Sankhyā **12,** 351.
Далэниус и Ходжес (Dalenius T. and Hodges J. L., Jr.) (1957), The choice of stratification points, Skand. Aktuartidskr. **40,** 198.
Далэниус и Ходжес (Dalenius T. and Hodges J. L., Jr.) (1959), Minimum variance stratification, J. Amer. Statist. Ass. **54,** 88.
Даниэлл (Daniell P. J.) (1946), Discussion on «Symposium on autocorrelation in time series», Suppl. J. R. Statist. Soc. **8,** 88.
Даниэлс (Daniels H. E.) (1956), The approximate distribution of serial correlation coefficients, Biometrika **43,** 169.
Даниэлс (Daniels H. E.) (1962), The estimation of spectral densities, J. R. Statist. Soc. **B, 24,** 185.
Данн (Dunn O. J.) (1961), Multiple comparisons among means, J. Amer. Statist. Ass. **56,** 52.
Дарлинг (Darling D. A.) (1952), On a test for homogeneity and extreme values, Ann. Math. Statist. **23,** 450 (исправление: **24,** 135).
Дас Гупта, Андерсон Т. и Мудхолкар (Das Gupta S., Anderson T. W. and Mudholkar G. S.) (1964), Monotonicity of the power functions of some tests of the multivariate linear hypothesis, Ann. Math. Statist. **35,** 200.
Дейли (Daly J. F.) (1940), On the unbiased character of likelihood-ratio-tests for independence in normal systems, Ann. Math. Statist. **11,** 1.
Дербин (Durbin J.) (1951), Incomplete blocks in ranking experiments, Brit. J. Psychol. (Statist. Sec.) **4,** 85.
Дербин (Durbin J.) (1953), Some results in sampling theory when the units are selected with unequal probabilities, J. R. Statist. Soc. **B, 15,** 262.
Дербин (Durbin J.) (1954), Non-response and call-backs in surveys, Bull. Int. Statist. Inst. **34**(2), 72.
Дербин (Durbin J.) (1958), Sampling theory for estimates based on fewer individuals than the number selected, Bull. Int. Statist, Inst. **36**(3), 113.
Дербин (Durbin J.) (1959a), A note on the application of Quenouille's method of bias reduction to the estimation of ratios, Biometrika **46,** 477.
Дербин (Durbin J.) (1959b), Efficient estimation of parameters in moving-average models, Biometrika **46,** 306.
Дербин (Durbin J.) (1960a), Estimation of parameters in time series regression models, J. R. Statist. Soc. **B, 22,** 139.
Дербин (Durbin J.) (1960b), The fitting of time-series models, Rev. Int. Statist. Inst. **28,** 233.
Дербин (Durbin J.) (1961), Efficient fitting of linear models for continuous stationary time series from discrete data, Bull. Int. Statist. Inst. 38(4), 273.
Дербин (Durbin J.) (1965), A method of sample selection with unequal probabilities without replacement (Abstract.), Ann. Math. Statist. **36,** 1327.
Дербин и Ватсон (Durbin J. and Watson G. S.) (1950—1), Testing for serial correlation in least squares regression, I, II, Biometrika **37,** 409; **38,** 159.

Джеймс (James A. T.) (1964), Distributions of matrix variates and latent roots derived from normal samples, Ann. Math. Statist. **35,** 475.
Джеймс (James A. T.) (1966), Inference on latent roots by calculation of hypergeometric functions of matrix argument, Proc. Symp. Multiv. Analysis, Academic Press, New York.
Дженкинс (Jenkins G. M.) (1954a), An angular transformation for the serial correlation coefficient, Biometrika **41,** 261.
Дженкинс (Jenkins G. M.) (1954b, 1956), Tests of hypothesis in the linear autoregressive model, I: Null hypothesis distributions in the Yule scheme, II: Null distributions for higher order schemes; non-null distributions, Biometrika **41,** 405; **43,** 186.
Дженкинс (Jenkins G. M.) (1961), General considerations in the analysis of spectra, Comments on the discussions of Messrs Tukey and Goodman, Technometrics **3,** 133, 229.
Джон (John S.) (1961), Errors in discrimination, Ann. Math. Statist. **32,** 1125.
Джонсон (Johnson N. L.) (1948), Tests of significance in the variate difference method, Biometrika **35,** 203.
Джонсон (Johnson N. L.) (1953), Some notes on the application of sequential methods in the analysis of variance, Ann. Math. Statist. **24,** 614.
Джонсон (Johnson N. L.) (1954), Sequential procedures in certain component of variance problems, Ann. Math. Statist. **25,** 357.
Джонсон (Johnson N. L.) (1957), Optimal sampling for quota fulfilment, Biometrika **44,** 518.
Диксон (Dixon W. J.) (1944), Further contributions to the problem of serial correlation, Ann. Math. Statist. **15,** 119.
Доксум (Doksum K.) (1967), Robust procedures for some linear models with one observation per cell, Ann. Math. Statist. **38,** 878.
Долби (Dolby J. L.) (1963), A quick method for choosing a transformation, Technometrics **5,** 317.
Дрэйпер (Draper N. R.) (1960a), Second order rotatable designs in four or more dimensions, Ann. Math. Statist. **31,** 23.
Дрэйпер (Draper N. R.) (1960b), Third order rotatable designs in three dimensions, Ann. Math. Statist. **31,** 865.
Дрэйпер (Draper N. R.) (1960c), A third order rotatable design in four dimensions, Ann. Math. Statist. **31,** 875.
Дрэйпер (Draper N. R.) (1961), Third order rotatable designs in three dimensions: some specific designs, Ann. Math. Statist. **32,** 910.
Дункан (Duncan D. B.) (1952), On the properties of the multiple comparisons test, Virginia, J. Sci. N. S., **3,** 49.
Дункан (Duncan D. B.) (1955), Multiple range and multiple F-tests, Biometrics **11,** 1.
Дункан (Duncan D. B.) (1957), Multiple range tests for correlated and heteroscedastic means, Biometrics **13,** 164.
Дэвид (David H. A.) (1963), The Method of paired Comparisons, Griffin, London.
Дэвид и Аренс (David H. A. and Arens B. E.) (1959), Optimal spacing in regression analysis, Ann. Math. Statist. **30,** 1072.
Дэвис (Davis H. T.) (1941), The analysis of economic time series, Cowles Comm. Res. Econ. (Bloomington, Indiana), Monogr. no. 6.
Дэвьес, Бокс, Коннор, Кузнн, Химсворс и Силлитто (Davies O. L. (also editor), Box G. E. P., Connor L. R., Cousins W. R., Himsworth F. R. and Sillitto G. P.) (1954), The Design and Analysis of Industrial Experiments, Oliver and Boyd, Edinburgh.
Дэмпстер (Dempster A. P.) (1958), A high dimensional two sample significance test, Ann. Math. Statist. **29,** 995.

Дэмпстер (Dempster A. P.) (1960—1), Random allocation design, I. On general classes of estimation methods, II. Approximate theory for simple random allocation, Ann. Math. Statist. **31**, 885; **32**, 387.
Дэмпстер (Dempster A. P.) (1966), Estimation in multivariate analysis, Proc. Symp. Multiv. Analysis, Academic Press, New York.
Дэрроч и Силви (Darroch J. N. and Silvey S. D.) (1963), On testing more than one hypothesis, Ann. Math. Statist. **34**, 555.
Елашов Дж., Елашов Р. и Голдмэн (Elashoff J. D., Elashoff R. M. and Goldman G. E.) (1967), On the choice of variables in classification problems with dichotomous variables, Biometrika **54**, 668.
Имхоф (Imhof J. P.) (1960), A mixed model for the complete three-way layout with two random-effects factors, Ann. Math. Statist. **31**, 906.
Имхоф (Imhof J. P.) (1962), Testing the hypothesis of no fixed main-effects in Scheffé's mixed model, Ann. Math. Statist. **33**, 1085.
Ито и Шулль (Ito K. and Schull W. J.) (1964), On the robustness of the T_0^2 test in multivariate analysis of variance when variance-covariance matrices are not equal, Biometrika **51**, 71.
Иэйтс (Yates F.) (1933), The analysis of replicated experiments when the field results are incomplete, Empire J. Exper. Agric. **1**, 129.
Иэйтс (Yates F.) (1934). The analysis of multiple classifications with unequal numbers in the different classes, J. Amer. Statist. Ass. **29**, 51.
Иэйтс (Yates F.) (1936a), Incomplete randomized blocks, Ann. Eugen **7**, 121.
Иэйтс (Yates F.) (1936b), A new method of arranging variety trials involving a large number of varieties, J. Agric. Sci. **26**, 424.
Иэйтс (Yates F.) (1937), The design and analysis of factorial experiments, Imp. Bur. Soil Sci. (Harpenden) Tech. Comm. no. 35.
Иэйтс (Yates F.) (1939), The recovery of inter-block information in variety trials arranged in three-dimensional lattices. Ann. Eugen. **9**, 136.
Иэйтс (Yates F.) (1940a), The recovery of inter-block information in balanced incomplete block designs, Ann. Eugen. **10**, 317.
Иэйтс (Yates F.) (1940b), Lattice squares, J. Agric. Sci. **30**, 672.
Иэйтс (Yates F.) (1960), Sampling Methods for Censuses and Surveys, 3rd edn. Griffin, London. (Русский перевод: Ф. Иейтс, Выборочный метод в переписях и обследованиях, М., Статистика, 1965).
Иэйтс и Гранди (Yates F. and Grundy P. M.) (1953), Selection without replacement from within strata with probability proportional to size, J. R. Statist. Soc. **B**, **15**, 253.
Калински (Caliński T.) (1966), On the distribution of the F-type statistics in the analysis a group of experiments, J. R. Statist. Soc. **B**, **28**, 526.
Карлин и Стадден (Karlin S. and Studden W. J.) (1966), Optimal experimental designs, Ann. Math. Statist. **37**, 783.
Кейлс (Keuls M.) (1952), The use of «Studentized range» in connection with an analysis of variance, Euphytica **1**, 112.
Кемпсорн (Kempthorne O.) (1952), The Design and Analysis of Experiments, Wiley, New York.
Кендалл (Kendall M. G.) (1947), The estimation of parameters in linear autoregressive time series, Econometrica Suppl. **17**, 44.
Кендалл (Kendall M. G.) (1954), Note on bias in the estimation of autocorrelation, Biometrika **41**, 403.
Кендалл (Kendall M. G.) (1957), The moments of the Leipnik distribution, Biometrika **44**, 270.
Кендалл (Kendall M. G.) (1961a), A theorem in trend analysis, Biometrika **48**, 224.
Кендалл (Kendall M. G.) (1961b), A Course in Multivariate Analysis, 2nd edn, Griffin, London.
Кендалл (Kendall M. G.) (1966), Discrimination and classification, Proc. Symp. Multiv. Analysis, Academic Press, New York.

Кендалл и Смит Б. (Kendall M. G. and Smith B. Babington) (1939), The problem of m rankings, Ann. Math. Statist. **10**, 275.
Кенуй (Quenouille M. H.) (1947a), Notes on the calculation of autocorrelations of linear autoregressive schemes, Biometrika **34**, 365.
Кенуй (Quenouille M. H.) (1947b), A large-sample test for the goodness of fit of autoregressive schemes, J. R. Statist. Soc. **110**, 123.
Кенуй (Quenouille M. H.) (1948), Some results in the testing of serial correlation coefficients, Biometrika **35**, 261.
Кенуй (Quenouille M. H.) (1949a), A method of trend elimination, Biometrika **36**, 75.
Кенуй (Quenouille M. H.) (1949b), The joint distribution of serial correlation coefficients, Ann. Math. Statist. **20**, 561.
Кенуй (Quenouille M. H.) (1953a), The Design and Analysis of Experiment, Griffin, London.
Кенуй (Quenouille M. H.) (1953b), Modifications to the variate-difference method, Biometrika **40**, 383.
Кенуй (Quenouille M. H.) (1956), Notes on bias in estimation, Biometrika **43**, 353.
Кенуй (Quenouille M. H.) (1957), The Analysis of Multiple Time-Series, Griffin, London.
Кенуй (Quenouille M. H.) (1958), Discrete autoregresive schemes with varying time-intervals, Metrika **1**, 21.
Кёртисс (Curtiss J. H.) (1943), On transformations used in the analysis of variance, Ann. Math. Statist. **14**, 107.
Кинг (King B.) (1967), Step-wise clustering procedures, J. Amer. Statist. Ass. **62**, 86.
Кифер (Kiefer J.) (1958), On the nonrandomized optimality and randomized nonoptimality of symmetrical designs, Ann. Math. Statist. **29**, 675.
Кифер (Kiefer J.) (1959), Optimum experimental designs, J. R. Statist. Soc. **B**, **21**, 272.
Кифер и Волфовиц (Kiefer J. and Wolfowitz J.) (1954), Optimum designs in regression problems, Ann. Math. Statist. **30**, 271.
Киш и Гесс (Kish L. and Hess I.) (1959), On variances of ratios and their differences in multistage samples, J. Amer. Statist. Ass. **54**, 416.
Клатуэрти и Левипкий (Clatworthy W. H. and Lewyckyj R. J.) (1968), см. Такэючи (1962).
Кокрэйн и Оркутт (Cochrane D. and Orcutt G. H.) (1949), Application of least-squares regression to relationships containing auto-correlated error terms, J. Amer. Statist. Ass. **44**, 32.
Кокрэн (Cochran W. G.) (1941), The distribution of the largest of a set of estimated variances as a fraction of their total, Ann. Eugen. **11**, 47.
Кокрэн (Cochran W. G.) (1947), Some consequences when the assumptions for the analysis of variance are not satisfied, Biometrics **3**, 22.
Кокрэн (Cochran W. G.) (1961), Comparison of methods for determining stratum boundaries, Bull. Int. Statist. Inst. **38(2)**, 345.
Кокрэн (Cochran W. G.) (1963), Sampling Techniques, 2nd edn. Wiley, New York. (Русский перевод: В. Кокрен, Методы выборочного обследования, М., «Статистика», 1976.)
Кокрэн (Cochran W. G.) (1964), On the performance of the linear discriminant function, Technometrics **6**, 179.
Кокрэн (Cochran W. G.) (1965), The planning of observational studies of human populations, J. R. Statist. Soc. **A**, **128**, 234.
Кокрэн и Блисс (Cochran W. G. and Bliss C. I.) (1948), Discriminant functions with covariance, Ann. Math. Statist. **19**, 151.
Кокрэн и Гопкинс (Cochran W. G. and Hopkins C. E.) (1961), Some classification problems with multivariate qualitative data, Biometrics **17**, 10.

Кокрэн и Кокс Г. (Cochran W. G. and Cox G. M.) (1957), Experimental Designs, 2nd edn, Wiley, New York.
Кокс (Cox D. R.) (1952), Sequential tests for composite hypotheses, Proc. Camb. Phil. Soc. **48**, 290.
Кокс (Cox D. R.) (1958a), Planning of Experiments, John Wiley, New York.
Кокс (Cox D. R.) (1958b), The interpretation of the effects of non-additivity in the latin square, Biometrika **45**, 69.
Кокс (Cox D. R.) (1961), Prediction by exponentially weighted moving averages and related methods, J. R. Statist. Soc. **B**, **23**, 414.
Кокс (Cox D. R.) (1966), The null distribution of the first serial correlation coefficient, Biometrika **53**, 623.
Кокэн и Хан (Kokan A. R. and Khan S.) (1967), Optimum allocation in multivariate surveys: an analytical solution, J. R. Statist. Soc. **B**, **29**, 115.
Кольер и Бейкер (Collier R. O. Jr. and Baker F. B.) (1966), Some Monte Carlo results on the power of the F-test under permutation in the simple randomized block design, Biometrika **53**, 199.
Константин (Constantine A. G.) (1966), The distribution of Hotelling's generalised T_0^2, Ann. Math. Statist. **37**, 215.
Консул (Consul P. C.) (1966), On the exact distribution of the likelihood ratio criteria for testing linear hypotheses about regression coefficients, Ann. Math. Statist. **37**, 1319.
Консул (Consul P. C.) (1967a), On the exact distributions of likelihood ratio criteria for testing independence of sets of variates under the null hypothesis, Ann. Math. Statist. **38**, 1160.
Консул (Consul P. C.) (1967b), On the exact distributions of the criterion W for testing sphericity in a p-variate normal distribution, Ann. Math. Statist. **38**, 1170.
Корин (Korin B. P.) (1968), On the distribution of a statistic used for testing a covariance matrix, Biometrika **55**, 171.
Корнфилд и Тьюки (Cornfield J. and Tukey J. W.) (1956), Average values of mean squares in factorials, Ann. Math. Statist. **27**, 907.
Коуден (Cowden D. J.) (1962), Weights for fitting polynomial secular trends, Univ. North Carolina Sch. Business Admin. Tech. Paper no 4.
Крускал (Kruskal J. B.) (1965), Analysis of factorial experiments by estimating monotone transformations of the data, J. R. Statist. Soc. **B**, **27**, 251.
Крэддок (Craddock J. M.) (1965), A meteorological application of principal component analysis, The Statistician **15**, 143.
Куад (Quade D.) (1967), Rank analysis of covariance, J. Amer. Statist. Ass. **62**, 1187.
Куп (Koop J. C.) (1963), On the axioms of sample formation and their bearing on the construction of linear estimators in sampling theory for finite universes, I, II, III, Metrika **7**, 81 and 165.
Купмэнс (Koopmans T. C.) (1942), Serial correlation and quadratic forms in normal variables, Ann. Math. Statist. **13**, 14.
Кширсагар (Kshirsagar A. M.) (1959), Bartlett decomposition and Wishart distribution, Ann. Math. Statist. **30**, 239.
Кширсагар (Kshirsagar A. M.) (1961), The non-central multivariate beta distribution, Ann. Math. Statist. **32**, 104.
Лаубшер (Laubscher N. F.) (1961), On stabilizing the binomial and negative binomial variances, J. Amer. Statist. Ass. **56**, 143.
Лахири (Lahiri D. B.) (1951), A method of sample selection providing unbiased ratio estimates, Bull. Int. Statist. Inst. **33**, (2), 133.
Левайн (Levine A.) (1966), A problem in minimax variance polynomial extrapolation, Ann. Math. Statist. **37**, 898.
Левен (Levene H.) (1952), On the power function of tests of randomness based on runs up and down, Ann. Math. Statist. **23**, 34.

Ледерманн (Ledermann W.) (1937), On the rank of the reduced correlational matrix in multiplefactor analysis, Psychometrika **2**, 85.
Лейпник (Leipnik R. B.) (1947), Distributon of the serial correlation coefficient in a circularly correlated universe, Ann. Math. Statist. **18**, 80.
Лейпник (Leipnik R. B.) (1958), Note on the characteristic function of a serial-correlation distribution, Biometrika **45**, 559.
Леман (Lehmann E. L.) (1959), Testing Statistical Hypotheses, Wiley, New York. (Русский перевод: Э. Леман, Проверка статистических гипотез, М., «Наука», 1964).
Леман (Lehmann E. L.) (1963a), Robust estimation in analysis of variance, Ann. Math. Statist. **34**, 957.
Леман (Lehmann E. L.) (1963b), Asymptotically nonparametric inference: an alternative approach to linear models, Ann. Math. Statist. **34**, 1494.
Леман (Lehmann E. L.) (1963c), Nonparametric confidence intervals for a shift parameter, Ann. Math. Statist. **34**, 1507.
Леман (Lehmann E. L.) (1964), Asymptotically nonparametric inference in some linear models with one observation per cell, Ann. Math. Statist. **35**, 726.
Линхэрт (Linhart H.) (1959), Techniques for discriminant analysis with discrete variables, Metrika **2**, 138.
Лоли (Lawley D. N.) (1939), A generalization of Fisher's z-test, Biometrika **30**, 180.
Лоли (Lawley D. N.) (1956a), Tests of significance for the latent roots of covariance and correlation matrices, Biometrika **43**, 128.
Лоли (Lawley D. N.) (1956b), A general method of approximating to the distribution of likelihood ratio criteria, Biometrika **43**, 295.
Лоли (Lawley D. N.) (1959), Tests of significance in canonical analysis, Biometrika **46**, 59.
Лоли (Lawley D. N.) (1963), On testing a set of correlation coefficients for equality, Ann. Math. Statist. **34**, 149.
Лоли и Максвелл (Lawley D. N. and Maxwell A. E.) (1963), Factor Analysis as a Statistical Method, Butterworths, London. (Русский перевод: Д. Лоули и А. Максвелл, Факторный анализ как статистический метод, М., «Мир», 1967 г.)
Ломницки (Lomnicki Z. A.) (1961), Tests for departure from normality in the case of linear stochastic processes, Metrika **4**, 37.
Лоу (Low L. Y.) (1964), Sampling variances of estimates of components of variance from a non-orthogonal two-way classification, Biometrika **51**, 491.
Любишев (Lubischew A. A.) (1962), On the use of discriminant functions in taxonomy, Biometrics **18**, 455.
Мак Дональд и Томпсон В. (McDonald B. J. and Thompson W. A.) (1967), Rank sum multiple comparisons in one- and two-way classifications, Biometrika **54**, 487.
Маллер (Muller E.-R.) (1965), A method of constructing balanced incomplete block designs, Biometrika **52**, 285.
Манн (Mann H. B.) (1945a), On a test for randomness based on signs of differences, Ann. Math. Statist. **16**, 193.
Манн (Mann H. B.) (1945b), Nonparametric tests against trend, Econometrica **13**, 245.
Манн (Mann H. B.) (1949), Analysis and Design of Experiments, Dover, New York.
Ман и Вальд (Mann H. B. and Wald A.) (1943), On the statistical treatment of linear stochastic difference equations, Econometrica **11**, 173.
Махаланобис (Mahalanobis P. C.) (1930), On tests and measures of group divergence, I, J. Proc. Asiat. Soc. Bengal **26**, 541.
Мёрфи (Murthy M. N.) (1957), Ordered and unordered estimators in sampling without replacement, Sankhyā **18**, 379.

Мёрфи (Murthy M. N.) (1963), Some recent advances in sampling theory, J. Amer. Statist. Ass. **58**, 737.
Мёрфи и Нанджамма (Murthy M. N. and Nanjamma N. S.) (1959), Almost unbiased ration estimates based on interpenetrating sub-sample estimates, Sankhyā **21**, 381.
Микки (Mickey M. R.) (1959), Some finite population unbiased ratio and regression estimators, J. Amer. Statist. Ass. **54**, 594.
Милл (Mill J. S.) (1843), A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence, and the Methods of Scientific Investigation, 2 vols, London.
Молдон (Mauldon J. G.) (1955), Pivotal quantities for Wishart's and related distributions, and a paradox in fiducial theory, J. R. Statist. Soc. **B**, **17**, 79.
Морэн (Moran P. A. P.) (1947—8), Some theorems on time series, I, II, The significance of the serial correlation coefficient, Biometrika **34**, 281; **35**, 255.
Морэн (Moran P. A. P.) (1949), The spectral theory of discrete stochastic processes, Biometrika **36**, 63.
Морэн (Moran P. A. P.) (1967), Testing for serial correlation with exponentially distributed variates, Biometrika **54**, 395.
Мостеллер и Ютц (Mosteller F. and Youtz C.) (1961), Tables of the Freeman-Tukey transformations for the binomial and Poisson distributions, Biometrika **48**, 433.
Мочли (Mauchly J. W.) (1940), Significance test for sphericity of a normal n-variate distribution, Ann. Math. Statist. **11**, 204.
Муд (Mood A. M.) (1951), On the distribution of the characteristic roots of normal second-moment matrices, Ann. Math. Statist. **22**, 266.
Мур и Уоллис (Moore G. H. and Wallis W. A.) (1943), Time series significance tests based on signs of differences, J. Amer. Statist. Ass. **38**, 153.
Муртейра (Murteira B.) (1951), Note on the variate differences of autoregressive series, Biometrika **38**, 479.
Мустафа (Mostafa M. G.) (1967), Note on testing hypotheses in an unbalanced random effects model, Biometrika **54**, 659.
Мэдоу (Madow W. G.) (1945), Note on the distribution of the serial correlation coefficient, Ann. Math. Statist. **16**, 308.
Мэдоу (Madow W. G.) (1948), On the limiting distributions of estimates based on samples from finite universes, Ann. Math. Statist. **19**, 535.
Мэнли (Manley G.) (1953 and later), The mean temperature of Central England, 1698—1952, Quart. J. R. Meteor. Soc. **79**, 242.
Мэра и Саранджи (Mehra K. L. and Sarangi J.) (1967), Asymptotic efficiency of certain rank tests for comparative experiments, Ann. Math. Statist. **38**, 90.
Мэрриотт и Поуп (Marriott F. H. C. and Pope J. A.) (1954), Bias in the estimation of autocorrelations, Biometrika **41**, 390.
Мэртн (Murty V. N.) (1961), An inequality for balanced incomplete block designs, Ann. Math. Statist. **32**, 908.
Нанджамма, Мёрфи и Сэти (Nanjamma N. S., Murthy M. N., and Sethi V. K.) (1959), Some sampling systems providing unbiased ratio estimators, Sankhyā **21**, 299.
Нарайн (Narain R. D.) (1950), On the completely unbiased character of tests of independence in multivariate normal systems, Ann. Math. Statist. **21**, 293.
Нейман (Neyman J.) (1934)*, On the two different aspects of the representative method; the method of stratified sampling and the method of purposive selection, J. R. Statist. Soc. **97**, 558.
Нейман (Neyman J.) (1938)*, Contribution to the theory of sampling human populations, J. Amer. Statist. Ass. **33**, 101.

Нейман и Скотт (Neyman J. and Scott E. L.) (1960), Correction for bias introduced by a transformation of variables, Ann. Math. Statist. **31**, 643.
Нерлов (Nerlove M.) (1964), Spectral analysis of seasonal adjustment procedures, Econometrica, **32**, 241.
Нётер (Noether G. E.) (1950), Asymptotic properties of the Wald-Wolfowitz test of randomness, Ann. Math. Statist. **21**, 231.
Нието и Паскуаль (Nieto de Pascual J.) (1961), Unbiased ratio estimators in stratified sampling, J. Amer. Statist. Ass. **56**, 70.
Нортон (Norton H. W.) (1939), The 7×7, squares. Ann. Eugen. **9**, 269 (список опечаток: **10**, 1).
Ньюмэн (Newman D.) (1939), The distribution of range in samples from a normal population, expressed in terms of an independent estimate of standard deviation, Biometrika **31**, 20.
Огава (Ogawa J.) (1961), The effect of randomization on the analysis of randomized block design, Ann. Inst. Statist. Math., Tokyo **13**, 105.
Огава (Ogawa J.) (1963), On the null-distribution of the F-statistic in a randomized balanced incomplete block design under the Neyman model, Ann. Math. Statist. **34**, 1558.
Олкин (Olkin I.) (1958), Multivariate ratio estimation for finite populations, Biometrika **45**, 154.
Парзен (Parzen E.) (1961), Mathematical considerations in the estimation of spectra. Comments on the discussions of Messrs Tukey and Goodman, Technometrics **3**, 167, 232.
Паттерсон (Patterson H. D.) (1950), Sampling on successive occasions with partial replacement of units, J. R. Statist. Soc. **B**, **12**, 241.
Патхак (Pathak P. K.) (1961a), Use of «order-statistic» in sampling without replacement, Sankhyā **A**, **23**, 409.
Патхак (Pathak P. K.) (1961b), On the evalution of moments of distinct units in a sample, Sankhyā **A**, **23**, 415.
Патхак (Pathak P. K.) (1962a), On simple random sampling with replacement, Sankhyā **A**, **24**, 287.
Патхак (Pathak P. K.) (1962b), On sampling with unequal probabilities, Sankhyā **A**, **24**, 315.
Патхак (Pathak P. K.) (1964a), On sampling schemes providing unbiased ratio estimators, Ann. Math. Statist. **35**, 222.
Патхак (Pathak P. K.) (1964b), On inverse sampling with unequal probabilities, Biometrika **51**, 185.
Патхак (Pathak P. K.) (1964c), Sufficiency in sampling theory, Ann. Math. Statist. **35**, 795.
Пиллаи (Pillai K. C. S.) (1956), On the distribution of the largest or the smallest root of a matrix in multivariate analysis, Biometrika **43**, 122.
Пиллаи (Pillai K. C. S.) (1964), On the distribution of the largest of seven roots of a matrix in multivariate analysis, Biometrika **51**, 270.
Пиллаи (Pillai K. C. S.) (1965), On the distribution of the largest characteristic root of a matrix in multivariate analysis, Biometrika **52**, 405.
Пиллаи (Pillai K. C. S.) (1966), On the non-central multivariate Beta distribution and the moments of traces of some matrices, Proc. Symp. Multiv. Analysis, Academic Press, New York.
Пиллаи (Pillai K. C. S.) (1967a), On the distribution of the largest root of a matrix in multivariate analysis, Ann. Math. Statist. **38**, 616.
Пиллаи (Pillai K. C. S.) (1967b), Upper percentage points of the largest roots of a matrix in multivariate analysis, Biometrika **54**, 189.
Пиллаи и Джаячандраи (Pillai K. C. S. and Jayachandran K.) (1967), Power comparisons of tests of two multivariate hypotheses based on four criteria, Biometrika **54**, 195.

Пирс (Pearce S. C.) (1963), The use and classification of non-orthogonal designs, J. R. Statist. Soc. A, **126**, 353.
Пирсон и Уилкс (Pearson E. S. and Wilks S. S.) (1933)*, Methods of statistical analysis appropriate for *k* samples of two variables, Biometrika **25**, 353.
Питмэн (Pitman E. J. G.) (1938), Significance tests which may be applied to samples from any population, III, The analysis of variance test, Biometrika **29**, 322.
Плэкетт (Plackett R. L.) (1950), Some theorems in least squares, Biometrika **37**, 149.
Плэкетт (Plackett R. L.) (1960), Models in the analysis of variance, J. R. Statist. Soc. **B**, **22**, 195.
Постен и Баргман (Posten H. O. and Bargmann R. E.) (1964), Power of the likelihood-ratio test of the general linear hypothesis in multivariate analysis, Biometrika **51**, 467.
Пури и Сен (Puri M. L. and Sen P. K.) (1967a), On robust estimation in incomplete block designs, Ann. Math. Statist. **38**, 1587.
Пури и Сен (Puri M. L. and Sen P. K.) (1967b), On some optimum nonparamtric procedures in two-way layouts, J. Amer. Statist, Ass. **62**, 1214.
Радж (Raj D.) (1956), Some estimators in sampling with varying probabilities without replacement, J. Amer. Statist. Ass. **51**, 269.
Радж (Raj D.) (1964), On double sampling for PPS estimation, Ann. Math. Statist. **35**, 900.
Радж и Хамис (Raj D. and Khamis S. H.) (1958), Some remarks on sampling with replacement, Ann. Math. Statist. **29**, 550.
Рао Дж. (Rao J. N. K.) (1963), On three procedures of unequal probability sampling without replacement, J. Amer. Statist. Ass. **58**, 202.
Рао Дж. (Rao J. N. K.) (1965), A note on estimation of ratios by Quenouille's method, Biometrika **52**, 647.
Рао Дж. (Rao J. N. K.) (1966), On the comparison of sampling with and without replacement, Rev. Int. Statist. Inst. **34**, 125.
Рао Дж. (Rao J. N. K.) (1967), The precision of Mickey's unbiased ratio estimator, Biometrika **54**, 321.
Рао Дж. и Уэбстер (Rao J. N. K. and Webster J. T.) (1966), On two methods of bias reduction in the estimation of ratios, Biometrika **53**, 571.
Рао Дж., Хартли и Кокрэн (Rao J. N. K., Hartley H. O. and Cochran W. G.) (1962), On a simple procedure of unequal probability sampling without replacement, J. R. Statist. Soc. **B**, **24**, 482.
Рао П. и Мудхолкар (Rao P. R. S. and Mudholkar G. S.) (1967), Generalised multivariate estimator for the mean of finite populations, J. Amer. Statist. Ass. **62**, 1009.
Рао C. (Rao C. R.) (1947), General methods of analysis for incomplete block designs, J. Amer. Statist. Ass. **42**, 541.
Рао C. (Rao C. R.) (1952), Advanced Statistical Methods in Biometric Research, Wiley, New York.
Рао C. (Rao C. R.) (1961), A study of BIB designs with replications 11 to 15, Sankhyā A, **23**, 117.
Рао C. (Rao C. R.) (1966), Discriminant function between composite hypotheses and related problems, Biometrika **53**, 339.
Рао C. и Слейтер (Rao C. R. and Slater P.) (1949), Multivariate analysis applied to differences between neurotic groups, Brit. J. Psychol. (Statist. Sec.) **2**, 17.
Рис (Rees D. H.) (1966), The analysis of variance of designs with many non-orthogonal classifications, J. R. Statist. Soc., **B**, **28**, 110.
Робсон (Robson D. S.) (1957), Applications of multivariate polykays to the theory of unbiased ration-type estimation, J. Amer. Statist. Ass. **52**, 511.

Робсон и Витхаясаи (Robson D. S. and Vithayasai C.) (1961), Unbiased componentwise ratio estimation, J. Amer. Statist. Ass. **56**, 350.
Розенблатт (Rosenblatt M.) (ed.) (1963), Proceedings of the Symposium on Time Series Analysis held at Brown University, June 11—14, 1962, Wiley, New York.
Рой Дж. и Шах (Roy J. and Shah K. R.) (1962), Recovery of interblock information, Sankhyā A, **24**, 269.
Рой C. (Roy S. N.) (1939), *p*-statistics or some generalizations in analysis of variance appropriate to multivariate problems, Sankhyā **4**, 381.
Рой C. (Roy S. N.) (1957), Some Aspects of Mutivariate Analysis, Wiley, New York.
Рой C. и Кобб (Roy S. N. and Cobb W.) (1960), Mixed model variance analysis with normal error and possibly non-normal other random effects, Part I, The univariate case, Ann. Nath. Statist. **31**, 939.
Романовский В. (Romanovsky V. I.) (1932), Sur la loi sinusoïdale limite, R. C. Circ. Mat. Palermo **56**, 82.
Романовский В. (Romanovsky V. I.) (1933), Sur une généralisation de la loi sinusoïdale limite, R. S. Circ. Mat. Palermo **57**, 130.
Росс (Ross A.) (1961), Variance estimates in «optimum» sample designs, J. Amer. Statist. Ass. **56**, 135.
Роудс (Rhodes E. C.) (1921), Smoothing, Tracts for Computers, no. 6, 1, Cambridge Univ. Press.
Рубин (Rubin J.) (1945). On the distribution of the serial correlation coefficient. Ann. Math. Statist. **16**, 211.
Себер (Seber G. A. F.) (1964a), Orthogonality in analysis of variance, Ann. Math. Statist. **35**, 705.
Себер (Seber G. A. F.) (1964b), Linear hypotheses and induced tests, Biometrika **51**, 41.
Сен А. (Sen A. R.) (1953), On the estimate of variance in sampling with varying probabilities, J. Indian. Soc. Agric. Statist. **5**, 119.
Сен П. (Sen P. K.) (1966), On nonparametric simultaneous çonfidence regions and tests for the one criterion analysis of variance problem, Ann. Inst. Statist. Math. Tokyo **18**, 319.
Сен П. (Sen P. K.) (1967), A note on the asymptotic efficiency of Friedman's χ_r^2-test, Biometrika **54**, 677.
Сен П. и Пури (Sen P. K. and Puri M. L.) (1967), On the theory of rank order tests for location in the multivariate one sample problem, Ann. Math. Statist. **38**, 1216.
Сёрл (Searle S. R.) (1958), Sampling variances of estimates of components of variance, Ann. Math. Statist. **29**, 167.
Силлитто (Sillitto G. P.) (1964) см. Такэючи (1962).
Симаик (Simaika J. B.) (1941), On an optimum property of two important statistical tests, Biometrika **32**, 70.
Сингх (Singh M. P.) (1967), The relative efficiency of some two-phase sampling schemes, Ann. Math. Statist. **38**, 937.
Сиотани (Siotani M.) (1964), Interval estimation for linear combinations of means, J. Amer. Statist. Ass. **59**, 1141.
Ситгривз (Sitgreaves R.) (1952), On the distribution of two random matrices used in classification procedures, Ann. Math. Statist. **23**, 263.
Слуцкий (Slutzky E.) (1937), The summation of random causes as the source of cyclic processes, Econometrica **5**, 105.
Смит K. (Smith K.) (1918), On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants, and the guidance they give towards a proper choice of the distribution of observations, Biometrika **12**, 1.
Смит C. (Smith C. A. B.) (1947), Some examples of discrimination, Ann. Eugen. **13**, 272 (список опечаток: **14**, opp. p. 279).

Снедекор (Snedecor G. W.) (1946), Statistical methods: applied to experiments in agriculture and biology, 4th edn, Iowa State College Collegiate Press, Ames, Iowa.
Спитволл (Spjøtvoll E.) (1967), Optimum invariant tests in unbalanced variance components models, Ann. Math. Statist. **38**, 422.
Спротт (Sprott D. A.) (1962), Listing of BIB designs from $r = 16$ to 20, Sankhyā **A**, **24**, 203.
Стейн (Stein C.) (1966), An approach to the recovery of inter-block information in balanced incomplete block designs, in F. N. David (ed.) Research Papers in Statistics (Wiley, New York).
Стивенс (Stevens W. L.) (1948), Statistical analysis of a non-orthogonal tri-factorial experiment, Biometrika **35**, 346.
Стьюарт (Stuart A.) (1952), The power of two differencesign tests, J. Amer. Statist. Ass. **47**, 416.
Стьюарт (Stuart A.) (1954), Asymptotic relative efficiencies of distribution-free tests of randomness against normal alternatives, J. Amer. Statist. Ass. **49**, 147.
Стьюарт (Stuart A.) (1956), The efficiencies of tests of randomness against normal regression, J. Amer. Statist. Ass. **51**, 285.
Стьюарт (Stuart A.) (1957), The efficiency of the records test for trend in normal regression, J. R. Statist. Soc. **B**, **19**, 149.
Стьюарт (Stuart A.) (1964), Multistage sampling with preliminary random stratification of first-stage units, Rev. Int. Statist. Inst. **32**, 193.
Стьюдент («Student») (1908), The probable error of a mean, Biometrika **6**, 1.
Сугияма (Sugiyama T.) (1967a), On the distribution of the largest latent root of the covariance matrix, Ann. Math. Statist. **38**, 1148.
Сугияма (Sugiyama T.) (1967b), Distribution of the largest latent root and the smallest latent root of the generalized B statistic and F statistic in multivariate analysis, Ann. Math. Statist. **38**, 1152.
Сэмпфорд (Sampford M. R.) (1962), Methods of cluster sampling with and without replacement for clusters of unequal sizes, Biometrika **49**, 27.
Сэмпфорд (Sampford M. R.) (1967), On sampling without replacement with unequal probabilities of selection, Biometrika **54**, 499.
Сэттерсвейт (Satterthwaite F. E.) (1941), Synthesis of variance, Psychometrika **6**, 309.
Сэттерсвейт (Sattertwaite F. E.) (1959), Random balance experimentation, Technometrics **1**, 111.
Сэшадри (Seshadri V.) (1966), Comparison of combined estimators in balanced incomplete blocks, Ann. Math. Statist. **37**, 1832.
Сюй (Hsu P. L.) (1939), On the distribution of roots of certain determinantal equations, Ann. Eugen. **9**, 250.
Такэючи (Takeuchi K.) (1962), A table of difference sets generating balanced incomplete blocks designs, Rev. Int. Statist. Inst. **30**, 361. Исправления: Силлитто (1964) и Клатуэрти и Левицкий (1968), в том же журнале **32**, 251 и **36**, 12.
Тамура (Tamura R.) (1966), Multivariate nonparametric several-sample tests, Ann. Math. Statist. **37**, 611.
Тин (Tin M.) (1965), Comparison of some ratio estimators, J. Amer. Statist. Ass. **60**, 294.
Тинтнер (Tintner G.) (1940), The variate difference method, Cowles Comm. Res. Econ. (Bloomington, Indiana) Monogr. no. 5.
Томпсон (Thompson W. A. Jr.) (1962), The problem of negative estimates of variance components, Ann. Math. Statist. **33**, 273.
Точер (Tocher K. D.) (1952), The design and analysis of block experiments, J. R. Statist. Soc. **B**, **14**, 45.
Трайон (Tryon R. C.) (1939), Cluster Analysis, Edwards Bros., Ann. Arbor, Michigan.

Тьюки (Tukey J. W.) (1949), One degree of freedom for non-additivity, Biometrics **5**, 232.
Тьюки (Tukey J. W.) (1951), Quick and dirty methods in statistics, II. Simple analyses for standart designs, Proc. 5th Annu. Conf. Amer. Soc. Qual. Contr. 189.
Тьюки (Tukey J. W.) (1952), Allowances for various types of error rates, Unpublished, Princeton University.
Тьюки (Tukey J. W.) (1953), The problem of multiple comparisons, Unpublished, Princeton University.
Тьюки (Tukey J. W.) (1956—7a), Variances of variance components, I. Balanced designs, II. The unbalanced single classification, Ann. Math. Statist. **27**, 722; **28**, 43.
Тьюки (Tukey J. W.) (1957b), On the comparative anatomy of transformations, Ann. Math. Statist. **28**, 602.
Уайз (Wise J.) (1955), The autocorrelation function and the spectral density function, Biometrika **42**, 151.
Уайз (Wise J.) (1956), Stationarity conditions for stochastic processes of the autoregressive and moving-average type, Biometrika **43**, 215.
Уайт (White J. S.) (1957), Approximate moments for the serial correlation coefficient, Ann. Math. Statist. **28**, 798.
Уайт (White J. S.) (1961), Asymptotic expansions for the mean and variance of the serial correlation coefficient, Biometrika **48**, 85.
Уилк и Кемпсорн (Wilk M. B. and Kempthorne O.) (1955), Fixed, mixed and random models, J. Amer. Statist. Ass. **50**, 1144.
Уилк и Кемпсорн (Wilk M. B. and Kempthorne O.) (1956), Some aspects of the analysis of factorial experiments in a completely randomized design, Ann. Math. Statist. **27**, 950.
Уилк и Кемпсорн (Wilk M. B. and Kemplthorne O.) (1957), Non-additivities in a latin square disign, J. Amer. Statist. Ass. **52**, 218.
Уилкинсон (Wilkinson G. N.) (1957), The analysis of covariance with incomplete data, Biometrics **13**, 363.
Уилкинсон (Wilkinson G. N.) (1958a), Estimation of missing values for the analysis of incomplete data, Biometrics **14**, 257.
Уилкинсон (Wilkinson G. N.) (1958b), The analysis of variance and derivation of standart errors for incomplete data, Biometrics **14**, 360.
Уилкинсон (Wilkinson G. N.) (1960), Comparison of missing value procedures, Australian J. Statist. **2**, 53.
Уилкс (Wilks S. S.) (1932)*, Certain generalizations in the analysis of variance, Biometrika **24**, 471.
Уилкс (Wilks S. S.) (1935)*, On the independence of k sets of normaly distributed statistical variables, Econometrica **3**, 309.
Уилкс (Wilks S. S.) (1946)*, Sample criteria for testing equality of means, equality of variances and equality of covariances in a normal multivariate distribution, Ann. Math. Statist. **17**, 257.
Уилльямс Е. (Williams E. J.) (1952), Some exact tests in multivariate analysis, Biometrika **39**, 17.
Уилльямс Е. (Williams E. J.) (1967), The analysis of association among many variates, J. R. Statist. Soc. **B**, **29**, 199.
Уилльямс У. (Williams W. H.) (1961), Generating undiased ration and regression estimators, Biometrics **17**, 267.
Уилльямс У. (Williams W. H.) (1962), On two methods of unbiased estimation with auxiliary variates, J. Amer. Statist. Ass. **57**, 184.
Уинтерс (Winters P. R.) (1960), Forecasting sales by exponentially weighted moving averages, Management Sci. **6**, 324.
Уиттекер (Whittaker E. T.) (1911), On the law whith governs the variations of S. S. Cygni, Month. Notes Roy, Astron. Soc. **71**, 686.

Уиттл (Whittle P.) (1951), Hypothesis Testing in Time Series Analysis, Almqvist & Wiksell, Uppsala.
Уиттл (Whittle P.) (1953a), The analysis of multiple stationary time series, J. R. Statist. Soc. **B, 15**, 125.
Уиттл (Whittle P.) (1953b), Estimation and information in stationary time-series, Ark. Mat. (Stockholm) **2**, 423.
Уиттл (Whittle P.) (1957), Curve and periodogram smoothing, J. R. Statist. Soc. **B, 19**, 38.
Уишарт (Wishart J.) (1928), The generalized product moment distribution in samples from a normal multivariate population, Biometrika **20A, 32**, 424.
Уишарт (Wishart J.) (1929), The correlation between product moments of any order in samples from a normal population, Proc. Roy. Soc. Edin. **49**, 78.
Уишарт (Wishart J.) (1948), Proofs of the distribution law of the second order moment statistics, Biometrika **35**, 55, 422.
Уишарт и Бартлетт (Wishart J. and Bartlett M. S.) (1932), The distribution of second order moment statistics in a normal system, Proc. Camb. Phil. Soc. **28**, 455.
Уишарт и Бартлетт (Wishart J. and Bartlett M. S.) (1933), The generalized product moment distribution in a normal system, Proc. Camb. Phil. Soc. **29**, 260.
Уокер А. (Walker A. M.) (1961), On Durbin's formula for the limiting generalized variance of a sample of consecutive observations from a moving-average process, Biometrika **48**, 197 (исправления: 476).
Уокер Г. (Walker G. T.) (1914), Correlation in seasonal variation of weather, III. On the criterion for the reality of relationships or periodicities, Indian Meteor. Dept. (Simla) Mem. **21** (9), 22.
Уокер Г. (Walker G. T.) (1931), On periodicity in series of related terms, Proc. Roy. Soc. **A, 131**, 518.
Уоллис и Мур (Wallis W. A. and Moore G. H.) (1941), A singnificance test for time series analysis, J. Amer. Statist, Ass. **36**, 401.
Уорд (Ward D. H.) (1963), Comparison of different systems of exponentially weighted prediction, The Statistician **13**, 173.
Уоркинг (Working H.)(1960), Note on the correlation of first differences of averages in a random chain, Econometrica **28**, 916.
Уэйгл (Wagle B. V.) (1962), Some Contributions to the Theory of Multivariate Analysis, Unpublished Ph. D. Thesis, Univ. London.
Уэлч (Welch B. L.) (1937), On the *z*-test in randomized blocks and latin squares, Biometrika **29**, 21.
Уэнг (Wang Y. Y.) (1967), A comparison of sevaral variance component estimators, Biometrika **54**, 301.
Фелледжи (Fellegi I. P.) (1963), Sampling with varying probabilities without replacement: rotating and non-rotating samples, J. Amer. Statist. Ass. **58**, 183.
Филлипс (Phillips A. W.) (1959), The estimation of parameters in systems of stochastic differential equations, Biometrika **46**, 67.
Финни (Finney D. J.) (1941), The joint distribution of variance ration based on a common error mean square. Ann. Eugen. **11**, 136.
Финни (Finney D. J.) (1952—), Statistical Method in Biological Assay, Griffin, London.
Финч (Finch P. D.) (1960), On the covariance determinants of moving-average and autoregressive models, Biometrika **47**, 194.
Фиск (Fisk P. R.) (1967), Stochastically dependent equations, Griffin, London.
Фишер (Fisher R. A.) (1929)*, Tests of significance in harmonic analysis, Proc. Roy. Soc. **A, 125**, 54.

Фишер (Fisher R. A.) (1935—), The Design of Experiments, Oliver and Boyd, Edinburgh.
Фишер (Fisher R. A.) (1936)*, The use of multiple measurements in taxonomic problems, Ann. Eugen. **7**, 179.
Фишер (Fisher R. A.) (1939)*, The sampling distribution of some statistics obtained from non-linear equations, Ann. Eugen. **9**, 238.
Фишер (Fisher R. A.) (1942)*, The theory of confounding in factorial experiments in relation to the theory of groups, Ann. Eugen. **11**, 341.
Фишер (Fisher R. A.) (1945)*, A system of confounding for factors with more than two alternatives, giving completely orthogonal cubes and higher powers, Ann. Eugen. **12**, 283.
Фортье и Соломон (Fortier J. J. and Solomon H.) (1966), Clustering procedures, Proc. Symp. Multiv. Analysis, Academic Press, New York.
Фостер (Foster F. G.) (1957—8), Upper percentage points of the generalized Beta distribution, II, III, Biometrika **44**, 441; **45**, 492.
Фостер и Рис (Foster F. G. and Rees D. H.) (1957), Upper percentage points of the generalized Beta distribution, I, Biometrika **44**, 237.
Фостер и Стьюарт (Foster F. G. and Stuart A.) (1954), Distribution-free tests in time-series based on the breaking of records, J. R. Statist. Soc. **B, 16**, 1.
Фридмэн М. (Friedman M.) (1937), The use of ranks to avoid the assumption of normality implicit in the analysis of variance, J. Amer. Statist. Ass. **32**, 675.
Фридмэн М. (Friedman M.) (1940), A comparison of alternative tests of significance for the problem of *m* rankings, Ann. Math. Statist. **11**, 86.
Фридмэн Х. и Рубин (Friedman H. P. and Rubin J.) (1967), On some invariant criteria for grouping data, J. Amer. Statist. Ass. **62**, 1159.
Фримэн Г. и Джефферс (Freeman G. H. and Jeffers J. N. R.) (1962), Estimation of means and standard errors in the analysis of non-orthogonal experiments by electronic computer, J. R. Statist. Soc. **B, 24**, 435.
Фримэн М. и Тьюки (Freeman M. F. and Tukey J. W.) (1950), Transformations related to the angular and the square root, Ann. Math. Statist. **21**, 607.
Хаавелмо (Haavelmo T.) (1943), The statistical implications of a system of simultaneous equations, Econometrica **11**, 1.
Ханани (Hanani H.) (1961), The existence and construction of balanced incomplete block designs, Ann. Math. Statist. **32**, 361.
Хансен и Гурвиц (Hansen M. H. and Hurwitz W. N.) (1943), On the theory of sampling from finite populations, Ann. Math. Statist. **14**, 333.
Хансен и Гурвиц (Hansen M. N. and Hurwitz W. N.) (1946), The problem of non-response in sample surveys, J. Amer. Statist. Ass. **41**, 517.
Хансен и Гурвиц (Hansen M. N. and Hurwitz W. N.) (1949), On the determination of optimum probabilities in sampling, Ann. Math. Statist. **20**, 426.
Хансен, Гурвиц и Мэдоу (Hansen M. H., Hurwitz W. N. and Madow W. G.) (1953), Sample Survey Methods and Theory, 2 vols, Wiley, New York.
Ханурав (Hanurav T. V.) (1967), Optimum utilization of auxiliary information: *πps* sampling of two units from a stratum, J. R. Statist. Soc. **B, 29**, 374.
Харвилл (Harville D. A.) (1967a), Estimability of variance components for the two-way classification with interaction, Ann. Math. Statist. **38**, 1508.
Харвилл (Harville D. A.) (1967b), Statistical dependence between random effects and the numbers of observations on the effects for the unbalanced one-way random classification, J. Amer. Statist. Ass. **62**, 1375.

Хармэн (Harman H. H.) (1960), Modern Factor Analysis, Univ. Chicago Press. (Русский перевод: Г. Харман, Современный факторный анализ, М., «Статистика», 1972).
Хартер (Harter H. L.) (1957), Error rates and sample sizes for range tests in multiple comparisons, Biometrics **13**, 511.
Хартли (Hartley H. O.) (1938), Studentization and large-sample theory, Suppl. J. R. Statist. Soc. **5**, 80.
Хартли (Hartley H. O.) (1955), Some recent developments in analysis of variance, Commun. Pure Appl. Math. **8**, 47.
Хартли (Hartley H. O.) (1959). Analytic studies of survey data in Istituto di Statistica (Roma), Volume in Onore di Corrado Gini, **1**, 213.
Хартли и Рао Дж. (Hartley H. O. and Rao J. N. K.) (1967), Maximum-likelihood estimation for the mixed analysis of variance model, Biometrika **54**, 93.
Хартли и Росс (Hartley H. O. and Ross A.) (1954), Unbiased ratio estimators, Nature **174**, 270.
Хейланд (Høyland A.) (1965), Robustness of the Hodges-Lehmann estimates for shift, Ann. Math. Statist. **36**, 174.
Хекст (Hext G.) (1964), A note on pre-whitening and recolouring, Stanford Univ. Dept. Statist. Tech. Rep. no. 13.
Хендерсон (Henderson C. R.) (1953), Estimation of variance and covariance components, Biometrics **9**, 226.
Хеннан (Hannan E. J.) (1960), The estimation of seasonal variation, Australian J. Statist. **2**, 1.
Хеннан (Hannan E. J.) (1961), Testing for a jump in the spectral function, J. R. Statist. Soc. **B**, **23**, 394.
Хигхем (Higham J. A.) (1882), On the adjustment of mortality tables, J. Inst. Actuar. **23**, 335.
Хигхем (Higham J. A.) (1885), On the graduation of mortality tables, J. Inst. Actuar. **25**, 245.
Хили и Тейлор (Healy M. J. R. and Taylor L. R.) (1962), Tables for power-law transformations, Biometrika **49**, 557.
Хиллс (Hills M.) (1966), Allocation rules and their error rates, J. R. Statist. Soc. **B**, **28**, 1.
Хиллс (Hills M.) (1967), Discrimination and allocation with discrete data, Applied Statistics **16**, 237.
Хинчин (Khintchin A. Ya.) (1932), Zu Birkhoff's Lösung des Ergodenproblems, Math. Ann. **107**, 485.
Хогг (Hogg R. V.) (1961), On the resolution of statistical hypotheses, J. Amer. Statist. Ass. **56**, 978.
Ходжес и Леман (Hodges J. L., Jr. and Lehmann E. L.) (1962), Rank methods for combination of independent experiments in analysis of variance, Ann. Math. Statist. **33**, 482.
Ходжес и Леман (Hodges J. L., Jr. and Lehmann E. L.) (1963), Estimates of location based on rank tests, Ann. Math. Statist. **34**, 598.
Холландер (Hollander M.) (1967), Rank tests for randomized blocks when the alternatives have an a priori ordering, Ann. Math. Statist. **38**, 867.
Холловэй и Данн (Holloway L. N. and Dunn O. J.) (1967), The robustness of Hotelling's T^2, J. Amer. Statist. Ass. **62**, 124.
Холт (Holt C. C.) (1957), Forecasting seasonals and trends by exponentially weighted moving averages, Carnegie Inst. Technol. Res. Memo. no. 52 (NONR 760(01)).
Хорвитц и Томпсон Д. (Horvitz D. G. and Thompson D. J.) (1952), A generalization of sampling without replacement from a finite universe, J. Amer. Statist. Ass. **47**, 663.
Хорснелл (Horsnell G.) (1953), The effect of unequal group variances on the F-test for the homogeneity of group means, Biometrika **40**, 128.

Хотеллинг (Hotelling H.) (1931), The generalization of Student's ratio, Ann. Math. Statist. **2**, 360.
Хотеллинг (Hotelling H.) (1936), Relations between two sets of variates, Biometrika **28**, 321.
Хотеллинг (Hotelling H.) (1951), A generalized T^2-test and measure of multivariate dispersion, Proc. 2nd Berkeley Symp. Math. Statist. and Prob., 23.
Хоув (Howe W. G.) (1955), Some contributions to factor analysis, U. S. Atom. Energy Comm. Rep. ORNL-1919.
Хоул (Hoel P. G.) (1937), A significance test for component analysis, Ann. Math. Statist. **8**, 149.
Хоул (Hoel P. G.) (1958), Efficiency problems in polynomial estimation, Ann. Math. Statist. **29**, 1134.
Хоул (Hoel P. G.) (1965a), Minimax designs in two dimensional regression, Ann. Math. Statist. **36**, 1097.
Хоул (Hoel P. G.) (1965b), Optimum designs for polynomial extrapolation, Ann. Math. Statist. **36**, 1483.
Хоул и Левайн (Hoel P. G. and Levine A.) (1964), Optimal spacing and weighting in polynomial prediction, Ann. Math. Statist. **35**, 1553.
Чако (Chacko V. J.) (1963), Testing homogeneity against ordered alternatives, Ann. Math. Statist. **34**, 945.
Чеймберс (Chambers J. M.) (1967), On methods of asymptotic approximation for multivariate distributions, Biometrika **54**, 367.
Чу (Chew V.) (1966), Confidence, prediction and tolerance regions for the multivariate normal distribution, J. Amer. Statist. Ass. **61**, 605.
Шатцов (Schatzoff M.) (1966a), Sensitivity comparison among tests of the general linear hypothesis, J. Amer. Statist. Ass. **61**, 415.
Шатцов (Schatzoff M.) (1966b), Exact distributions of Wilks's likelihood ratio criterion, Biometrika **53**, 347.
Шах (Shah K. R.) (1964), Use of inter-block information to obtain uniformly better estimators, Ann. Math. Statist. **35**, 1064.
Шеффе (Scheffé H.) (1953), A method for judging all contrasts in the analysis of variance, Biometrika **40**, 87.
Шеффе (Scheffé H.) (1956a), A «mixed model» for the analysis of variance, Ann. Math. Statist. **27**, 23.
Шеффе (Scheffé H.) (1956b), Alternative models for the analysis of variance, Ann. Math. Statist. **27**, 251.
Шеффе (Scheffé H.) (1959), The Analysis of Variance, Wiley, New York. (Русский перевод: Г. Шеффе, Дисперсионный анализ, М., Физматгиз, 1963).
Шискин (Shiskin J.) (1955), Seasonal computations on Univac, Amer. Statistician **9** (**1**), 19.
Шискин и Эйзенпресс (Shiskin J. and Eisenpress H.) (1957), Seasonal adjustments by electronic computer methods, J. Amer. Statist. Ass. **52**, 415.
Шметтерер (Schmetterer L.) (1960), On a problem of J. Neyman and E. Scott, Ann. Math. Statist. **31**, 656.
Шорак (Shorack G. R.) (1967), Testing against ordered alternatives in Model I analysis of variance: normal theory and non-parametric, Ann. Math. Statist. **38**, 1740.
Шривастава и Бознвич (Srivastava S. R. and Bozivich H.) (1962), Power of certain analysis of variance test procedures involving preliminary tests, Bull. Int. Statist. Inst. **39** (**3**), 133.
Эйзенпресс (Eisenpress H.) (1956), Rergession techniques applied to seasonal corrections and adjustments for calendar shifts, J. Amer. Statist. Ass. **51**, 615.
Эйзенхарт (Eisenhart C.) (1947), The assumptions underlying the analysis of variance, Biometrics **3**, 1.

Экман (Ekman G.) (1959), An approximation useful in univariate stratification, Ann. Math. Statist. **30**, 219.
Эндрюс (Andrews F. C.) (1954), Asymptotic behaviour of some rank tests for analysis of variance, Ann. Math. Statist. **25**, 724.
Энскомб (Anscombe F. J.) (1948), The transformation of Poisson, binomial and negative binomial data, Biometrika **35**, 246.
Энскомб (Anscombe F. J.) (1959), Quick analysis methods for random balance screening experiments, Technometrics **1**, 195.
Энскомб (Anscombe F. J.) (1961), Examination of residuals, Proc. 4th Berkeley Symp. Math. Statist. and Prob. **1**, 1.
Энскомб и Тьюки (Anscombe F. J. and Tukey J. W.) (1963), The examination and analysis of residuals, Technometrics **5**, 141.
Эштон, Хили и Липтон (Ashton E. H., Healy M. J. R. and Lipton S.) (1957), The descriptive use of discriminant functions in physical anthropology, Proc. Roy. Soc. **B**, **146**, 552.
Юдэн, Кемпсорн, Тьюки, Бокс, Хантер (Youden W. J., Kempthorne O., Tukey J. W., Box G. E. P. and Hunter J. S.) (1959), Discussion of the papers of Messrs Satterthweite and Budne, Technometrics **1**, 157 (authors' response to discussions: 184).
Юл (Yule G. U.) (1926), Why do we sometimes get nonsense-correlations between timeseries? — A study in sampling and the nature of time-series, J. R. Statist. Soc. **89**, 1.
Юл (Uule G. U.) (1927), On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers, Phil. Trans. **A**, **226**, 267.
Янг (Young D. H.) (1961), Quota fulfilment using unrestricted random sampling, Biomeyrika **48**, 333.
C-E-I-R (1968), Mathematical model Building in Economics and Industry, Griffin, London.

Собрания статей

R. A. Fisher, Contributions to Mathematical Statistics, Wiley, New York, 1950.
A Selection of Early Statistical Papers of J. Neyman, Cambridge Univ. Press, 1967.
The Selected Papers of E. S. Pearson, Cambridge Univ. Press, 1966.
«Student's» Collected Papers, Biometrika Office, University College London, 1942.
Selected Papers in Statistics and Probability by Abraham Wald, McGraw-Hill, New York, 1955; reprinted by Stanford Univ. Press, 1957.
S. S. Wilks, Collected Papers, Contributions to Mathematical Statistics, Wiley, New York, 1967.

УКАЗАТЕЛЬ

Автоковариации (autocovariance) 557; спектр, как их производящая функция 567; см. Автокорреляции, Временные ряды
Автокорреляции (autocorrelation) 557; в эргодических процессах 562—565; спектр, как их производящая функция 562, 567; автокорреляционная функция для непрерывных рядов 579; частные автокорреляции 583—585; смещение при оценивании 593—596; регрессия с автокоррелированными ошибками 677—679; см. Сериальные корреляции
Авторегрессионные ряды (autoregressive series) 573—578; и скользящие средние 574, 647—649, 655—660; уравнения Юла — Уокера 574, (упражнение 50.5) 684; дисперсия (упражнение 47.7) 586—587; представление в виде комбинированной последовательности рядов Маркова и Юла (упражнение 47.12) 587; матричное представление (упражнения 47.16—47.17) 588; теория сериальных корреляций для больших выборок 592—593; переменные разности (упражнения 48.15—48.16) 618; влияние начального значения на оценку 644—647; оценивание 649—653; частные автокорреляции 653; критерии согласия 653—654; со скользящими средними в качестве ошибок 660—661; схема, смешанная с регрессией 679—681; см. Марковские ряды, Юла ряды
Аддитивность (additivity) 28; критерий Тьюки ее проверки 43, (пример 35.3) 41; 193; преобразования ведущие к а. 141
Адичи (Adichie J. N.), устойчивость оценок в ДА 163
Андерсен (Andersen S. L.), устойчивость процедур ДА 147
Андерсон О. (Anderson O.), метод переменных разностей 535, 538
Андерсон Р. (Anderson R. L.), метод Иэйтса взвешенных квадратов от средних значений 50; ортогональные полиномы в ДА 79; *ковариационный* анализ 82; распределение сериальных корреляций 598—602; автокоррелированные ошибки 677, 679
Андерсон Т. (Anderson T. W.), проверка степени полиномиальной регрессии (упражнение 37.3) 169; нецентральное распределение Уишарта 367; распределение обобщенной дисперсии (упражнение 41.9) 370; критерии независимости 362; многомерный анализ 395; функции мощности для многомерных критериев 396; многомерная регрессия (упражнения 42.15—42.16) 398—399; собственные числа 413; дискриминация (упражнения 44.2—44.3) 471—472; автокоррелированные ошибки 679
Аренс (Arens B. E.), планирование полиномиальной регрессии 234
Арксинуса преобразования (angular transformations 143, (упражнения 37.4—37.5) 169—170
Армитедж (Armitage P.), расслоенный выбор 259, 262, (упражнения 39.14—39.15) 297
Арнольд (Arnold H. J.), влияние перестановок на распределение T^2 396
Архаров П. В., предельные теоремы для собственных чисел 413 (сноска)
Атикулла (Atiqullah M.), устойчивость процедур ДА 156
Афифи (Afifi A. A.), потерянные наблюдения 168 (упражнение 37.21) 174
Балакришнан (Balakrisnan T. R.), формирование слоев 268
Балмер (Bulmer M. G.), доверительные интервалы в модели II ДА 110, упражнение 36.15) 128
Баидерджи (Banerjee K. S.), разложение квадратичных форм от нормальных переменных 16
Бакрофт (Bancroft T. A.), метод Иэйтса взвешенных квадратов от средних значений 50; ортогональные полиномы в ДА 79; ковариационный анализ 82; объединения в модели II ДА (пример 36.6) 105
Баргман (Barqmann R. E.), мощность критерия ОП 396
Барнард Г. (Barnard G. A.), экспоненциальные веса 683
Барнард М. (Barnard M. M.), данные о черепах египтян (пример 42.3) 390—393.
Барнетт (Barnett V. Д.), канонический анализ исследований в образовании 426
Барра (Barra J. R.), обзор по латинским и ортогональным квадратам 200
Бартлетт (Bartlett M. S.), преобразования 135—136, 138, (упражнение 37.6) 170; асимптотическое распределение корреляционного определителя (пример 41.4) 353—354; распределение выборочной корреляции (упражнение 41.13) 370—371; разложение (упражнение 41.16) 371; анализ данных о черепах египтян (пример 42.3) 390—393; проверка гипотез о собственных числах 410; канонические корреляции 426; факторный анализ 432; дискриминация 454, 467; стандартные ошибки сериальных корреляций 591, (упражнения 48.2, 48.8—48.9) 615—616;

Экман (Ekman G.) (1959), An approximation useful in univariate stratification, Ann. Math. Statist. **30**, 219.
Эндрюс (Andrews F. C.) (1954), Asymptotic behaviour of some rank tests for analysis of variance, Ann. Math. Statist. **25**, 724.
Энскомб (Anscombe F. J.) (1948), The transformation of Poisson, binomial and negative binomial data, Biometrika **35**, 246.
Энскомб (Anscombe F. J.) (1959), Quick analysis methods for random balance screening experiments, Technometrics **1**, 195.
Энскомб (Anscombe F. J.) (1961), Examination of residuals, Proc. 4th Berkeley Symp. Math. Statist. and Prob. **1**, 1.
Энскомб и Тьюки (Anscombe F. J. and Tukey J. W.) (1963), The examination and analysis of residuals, Technometrics **5**, 141.
Эштон, Хили и Липтон (Ashton E. H., Healy M. J. R. and Lipton S.) (1957), The descriptive use of discriminant functions in physical anthropology, Proc. Roy. Soc. **B, 146**, 552.
Юдэн, Кемпсорн, Тьюки, Бокс, Хантер (Youden W. J., Kempthorne O., Tukey J. W., Box G. E. P. and Hunter J. S.) (1959), Discussion of the papers of Messrs Satterthweite and Budne, Technometrics **1**, 157 (authors' response to discussions: 184).
Юл (Yule G. U.) (1926), Why do we sometimes get nonsense-correlations between timeseries? — A study in sampling and the nature of time-series, J. R. Statist. Soc. **89**, 1.
Юл (Uule G. U.) (1927), On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers, Phil. Trans. A, **226**, 267.
Янг (Young D. H.) (1961), Quota fulfilment using unrestricted random sampling, Biomeyrika **48**, 333.
C-E-I-R (1968), Mathematical model Building in Economics and Industry, Griffin, London.

Собрания статей

R. A. Fisher, Contributions to Mathematical Statistics, Wiley, New York, 1950.
A Selection of Early Statistical Papers of J. Neyman, Cambridge Univ. Press, 1967.
The Selected Papers of E. S. Pearson, Cambridge Univ. Press, 1966.
«Student's» Collected Papers, Biometrika Office, University College London, 1942.
Selected Papers in Statistics and Probability by Abraham Wald, McGraw-Hill, New York, 1955; reprinted by Stanford Univ. Press, 1957.
S. S. Wilks, Collected Papers, Contributions to Mathematical Statistics, Wiley, New York, 1967.

УКАЗАТЕЛЬ

Автоковариации (autocovariance) 557; спектр, как их производящая функция 567; см. Автокорреляции. Временные ряды
Автокорреляции (autocorrelation) 557; в эргодических процессах 562—565; спектр, как их производящая функция 562, 567; автокорреляционная функция для непрерывных рядов 579; частные автокорреляции 583—585; смещение при оценивании 593—596; регрессия с автокоррелированными ошибками 677—679; см. Сериальные корреляции
Авторегрессионные ряды (autoregressive series) 573—578; и скользящие средние 574, 647—649, 655—660; уравнения Юла — Уокера 574, (упражнение 50.5) 684; дисперсия (упражнение 47.7) 586—587; представление в виде комбинированной последовательности рядов Маркова и Юла (упражнение 47.12) 587; матричное представление (упражнения 47.16—47.17) 588; теория сериальных корреляций для больших выборок 592—593; переменные разности (упражнения 48.15—48.16) 618; влияние начального значения на оценку 644—647; оценивание 649—653; частные автокорреляции 653; критерии согласия 653—654; со скользящими средними в качестве ошибок 660—661; схема, смешанная с регрессией 679—681; см. Марковские ряды, Юла ряды
Аддитивность (additivity) 28; критерий Тьюки ее проверки 43, (пример 35.3) 41; 193; преобразования ведущие к а. 141
Адичи (Adichie J. N.), устойчивость оценок в ДА 163
Андерсен (Andersen S. L.), устойчивость процедур ДА 147
Андерсон О. (Anderson O.), метод переменных разностей 535, 538
Андерсон Р. (Anderson R. L.), метод Уэйтса взвешенных квадратов от средних значений 50; ортогональные полиномы в ДА 79; *ковариационный* анализ 82; распределение сериальных корреляций 598—602; автокоррелированные ошибки 677, 679
Андерсон Т. (Anderson T. W.), проверка степени полиномиальной регрессии (упражнение 37.3) 169; нецентральное распределение Уишарта 367; распределение обобщенной дисперсии (упражнение 41.9) 370; критерии независимости 382; многомерный анализ 395; функции мощности для многомерных критериев 396; многомерная регрессия (упражнения 42.15—42.16) 398—399; собственные числа 413; дискриминация (упражнения 44.2—44.3) 471—472; автокоррелированные ошибки 679
Аренс (Arens B. E.), планирование полиномиальной регрессии 234
Арксинуса преобразования (angular transformations 143, (упражнения 37.4—37.5) 169—170
Армитедж (Armitage P.), расслоенный выбор 259, 262, (упражнения 39.14—39.15) 297
Арнольд (Arnold H. J.), влияние перестановок на распределение T^2 396
Архаров П. В., предельные теоремы для собственных чисел 413 (сноска)
Атикулла (Atiqullah M.), устойчивость процедур ДА 156
Афифи (Afifi A. A.), потерянные наблюдения 168 (упражнение 37.21) 174
Балакришнан (Balakrisnan T. R.), формирование слоев 268
Валмер (Bulmer M. G.), доверительные интервалы в модели II ДА 110, упражнение 36.15) 128
Бандерджи (Banerjee K. S.), разложение квадратичных форм от нормальных переменных 16
Бакрофт (Bancroft T. A.), метод Уэйтса взвешенных квадратов от средних значений 50; ортогональные полиномы в ДА 79; ковариационный анализ 82; объединения в модели II ДА (пример 36.6) 105
Варгман (Barqmann R. E.), мощность критерия ОП 396
Барнард Г. (Barnard G. A.), экспоненциальные веса 683
Барнард М. (Barnard M. M.), данные о черепах египтян (пример 42.3) 390—393.
Барнетт (Barnett V. Д.), канонический анализ исследований в образовании 426
Барра (Barra J. R.), обзор по латинским и ортогональным квадратам 200
Бартлетт (Bartlett M. S.), преобразования 135—136, 138, (упражнение 37.6) 170; асимптотическое распределение корреляционного определителя (пример 41.4) 353—354; распределение выборочной корреляции (упражнение 41.13) 370—371; разложение (упражнение 41.16) 371; анализ данных о черепах египтян (пример 42.3) 390—393; проверка гипотез о собственных числах 410; канонические корреляции 426; факторный анализ 432; дискриминация 454, 467; стандартные ошибки сериальных корреляций 591, (упражнения 48.2, 48.8—48.9) 615—616;

критерий для ординаты спектра 630; сглаживание спектра (пример 49.4) 635; многомерные временные ряды 676
Бартоломью (Bartholomew D. J.), альтернативы об упорядочениях (упражнение 35.15) 86—87
Басу (Basu D.), достаточность в теории выборочных обследований 245
Беверидж (Beveridge W.), индексы цен на пшеницу (таблица 47.1) 560—561; периодограмма 627
Белц (Belz M. H.), доказательство метода всех сравнений Шеффе (упражнение 35.19) 87
Бенард (Benard A.), ранговые критерии для СНБ 221
Бенкен (Behnken D. W.), ротатабельные планы 230
Беннет (Bennett B. M.), многомерные критерии средних 396, (упражнение 42.12) 398
Бёрмэн (Burman J. P.), сезонные изменения 551
Биггерс (Biggers J. D.), потерянные наблюдения 167
Бикел (Bickel P. J.), устойчивость оценок в ДА 163
Бил (Beale E. M. L.), модифицированная оценка-отношение (упражнение 40.14) 338; отбрасывание величин 414
Биномиальное распределение (binomial distribution), преобразования выборочных статистик 143, (упражнения 37.4, 37.6) 114—115; выбор (упражнения 39.13, 39.23) 297, 298; см. Отрицательное биномиальное распределение
Биркгоф (Birkhoff G. D.), эргодическая теорема 559, 562
Блисс (Bliss G. I.), дискриминация 461
Блоки (blocks) 182; и слои 262; см. Эксперименты, Случайные блоки
Блом (Blom G.), преобразования 141
Блэкис (Blackith R. E.), дискриминация 454
Блэкмэн (Blackman R. B.), спектральный анализ 636, 637, 638, (упражнение 49.5) 640
Бозе (Bose R. C.), ошибочность предположения Эйлера о греко-латинских квадратах 199; построение латинских квадратов 200; неравенство для разложимых СНБ 209, (упражнение 38.18) 238; построение СНБ 209—211; планы связанных парных сравнений 221—222; ЧСНБ 223; ротатабельные планы 230; D^2-статистика Махаланобиса 368; нецентральное распределение вектора средних (упражнение 41.1) 368
Бозивич (Bozivich H.), объединение в модели II ДА (пример 36.6) 103
Бокс (Box G. E. P.), приближенный F-критерий в ДА (упражнение 36.7) 126—127; распределение линейной функции от χ^2 величин (упражнение 36.14) 127—128; преобразования 129—133, (упражнения 37.1, 37.9) 168—170; анализ остатков 145; устойчивость процедур ДА 147, (упражнения 37.10—37.12) 171; случайные сбалансированные эксперименты 191; смешивание 229; эволюционная операция 230; асимптотические разложения критерия ОП 378—380; экспоненциальные веса 683
Борер (Bohrer R.), доверительные интервалы для сравнений в ДА 78
Браду (Bradu D.), неортогональная аддитивная трехфакторная перекрестная классификация 62
Брант (Brunt D.), данные об осадках (таблица 45.2) 477
Браун Г. (Brown G.W.), медианные критерии 161—162, (упражнения 37.18—37.20) 117—118
Браун Р. (Brown R. G.), экспоненциальные веса 683
Бриллинджер (Brillinger D. R.), полиспектр 670
Брюэр (Brewer K. R. W.), выбор с неравными вероятностями 247
Будиэ (Budne T. A.), случайные сбалансированные эксперименты 191
Бхапкар (Bhapkar V. P.), медианные критерии 163
Бхаттачария (Bhattacharyya G. K.), эффективность многомерных свободных от распределения критериев 396
Бхучонгкул (Bhuchongkul S.), устойчивость оценок в ДА 164
Бэбингтон Смит (Smith B. Babington), использование рангов при ДА (упражнение 37.14) 172
Бэйкер (Baker F. B.), устойчивость F-критериев для случайных блоков 203

Вайда (Vajda S.), латинские квадраты 200; планы СНБ 211; планы ЧСНБ 223
Вальд (Wald A.), доверительные интервалы в несбалансированной модели II ДА 113, (упражнение 36.16) 128; дискриминация 471, (упражнение 44.2) 471—472); критерий ранговой сериальной корреляции 498; метод НК для авторегрессионных рядов 650, 679
Ван Элтери (Van Eitern P.), ранговые критерии для СНБ 220—221
Ватсон (Watson G. S.), устойчивость процедур ДА 147, (упражнения 37.10—37.12) 171; совместное распределение сериальных корреляций 602, 615; регрессия с автокоррелированными ошибками 677
Вейсман (Wijsman R.), разложение Бартлетта (упражнения 41.17—41.18) 371
Веса (weights), их выбор в ДА для линейной модели 44—46
Взаимодействия (interactions) в ДА 28, 58; критерий Тьюки для них (пример 35.3) 41, 43; независимость их равенства нулю от системы весов 45; независимые и связанные 115—116; объекта и обработки 123
Вилкоксон (Wilcoxon F.), критерий 110, 203, 220
Винер (Wiener N.), автокорреляционная функция 581
Витхаясаи (Vithayasai C.), оценки-отношения 318
Вложенные гипотезы (nested hypotheses) 132, (упражнения 37.1—37.3) 168—169
Возвращения (replacement), выбор с в. и без в. 240; см. Обследования
Волд (Wold H.), ряды скользящих средних 572, 669; ряды авторегрессии 575, (упражнение 47.6) 586; модели «причинных цепочек» 680, 681
Волфовиц (Wolfowitz J.), планирование регрессии 234; критерий фаз 490; критерий ранговой сериальной корреляции 498
Вос (Vos J. W. E.), выбор, одновременный по времени и пространству (упражнение 40.10) 337

Временные ряды (time-series) 474—684; общая часть 474—481; составляющие в. р. 482—483, 506; критерий случайности 485—499; скользящие средние 506—554; стационарные 555—556; эргодические 559, 562; интенсивность 566; периодограмма 566; ряды скользящих средних 567—573; ряды авторегрессии 573—578; непрерывные ряды 578—582; фильтры и передаточные функции 582—583; бесконечные и циклические процессы 584—585; сериальные корреляции 499—501, 590—619; спектральная теория 562—566, 620—643; оценки и критерии для авторегрессионных рядов и рядов скользящих средних 644—562, 677—681; многомерные 662—676; системы уравнений 676—677; прогнозирование 681—683; см. Авторегрессионные, Марковские, Скользящих средних, Юла ряды, Автокорреляции, Переменных разностей метод, Сериальные корреляции, Случайности критерий, Спектр, Тренд
Выбор с отбрасыванием (rejective sampling) 251
Выбор с равными долями (ВРД) (uniform sampling fraction) 259
Выборка по группам (quota sampling), (упражнение 40.11) 337—338
Выборочное среднее (sample mean), его моменты 243—245
Выборочные обследования (sample surveys), см. Обследования
Выровненные ранговые критерии (aligned rank tests), в ДА 161

Габриэл (Gabriel K. R.), ДА средних значений в клетке 62; одновременные и ступенчатые критерии в ДА 72—74, 78, (упражнения 35.12, 35.17—35.18) 87
Гаек (Hájek J.), предельная нормальность 244—245; выбор с отбрасыванием 251
Гамма-распределение (gamma distribution), логарифмическое преобразование (пример 37.2) 136—137
Гардинер (Gardiner D. A.), ротатабельные планы 230
Гармонический анализ (harmonic analysis), см. Спектр
Гассиер (Gassner B. J.), планы СНБ (упражнение 38.20) 239
Гаучи (Gautschi W.), полнота расширенного экспоненциального семейства 96, 103
Гейлор (Gaylor D. W.), планирование полиномиальной регрессии 234
Гейссер (Geisser S.), многомерное нормальное распределение (упражнение 41.5) 369
Генеральное среднее (general mean), 23, 24, 28
Гербах (Herbach L. H.), проверка гипотез в модели II ДА 104, (упражнения 36.5—36.6), 126
Гесс (Hess I.), формирование слоев 268; оценки-отношения 318
Гест (Guest P. G.), планирование полиномиальной регрессии 233—234
Гири (Geary R. C.), распределение отношения 610
Гиршик (Girshick M. A.), собственные числа 365, 411—412, (упражнение 43.9) 435; канонические корреляции (упражнение 43.8) 435, (упражнение 43.10) 435
Главные компоненты (principal components) 402, см. Главных компонент анализ
Главные эффекты в ДА (main effects in AV), 28, 58
Главных компонент анализ (component analysis) 400—408; собственные числа и вектора, геометрическая интерпретация 401—406; нормировка 406; проверка гипотез о собственных числах 408—412; индексы деловой активности 413; метеорологические данные (пример 43.4) 414—419; см. Факторный анализ
Глезер (Gleser L. J.), критерий сферичности 383
Глейзберг (Gleissberg W.), критерий случайности 490
Гнездовая классификация (nested classification) 52 (сноска); см. Классификация иерархическая
Годэмб (Godambe V. P.), линейные оценки в выборочных обследованиях 252
Голдмэн (Goldman G. E.), дискриминация 458
Гомоскедастичность (homoscedasticity), критерий ОП ее проверки 132, (упражнения 37.1—37.3) 168—169; см. Однородность
Гопкинс (Hopkins C. E.), дискриминация 455, (пример 44.5) 457—458
Госсет (Gosset W. S.), см. Стьюдент
Гранди (Grundy P. M.), выбор с неравными вероятностями 250, (упражнение 39.11) 296—297
Грейбилл (Graybill F. A.), разложение квадратичных форм от нормальных переменных 16; модель II ДА 90, 95—98; информация между блоками 220
Греко-латинские квадраты (Graeco — Latin squares) 196—201, (упражнение 36.6) 235—236; при факторных экспериментах 226
Гренандер (Grenander U.), спектральный анализ 638; регрессия с автокоррелированными ошибками 679
Гринберг (Greenberg V. L.), устойчивость оценок в ДА 164
Группировка (clustering), при планировании выборки 269—271, 279, (упражнение 39.22) 298; при классификации 467—471
Грэндейдж (Grandage A. H. E.), ротатабельные планы 230
Грэнджер (Granger C.), спектральный анализ 638, 640, 670
Гуд (Good I. J.), циркулянтная матрица 544, (упражнение 46. 13) 553
Гудмэн (Goodman L. A.), несмещенные оценки-отношения 305, 307, 309
Гуерин (Guerin R.), обзор по СНБ и ЧСНБ 208—209, 223
Гурвиц (Hurwitz W. N.), выборочные обследования 241; выбор с пропорциональными объему вероятностями 281; подбор вероятностей извлечения 291; уклонение от ответа при обследованиях (упражнение 40.7) 335—336
Гхош Б. (Ghosh B. K.), последовательные процедуры ДА 121
Гхош М. (Ghosh M. N.), критерий Тьюки проверки аддитивности 43
Гхош С. (Ghosh S. P.), формирование слоев 268; двухэтапный выбор с группировкой 325

ДА, см. Дисперсионный анализ
Далениус (Dalenius T.), формирование слоев 265—268, (упражнение 39.17) 298

Даннэлл (Danieli P. J.), сглаживание спектра (пример 49.5) 636
Даниэлс (Danieis H. E.), приближения распределений сериальных корреляций 609—615; спектральный анализ 638
Данн (Dunn O. J.), доверительные интервалы для сравнений в ДА 78, (упражнение 35.14) 86; устойчивость критерия T^2 396
Дарлинг (Darling D. A.), объединение критериев ДА 68
Дас Гупта (Das Gupta S.), функция мощности многомерных критериев 396
Двумерное нормальное распределение (bivariate normal distribution), стабилизирующее дисперсию преобразование выборочного коэффициента корреляции (пример 37.3) 130—131; распределение Уишарта выборочных дисперсий и ковариаций (пример 41.1) 348; моменты выборочных ковариаций (пример 41.2) 349; нулевое распределение собственных чисел матрицы рассеяния (пример 41.7) 366—367; распределение выборочной корреляции (упражнение 41.13) 370—371
Двухфакторный перекрестный ДА, см. Классификация по двум признакам, перекрестная
Двухэтапная выборка (two-phase sampling) 319—325; с расслоением 319—322; функция стоимости 322—324; д. в. для оценок-отношений 324—326; для регрессионных оценок 325
Дейли (Daly J. F.), несмещенность критерия ОП независимости 382
Дербин (Durbin J.), ранговые критерии для СНБ 220, (упражнение 36.17) 238; оценки дисперсии в многошаговом выборе 286, 289, 294, 319 (упражнения 39.26—39.27) 298—299; выбор с неравными вероятностями (упражнения 39.2, 39.4) 294, 295; уменьшение смещения для оценок-отношений 308, 309; асимптотическая линейность оценок-отношений и регрессионных 318; области изучения 327, (упражнения 40.15—40.16) 339; уклонение от ответа при обследованиях (упражнение 40.8) 336; распределение сериальных корреляций 602; спектр и сезонные изменения 638; скользящие средние и авторегрессия 655—662, (упражнение 50.6) 684; регрессия с автокоррелированными ошибками 677—678; смешанная регрессионно-авторегрессионная схема 679; уравнения Юла — Уокера (упражнение 50.5) 684
Джаячандран (Jayachandran K.), мощности критериев для вектора средних и критерии независимости 396
Джеймс (James A. T.), собственные числа 368
Дженкинс (Jenkins G. M.), совместное распределение сериальных корреляций 598, 615, (упражнения 48.18—48.19) 618—619; распределение сериальных корреляций в марковском случае 607, 609, (упражнение 48.17) 618; спектральный анализ 636; экспоненциальные веса 683
Джефферс (Jeffers J. N. R.), неортогональная трехфакторная перекрестная классификация 62
Джон (John S.), дискриминация (пример 44.5) 457—458, (упражнение 44.2) 471—472
Джонсон А. (Johnson A. H. L.) преобразования квадратного корня (упражнение 37.15) 172
Джонсон Н. (Jonson N. L.), последовательные процедуры ДА 121; выборка по группам (упражнение 40.11) 337—338; метод переменных разностей 539
Диксон (Dixon W. J.), сериальные корреляции 604—605, (упражнение 48.20) 619
Дискриминация и классификация (discrimination and classification); 437—473; линейная дискриминация 438—443; геометрическая интерпретация 441, 453; квадратичная 322; критерии 448—449; с функцией стоимости 450; для k генеральных совокупностей 450—454; качественные данные 454—456; отказ от решения 455—456; смещение при оценивании вероятностей ложной классификации 456—458; избыточные переменные 468; стандартные ошибки коэффициентов 458—460; свободные от распределения методы 461—466; различия в характеристиках рассеяния 466—467; классификация 467—471
Дисперсионный анализ (ДА) (analysis of variance) в линейной модели (модель I) 11—87; определение 13, 18, 88; геометрическая интерпретация 24, 64; главные эффекты и взаимодействия 28, 58; выбор весов 44—46; пустые клетки 51; объединение критериев 64—69; множественные сравнения 69—78; ортогональные полиномы 79; альтернативы об упорядочениях 79; ковариационный анализ 79—82; в модели II 88—115, 117, 120; несмещенные квадратичные оценки 91—93; достаточные статистики 93—96, (упражнение 36.12) 127; соответствие с моделью I 96—98; проверка гипотез в модели II 102—106, 107—108, (упражнения 36.1—36.6, 36.8) 125—127; мощность, доверительные интервалы, отрицательные значения оценок в модели II 109—110, (упражнения 36.11, 36.15—36.16) 127—128; несбалансированный случай в модели II 110—113; обобщение моделей 113—121; смешанные модели 118—121; последовательные процедуры 121; размещение экспериментальных объектов, рандомизация 121—124; выбор модели 124—125; преобразования 129—142; анализ остатков 143—146; устойчивость 146—156; свободные от распределения методы 156—161; медианные критерии 161—164; потерянные наблюдения 164—168, (упражнение 37.21) 174; для блочных экспериментов 188—190; для случайных блоков 192—193; для латинских квадратов 196, (упражнение 38.6) 162; для СНБ 213; с использованием информации между блоками 218; в факторных экспериментах 226; см. Классификация по одному признаку и т. д.
Доксум (Doksum K.), устойчивость оценок в ДА 164; проверка гипотез против альтернатив об упорядочениях 203
Долби (Dolby J. L.), преобразования 131
Дополнительная информация (supplementary information) 301—339; см. Двухэтапная выборка, Оценки отношения, Регрессионные оценки
Достаточность (sufficiency), в модели II ДА 95—96, 112—113, (упражнение 36.12) 127; в теории выборочных обследований 245—246, 254, (упражнение 39.1) 294, (упражнения 39.30—39.31) 299—300
Дрэйпер (Draper N. R.), ротатабельные планы 230
Дункан (Duncan D. B.), критерии стюдентизированного размаха в ДА 73—74

Дэвид (David H. A.), парные сравнения 221—222; планирование полиномиальной регрессии 234
Дэвис (Davis H. T.), критерий в гармоническом анализе 630
Дэвьес (Davies O. L.), смешивание 229, эволюционная операция 230
Дэмпстер (Dempster A. P.), случайный выбор планов 190; многомерный критерий для двух выборок 359; многомерные оценки 372, 408; канонические корреляции 427
Дэрроч (Darroch J. N.), мощность многомерных критериев 396

Елашов Дж. (Elashoff J. D.), дискриминация 458
Елашов Р. (Elashoff R. M.), потерянные наблюдения 168, (упражнение 37.21) 174; дискриминация 458

Зейков (Zaycoff R.), метод переменных разностей 538
Знаков разностей критерий (difference-sign test) 491—494, 498, (упражнения 45.2—45.3, 45.6) 502—503

Иерархическая классификация, см. Классификация
Измерений ошибки (technical errors) 123
Имхоф (Imhof J. P.), смешанные модели в ДА 118, 121
Индекс деловой активности (index number), анализ главных компонент 413
Интенсивность (intensity) 566, см. Временные ряды
Информация между блоками (inter-block information) 213—220, (упражнения 38.14—38.16) 237—238
Инциденций матрица (incidente matrix), для эксперимента 183
Ито (Ito K.), устойчивость критерия T_0^2 396
Иэйтс (Yates F.), метод взвешенных квадратов от средних значений 50, (упражнения 35.5—35.7) 84; потерянные наблюдения 164, 168; планы СНБ 207; информация между блоками 216, 219; решетчатые планы 223; смешивание 229; обследования 241; выбор с неравными вероятностями 250; систематический выбор 270; оценки дисперсии в многошаговом выборе 286, 289, 294, 319; эффективность многошагового выбора 293; двухэтапный выбор 325; области изучения 327; последовательные возможности выбора (упражнения 40.9—40.10) 336—337

Калински (Caliński T.), анализ группы экспериментов 121
Канонические величины (canonical variables) 400—436; корреляции 419—427, (упражнения 43.5—43.8, 43.10) 435; стандартные ошибки 434—435; см. Главных компонент анализ, Собственные числа, Факторный анализ
Карлин (Karlin S.), оптимальные эксперименты 190
Картер (Carter R. L.), ротатабельные планы 230
Квадратичные формы (quadratic forms), их разложение 13—17
Квадратного корня (square root) преобразования, (пример 37.1) 135—136, 143, (упражнения 37.15—37.17) 172—173
Квазислучайный выбор (quasi-random sampling) 270
Кейлс (Keuls M.), критерий стьюдентизированного размаха в ДА 73—74, (упражнение 35.10) 85
Келли (Kelley T. L.), анализ главных компонент для психологических данных (пример 43.5) 423
Кемпсорн (Kempthorne O.), модели в ДА 115; полная рандомизация 123—124; случайные сбалансированные эксперименты 191; математическое ожидание СрК для случайных блоков и латинских квадратов 202; ДА для планов СНБ 213; анализ ЧСНБ 223; решетчатые планы 225; смешивание 229; последовательность экспериментов 229
Кендалл Д. (Kendall D. G.), логарифмическое преобразование 138
Кендалл М. (Kendall M. G.), использование рангов в ДА (упражнение 37.14) 172; n-мерная геометрия 345 (сноска); вычисления при анализе главных компонент 406; ранги при анализе главных компонент 413; отбрасывание величин 414; факторный анализ 432; классификация (пример 44.7) 470—471, пределы весов скользящих средних 511 (сноска); смещение для сериальных корреляций 596, (упражнения 48.4, 48.11) 615; 616; распределение сериальных корреляций в марковском случае 607
Кенуй (Quenouille M. H.), последовательность экспериментов 292; метод уменьшения смещения 308, 372, 427, 596; метод переменных разностей 542, (упражнения 46.7—46.11) 552—553; выравнивание тренда 542—544, 546, (упражнение 46.12) 553; теория больших выборок сериальных корреляций для рядов авторегрессии 593, (упражнение 48.14) 617—618; ненулевое распределение сериальных корреляций в марковском случае 607; преобразованное (упражнение 48.13) 617; совместное распределение сериальных корреляций 615; устойчивость теории сериальных корреляций 615; неравные временные интервалы во временных рядах 640; частные корреляции и критерий согласия для авторегрессионных рядов 653; многомерные временные ряды 662—666, 670, 676; ряды с общими ошибками (упражнение 50.8) 684
Кёртисс (Curtiss J. H.), преобразования 139
Кинг (King B.), методы формирования групп 469
Кифер (Kiefer J.), оптимальные эксперименты 190; планирование регрессии 231, 234—235
Киш (Kish L.), оценки дисперсии (упражнение 39.16) 297; оценки-отношения 318
Классификация, см. Дискриминация и классификация
Классификация гнездовая (nested classification) 52 (сноска); см. Классификация иерархическая
Классификация иерархическая (hierarchical classification), в модели I ДА 52—57; двухфакторная (пример 35.5) 53; трехфакторная 56—57, (упражнение 35.3)

84; в модели II 108—109, сбалансированная двухфакторная (пример 36.7) 108; сбалансированная трехфакторная (пример 36.8) 109
Классификация многофакторная (multiway classification), в модели I ДА 57—64, (упражнение 35.6) 84; в модели II 106—109, (упражнение 36.8) 127; критерии перестановок 160—161
Классификация по двум признакам, перекрестная (two-way cross-classification), в модели I ДА 25—52; пропорциональные частоты 34, (пример 35.2) 36; сбалансированная (пример 35.3) 41, (упражнение 35.2) 83—84; непропорциональные частоты 46, (пример 35.4) 47, 48—50, (упражнения 35.4—35.5, 35.7) 84—85, (упражнения 37.7—37.8) 170; пустые клетки 51; сбалансированная в модели II (примеры 36.4, 36.6) 100, 105, (упражнение 36.6) 126; несбалансированная в модели II 113; в общей модели 115—121; критерии перестановок 157—160; медианные критерии 161—163, (упражнения 37.18—37.20) 173
Классификация по одному признаку (one-way classification), в модели I ДА (пример 35.1) 19, (упражнение 35.1), 83; сбалансированная в модели II, (примеры 36.1—36.3, 36.5) 90, 94, 98, 104, (упражнения 36.1, 36.3, 36.5, 36.10—36.11, 36.13) 125—127; несбалансированная модель II 110—113, (упражнения 36.12—36.13) 127
Классификация смешанная (mixed classification), перекрестная и иерархическая 57—58
Классификация трехфакторная перекрестная (three-way cross-classification), в модели I ДА 58—64; сбалансированная (пример 35.6) 62, (упражнение 35.8) 85; непропорциональные частоты 62; сбалансированная в модели II 105—107; сбалансированная в смешанной модели 118; критерии перестановок 160—161
Классификация, упорядоченная и метрическая (ordered and metrical) 79, (упражнение 35.15) 86—87
Классифицированные данные (classified data) 26, ДА для них, см. Классификация иерархическая, многофакторная и т. д.
Клатуэрти (Clatworthy W. H.), таблицы СНБ 211; ЧСНБ 223
Кобб (Cobb W.), смешанная модель ДА 121
Ковариационный анализ (analysis of covariance) 79—82, 157, 301 (сноска)
Кокрэйн (Cochrane D.), регрессия с автокоррелированными ошибками 677—678
Кокрэн (Cochran W. G.), разложение СК 12; объединение критериев ДА 68; устойчивость процедур ДА 147; статистические выводы в выборочных обследованиях 176; планы СНБ 211; ДА для планов СНБ 213; решетчатые планы 225; смешивание 229; последовательность экспериментов 229; выборочные обследования 241; формирование слоев 267—268; (упражнение 39.18) 298; систематический выбор 270; оценки-отношения и регрессионные 318; двухэтапный выбор 325; области изучения 327; случайное формирование слоев (упражнение 40.6) 335; дискриминация 455, (пример 44.5) 457—458, 458, 461, (упражнения 44.6—44.9) 473
Кокс Г. (Cox G. M.), планы СНБ 211; ДА для планов СНБ 213; решетчатые планы 225; смешивание 229; последовательность экспериментов 229
Кокс Д. (Cox D. R.), ковариационный анализ 80; последовательные процедуры ДА 121; преобразования 129—133, (упражнение 37.1) 168—169; анализ остатков 145; статистические выводы из экспериментов 181; математическое ожидание СрК для латинских квадратов 202; сериальные корреляции 615; экспоненциальные веса 683
Кокэн (Kokan A. R.), расслоенный выбор 264
Кольер (Kollier R. O. Jr.), устойчивость F-критериев для случайных блоков 203
Компоненты дисперсии (components of variance) 90; см. ДА в модели II; канонические, см. Главных компонент анализ
Конечные совокупности (finite populations) 240; см. Обследования
Коннор (Connor L. R.), смешивание 229, эволюционная операция 230
Константин (Constantine A. G.), распределение T_0^2 критерия 396
Консул (Consul P. C.), распределение критерия ОП 382, 383, 393
Корин (Korin B. P.), критерий ОП для матрицы рассеяния 383
Корнфилд (Cornfield J.), математические ожидания СрК в ДА 107; модели в ДА 115, 118
Коррелограмма (correlogram) 601, 557; к. и спектр 563—565; см. Временные ряды
Корреляции коэффициент (correlation coefficient), стабилизирующее дисперсию преобразование для выборки из нормальной совокупности (пример 37.3) 139—140; канонический 419—427; см. Автокорреляции, Сериальные корреляции
Корреляционный определитель (correlation determinant), моменты и асимптотическое распределение (примеры 41.4—41.5) 352—355; в критерии ОП проверки независимости 382
Коуден (Kowden D. J.), скользящие средние 517
Круг Г. К., планирование экспериментов 230 (сноска)
Крускал (Kruskal J. B.), монотонные преобразования 133
Крэддок (Craddock J. M.), анализ главных компонент для метеорологических данных (пример 43.4) 414—419
Крэйг (Craig A. T.), независимость квадратичных форм 14
Куад (Quade D.), ранговые критерии в ковариационном анализе 157
Кузин (Cousins W. R.), смешивание 229, эволюционная операция 230
Куп (Koop J. C.), линейные оценки в выборочных обследованиях 252
Купмэнс (Koopmans T. C.), сериальные корреляции 605
Кширсагар (Kshirsagar A. M.), многомерное бетта-распределение 368, разложение Бартлетта (упражнения 41.16—41.17) 371

Латинские квадраты (latin squares) 196—201, (упражнения 38.5—38.7) 235—236; устойчивость к нормальной теории 201—204; при факторных экспериментах 226
Лаубшер (Laubscher N. F.), преобразования в ДА (упражнения 37.4—37.5) 169—170

Лахири (Lahiri D. B.), выбор с неравными вероятностями (упражнение 39.11) 296—297; устранение смещения для оценок-отношений 317
Левайн (Levine A.), планирование полиномиальной регрессии 234
Левен (Levene H.), критерий случайности 490, 491, (упражнения 45.6—45.7) 503
Левицкий (Lewyckyj R. J.), таблицы СНБ 211
Ледерман (Ledermann W.), анализ главных компонент 401
Лейпник (Leipnik R. B.), сериальные корреляции 605—607, 609
Леман (Lehmann E. L.), выровненные ранговые критерии 161; устойчивость оценок в ДА 163—164; многомерные критерии 395
Линейная модель (linear model), ДА в ней (модель 1) 11—87; разложение нецентральных квадратичных форм 13—17; устранение сингулярности 27—28; выбор весов 44—46; общий случай непропорциональных частот 62; объединение критериев 64—69; множественные сравнения 69—78; ковариационный анализ 79—82; расширение для учета новых параметров 81—82; преобразования к ней 129—134; потерянные наблюдения 164—168, (упражнение 37.21) 174; для блочных экспериментов 184—190; для выбора без возвращения 241—243; многомерный случай 385—389, (пример 42.3) 390—393; см. Дисперсионный анализ, Классификация
Линхэрт (Linhart H.), дискриминация 455
Липтон (Lipton S.), данные о представителях мужского пола (пример 42.4) 394—395
Логарифмические преобразования (logarithmic transformations), (примеры 37.2, 37.4) 136—137, 140, 143
Логит преобразования (logit transformations) 142
Лоли (Lawley D. N.), распределение статистики ОП 380; критерий T_0^2 396; анализ главных компонент (пример 43.2) 406—408, 410; канонические корреляции 425, 426; факторный анализ 429—432, (упражнение 43.11) 436
Ломницки (Lomnicki Z. A.), критерий нормальности стационарного процесса (упражнение 49.8) 641
Лоу (Low L. Y.) оценки в несбалансированной модели II ДА 113
Льюис (Lewis T.), канонический анализ исследований в образовании 426
Любишев (Lubischew A. A.), дискриминация 467

Мак Дональд (McDonald B. J.), методы множественных сравнений ранжированных сумм 74
Максвелл (Maxwell A. E.), анализ главных компонент (пример 43.2) 406—408; факторный анализ 429, 432, (упражнение 43.11) 436
Маллер (Muller E. R.), планы СНБ 211
Малютов М. Б., планирование оптимального эксперимента 230 (сноска)
Манн Д. (Mann D. W.), отбрасывание величин 414
Манн Х. (Mann H. B.), латинские квадраты 200; построение СНБ 209; смешивание 229; критерий знаков разностей 494, (упражнение 45.3) 502; критерий ранговой корреляции 496; метод НК для авторегрессионных рядов 650, 679
Марковские ряды (Markoff series), автокорреляции (пример 47.2) 558; обратные соотношения (пример 47.3) 558; коррелограмма и спектр (пример 47.7) 575—576; частные автокорреляции 583—584; семиинварианты и нормальность (упражнение 47.2) 586, (упражнение 47.3) 586; группы членов ряда (упражнение 47.15) 588; стандартные ошибки сериальных корреляций (пример 48.3) 592; ковариации сериальных корреляций (пример 48.4) 592—593; смещение в оценках сериальных корреляций (пример 48.7) 595—596, (упражнения 48.4—48.5) 615; учет более высоких порядков 696; «ненулевое» распределение сериальных корреляций 606—607, 611—615, (упражнения 48.13, 48.17) 617, 618; влияние начального значения 645—647; многомерные 676; в прогнозировании 682
Марсалья (Marsaglia G.), разложение квадратичных форм от нормальных переменных 16
Матрица рассеяния (dispersion matrix), ее собственные числа 361; критерий ОП для м. р. 383; оценки м. р. с использованием СК остатков 386, 387—389; см. Обобщенная дисперсия, Собственные числа
Махаланобис (Mahalanobis P. C.), D^2-статистика и обобщенное расстояние 367—368
МД, см. Минимальная дисперсия
Медианные критерии (median tests), в ДА 161—164, (упражнения 37.18—37.20) 173
Мешающие факторы (nuisance factors) 182; два 193; три или более 198—201
Мёрфи (Murthy M. N.), выборочные обследования 241; достаточность при обследованиях 254, (упражнения 39.30—39.31) 299—300; несмещенные оценки-отношения 318, (упражнение 40.1) 334
МЗР, см. Минимальная значимая разность
Микки (Mickey M. R.), несмещенные регрессионные оценки 313
Милл (Mill J. S.), правила экспериментальных исследований 176
Минимальная дисперсия (МД) (minimum variance), размещение по слоям с МД 260, 263
Минимальная значимая разность (МЗР) (least significant difference), критерий МЗР в ДА 70—71
Многомерный анализ (multivariate analysis) 340—473; для временных рядов 662—676; см. Главных компонент анализ, Дискриминация и классификация, Канонические величины, Корреляционный определитель, Матрица рассеяния, Независимость, Обобщенная дисперсия, Однородность, Регрессия, Собственные числа, Сферичности критерий, Уишарта распределение, Факторный анализ
Многошаговый выбор (multi-stage sampling) 272—294, (упражнения 39.24, 39.28, 39.29) 298, 299; оценки 273; м. в. с равными вероятностями 274—280; с неравными вероятностями 280—284; оценки значения дисперсии 284—289, 318—319; функция стоимости и МД 290; подбор вероятностей 291—293; эффективность 293; расслоение 294; оценки-отношения и регрессионные 318—319; области изучения 331—333

Многоэтапная выборка (multi-phase sampling), 325
Множественные сравнения (multiple comparisons), в ДА 69—78
Молдон (Mauldon J. G.), парадоксы многомерных оценок 396
Морэн (Moran P. A. P.), теорема Слуцкого о синусоидальном пределе 571; моменты сериальных корреляций 596—598, 615, (упражнения 46.5—48.6, 48.12, 48.21) 615—617, 619
Мостеллер (Mosteller F.), таблицы преобразований 136, (упражнение 37.4) 169
Мочли (Mauchly J. W.), критерий сферичности 383
Муд (Mood A. M.), медианные критерии 161—162, (упражнения 37.18—37.20) 173; собственные числа 365—366
Мудхолкар (Mudholkar G. S.), многомерные оценки при обследованиях 308; функции мощности многомерных критериев 396
Мур (Moore A. M.), критерии случайности 490, 492
Муртейра (Murteira B.), метод переменных разностей (упражнения 48.15—48.16) 618
Мустафа (Mostafa M. G.), критерий в несбалансированной модели II ДА 113
Мэдоу (Madow W. G.), выборочные обследования 241; предельная нормальность 244—245; сериальные корреляции 605—606
Мэнли (Manley G.), метеорологические данные (пример 43.4) 414
Мэра (Mehra K. L.), выровненные ранговые критерии 161
Мэрриотт (Marriott F. H. C.), смещение в оценках сериальных корреляций 596
Мэртн (Murty V. N.), неравенство для СНБ (упражнение 38.13) 237

Наименьшие квадраты (НК) (least squares), при выборе без возвращения 241—243; в дискриминации 449, (упражнения 44.10—44.11) 473; для скользящих средних 505—506, 517—518, 543; в авторегрессионных рядах 650, 679; см. Линейная модель
Налимов В. В., планирование экстремальных экспериментов 230 (сноска)
Нанджамма (Nanjamma N. S.), несмещенные оценки-отношения 318, (упражнение 40.1) 334
Нарайн (Narain R. D.), несмещенность критерия ОП независимости 382
Независимость (independence), критерий ОП ее проверки (упражнения 41.10—41.11) 370, 381—382, 395—396, (упражнение 42.4) 396—397; мощность различных критериев 396; эквивалентность проверке равенства собственных чисел 409
Нейман (Neyman J.), смещение, вызванное преобразованием 142; расслоенный выбор 259; двухэтапная выборка 319
Неравные вероятности (unequal probabilities), при выборе без возвращения 246—254, (упражнения 39.2—39.11) 294—295, (упражнения 39.30—39.31) 299—300; линейные оценки 249—253; с возвращением 254—256; н. в. и расслоение 256—258; и группировка 269—271; и многошаговый выбор 271—273, 280—294; выбор с пропорциональными объему вероятностями 281—284; оценка значения дисперсии 284—289; их подбор для МД 291—293; их подбор для устранения смещения 317—318; двухэтапная выборка для их определения 325; см. Многошаговый выбор, Обследования, Расслоенный выбор
Нерлов (Nerlove M.), спектральный анализ 637
Нецентральные квадратичные формы (non-central quadratic forms), их разложение 13—17
Нётер (Noether G. E.), ранговые критерии для СНБ 220; критерий ранговый сериальной корреляции 498
Нито де'Паскуаль (Nieto de Pascual J.), несмещенные оценки-отношения 304, 306, 309, 318
НК, см. Наименьшие квадраты
Нормализующие преобразования (normalizing transformations) 139—141
Нормальное распределение (normal distribution), логарифмическое преобразование выборочной дисперсии (примеры 37.2, 37.4) 136—137, 140; преобразование извлечения квадратного корня для выборочной дисперсии (пример 37.5) 140—141; см. Двумерное нормальное, Многомерный анализ
Нормальные метки (normal scores), преобразования к ним 142; их использование в ДА 157, 159—161, 203
Нортон (Norton H. W.), обзор по латинским квадратам 200
Ньюмэн (Newman D.), критерий стьюдентизированного размаха в ДА 73, 74, (упражнение 36.10) 85
Нэйр (Nair K. R.), ЧСНБ 223

Области изучения (domain of study) 326—333; пересекающие слои 327—331; (упражнения 40.15—40.16) 339; при многошаговом выборе 331—333
Обобщенная дисперсия (generalized variance) (определитель рассеяния), распределение и моменты (пример 41.3) 350—352, (пример 41.5) 354—355, (упражнения 41.8—41.9) 370; ее оценки 372
Обработки (treatments) 182; ДА для них 226
Обратный выбор (inverse sampling), (упражнение 40.3) 334
Обследования (surveys), сравнение с экспериментами 175—176, 262; теория 240—339; случайный выбор без возвращения 241—243; моменты выборочного среднего 243—245; достаточность 245—246; см. Двухэтапная выборка, Многошаговый выбор, Неравные вероятности, Области изучения, Оценки-отношения, Расслоенный выбор, Регрессионные оценки
Объединение критериев ДА (combination of AV tests) 64—69
Объединения процедуры (pooling procedures) в ДА, (пример 36.6) 105; при регрессии (упражнение 37.3) 169
Объекта ошибки (unit errors) 123
Огава (Ogawa J.), устойчивость F-критериев для случайных блоков и СНБ 203, 220
Одновременные процедуры проверки (simultaneous test procedures) 71—78, (упражнения 35.11—35.14, 35.16—36.17, 36.19) 85—87
Однородность (homogeneity), критерий ОП ее проверки 373—380, (пример 42.1) 381, (пример 42.2) 383—385, (упражнения 42.1—42.3) 396; см. Гомоскедастичность

Однофакторный ДА в модели I, см. Классификация по одному признаку
Олкин (Olkin J.), многомерные оценки при обследованиях 308, 318
ОП, см. Отношение правдоподобия
Опознаваемость элементов (recognizible individuals), в теории выборочных обследований 240, 245—246, 251—252
Оркутт (Orcutt G. H.), регрессия с автокоррелированными ошибками 677—678
Ортогональные квадраты (orthogonal squares) 199, (упражнения 38.6—38.7) 235—236; при факторных экспериментах 226
Остатки (residuals), их анализ 143—146; их матрица рассеяния 386, 387—389
Отношение правдоподобия (ОП) (Likelihood ratio), критерий ОП, в модели II ДА (упражнения 36.5—36.6) 126; для вложенных гипотез 132, (упражнения 37.1—37.3) 168—169; в многомерном анализе 373—399; см. Матрица рассеяния, Независимость, Однородность, Регрессия, Сферичность
Отрицательное биномиальное распределение (negative binomial bistribution), преобразования арксинуса 143, (упражнение 37.5) 169—170
Оценка-произведение (product estimator), (упражнение 40.2) 334
Оценки-отношения (ratio estimators), смещенность 301—302, 308—311, 317—318; состоятельность 302; модификация 303; (упражнения 40.1—40.2) 334, (упражнения 40.13, 40.14) 338; сравнение дисперсий 304—311; при расслоенном и многошаговом выборе 318—319; асимптотическая линейность 318; при двухэтапном выборе 324—325
Ошибки объекта, взаимодействий и измерений (unit, interactive and technical errors) 123

Парзен (Parzen E.), спектральный анализ 636, 637, (упражнения 49.6—49.7) 640—641
Паркер (Parker R.), ошибочность предположения Эйлера о греко-латинских квадратах 199
Парные сравнения (paired comparisons) 221
Паттерсон (Patterson H. D.), последовательная возможность выбора (упражнения 40.9—40.10) 336—337
Патхак (Pathak P. K.), достаточность в теории выборочных обследований 245—246, (упражнение 39.30) 299—300; 318, (упражнение 40.3) 334—335
Передаточные функции (transfer function) 582—583; см. Временные ряды
Переменных разностей метод (variate-difference method) 531—542, (упражнения 46.7—46.11) 552—553, (упражнения 48.15—48.16) 618
Перестановок критерии (permutation tests), в ДА 157—161, 203, 220
Периодограмма (periodogram) 566, см. Временные ряды, Спектр
Пиллаи (Pillai K. C.), распределение собственных чисел 367, 368; мощности критериев для вектора средних и критерии независимости 396
Пирс (Pearce S. C.), обзор по неортогональному ДА 62
Пирсон (Pearson E. S.), критерии однородности (пример 42.2) 383
Питмэн (Pitman E. J. G.), критерии перестановок 159, (упражнение 37.13) 171
Планирование (design), его задачи 175—176; уравнение плана 206; см. Обследования, Эксперименты
Плиз (Please N. W.), дискриминация 467
Плэкетт Plackett R. L.), модели в ДА 115, 124; взятые дважды наблюдения в ДА 168
Поверхности отклика (response surfaces), 230
Полиномы (polinomial), проверка степени при регрессии (упражнение 37.3) 169; планирование регрессии 230—236
Полная рандомизация (complete randomization) 122
Полнота (completness) расширенного экспоненциального семейства 103
Последовательные возможности выбора (sampling on successive occasions), (упражнения 40.9—40.10) 336—337
Последовательные процедуры ДА (sequential AV procedures) 121
Постен (Posten H. O.), мощность критерия ОП 396
Потерянные наблюдения (missing observations) 164—168, (упражнение 37.21) 174
Поуп (Pope J. A.), смещение в оценках сериальных корреляций 596
Предпочтения (preference), эксперименты с п. 221
Преобразования (transformations), к нормальной линейной модели 129—134; их цели 133—134; монотонные 133; стабилизирующие дисперсии 134—139; нормализующие 139—141; ведущие к аддитивности 141—142; устраняющие смещение 142—143; анализ остатков 142—145; см. Арксинуса, Квадратного корня и Логарифмические преобразования
Пробит преобразования (probit transformations) 142
Пропорциональные объему вероятности (probabilities proportional to size) 280—284, 293; см. Неравные вероятности, Обследования
Пропорциональный выбор 259; см. ВРД
Пуассона распределение (Poisson distribution), извлечение квадратного корня (пример 37.1) 135—136, упражнения 37.15—37.16) 172—173; выбор (упражнения 39.13. 39.23) 297, 298
Пурн (Puri M. L.), устойчивость оценок в ДА 164; многомерный ранговый критерий 396
Пустые клетки (empty cells), при двухфакторной перекрестной классификации 51—52

Равновзвешивающий выбор self-weighting sampling) 280—281, 291—292, 331
Радж (Raj D.), достаточность при обследованиях 246, (упражнение 39.1) 294; неравные вероятности 253—254, (упражнения 39.9—39.10) 296; двухэтапный выбор с подбором вероятностей 325
Раджалакшман (Rajalakshman D. V.), многомерные временные ряды 676
Разложение нецентральных квадратичных форм (decomposition of non-central quadratic forms) 13—17
Размах (range), критерий выборочного размаха в ДА 71—74
Ранги (ranks), преобразования 142; их использование в ДА 156—157, 159—161, (упражнения 37.13—37.14) 171—172, 203; их использование в ДА для СНБ 220;

критерин во временных рядах, основанные на ранговой корреляции 494—498; критерий ранговой сериальной корреляции (упражнение 45.5) 502
Рандомизация полная (complete randomization), 122, 124; в экспериментах 177—134
Рао Дж. (Rao J. N. K.), метод МП в смешанной модели ДА 121; выбор с возвращением и без 246; неравные вероятности 251; уменьшение смещения для оценок-отношений 308; случайное формирование слоев, (упражнение 40.6) 335
Рао П. (Rao P. S. R. S.), многомерные оценки при обследованиях 308
Рао С. (Rao C. R.), планы СНБ 211; ДА для планов СНБ 213; анализ ЧСНБ 223; дискриминация 448, 451, 461, (пример 44.4) 452—453
Расслоенный выбор (stratified sampling) 256—269, (упражнения 39.13—39.15, 39.17—39.21) 297—298; обоснование 256—258; выбор объема слоев 259—262; слои и блоки 262; размещение с МД при фиксированной стоимости 263—264; формирование слоев 264—268, (упражнения 40.4—40.6) 335; оценка его влияния 268—269; р. в. и группировка 271; при многошаговом выборе 294; оценки-отношения и регрессионные 318; с двухэтапной выборкой 319—322; области изучения 330—333, (упражнения 40.15—40.16) 339; выборка по группам (упражнение 40.11) 337—338
Расщепленные планы (split-plot designs) 229
Реализации (realization) случайного процесса 556
Регрессионные оценки (regression estimators) 311—313; несмещенность 313—316; р. о. при расслоении и многошаговом выборе 318—319; асимптотическая линейность 318; при двухэтапной выборке 319—325
Регрессия (regression), проверка степени полинома (упражнение 37.3) 169; преобразования (упражнение 37.9) 170; планирование р. 230—235; в многомерном анализе 385—390, (пример 42.3) 390—393, (упражнения 42.15—42.16) 398—399; с автокоррелированными ошибками 677—679; см. Авторегрессионные ряды
Рекордных значений критерий (records test) 497, (упражнения 45.8—45.9) 603—504
Решетчатые планы (lattice designs) 223—225
Рис (Rees D. H.), неортогональная аддитивная трехфакторная перекрестная классификация 62; распределение собственных чисел 367, (пример 42.4) 394—395
Робсон (Robson D. S.), оценки-отношения 305, 318; оценка-произведение (упражнение 40.2) 334
Розенблатт (Rosenblatt M.), спектральный анализ 638; регрессия с автокоррелированными ошибками 679
Рой Дж. (Roy J.), информация между блоками 220
Рой С. (Roy S. N.), смешанные модели ДА 121; распределение собственных чисел 365, 367; D^2-статистика Махаланобиса 368
Романовский В., синусоидальный предел Слуцкого 571
Росс (Ross A.), размещение при расслоенном выборе (упражнение 39.20) 298; несмещенные оценки-отношения 303
Ротатабельные планы (rotatable designs) 230
Роудс (Rhodes E. C.), выравнивание тренда 542
Рубин Дж. (Rubin J.), методы формирования групп 469
Рубин Х. (Rybin H.), сериальные корреляции 605

Саранджи (Sarangi J.), выровненные ранговые критерии 161
Сбалансированная классификация (balanced classification) (с равным числом наблюдений в клетках) 40—41; важность в модели II, а не в модели I 94; устойчивость 158; см. Классификация
Сбалансированные неполные блоки (СНБ) (balanced incomplete blocks) 207—213, 217—223, (упражнения 38.10—38.13, 38.16—38.18) 236—238; неприводимые 209; симметричные, дуальные, разложимые, афиноразложимые 209; дополнительный план 210; ДА 210—213; с информацией между блоками 220; перестановочные распределения 220; эксперименты с предпочтением 221; парные сравнения 221—222; смешивание 226—227
Свини (Sweeny H. C.), планирование полиномиальной регрессии 234
Свободные от распределения методы (distribution-free methods), в ДА 156—164; при многомерной проверке параметров расположения 396; в дискриминации 449, 461, (упражнения 44.10—44.11) 473
Связанные взаимодействия (tied interactions) 115—116
Себер (Seber G. A. F.), ортогональность в ДА 60; мощность многомерных критериев 396
Сезонные изменения (seasonal variation) 482—484, 546—551, 556; с. и. и спектр 638; см. Скользящие средние, Тренд
Сен А. (Sen A. R.), схемы выбора с неравными вероятностями (упражнения 39.5—39.6) 295
Сен П. (Sen P. K.), устойчивость оценок в ДА 164; эффективность критериев перестановок для случайных блоков 220; многомерный ранговый критерий 396
Сериальные корреляции (serial correlation), использующие ранги 498, (упражнение 45.5) 502; определение 499—501, (упражнения 45.10—45.11) 504; с. к. и дисперсии разностей 539—540; с. к. и метод переменных разностей 542, (упражнения 46.10—46.11) 552—553; теория для больших выборок 590—593, (упражнения 48.3, 48.9—48.10) 615—616; смещение 593—596; точные выражения для моментов 596—598, (упражнения 48.5—48.6, 48.12, 48.18—48.19) 615—619; распределение в нормальном случае 598—615, (упражнение 48.20) 619; преобразования 609, (упражнение 48.13) 617; см. Автокорреляции
Сёрл (Searle S. K.), оценки в несбалансированной модели II ДА 113
Силви (Silvey S. D.), мощность критериев 396
Силлитто Sillitto G. P.), таблицы СНБ 211
Симаик (Simaika J. B.), мощность критерия T^2 395
Сингх (Singh M. P.), двухэтапная выборка 325
Сиотани (Siotani M.), доверительные интервалы для сравнений в ДА 78
Систематический выбор (systematic sampling) 270

Ситгривз (Sitgreaves R.), дискриминация (упражнение 44.2) 471—472
Скользящие средние (moving average) 506—554; как полиномы метода НК 505—506; формулы коэффициентов до пятой степени 508—509; формулы в терминах разностей 511—512; формулы Спенсера для 15 и 21 точек (примеры 46.3—46.4) 514—515; с. с. и краевые эффекты 515—517; использование ортогональных полиномов 517—518; см. Сезонные изменения, Скользящих средних ряды, Тренд
Скользящих средних ряды (moving average series), 567—573, с. с. р. и ряды авторегрессии 573, 647—649; оценки и критерии согласия 655—660, (упражнение 50.6) 684; с. с. р. как ошибки в авторегрессионных рядах 660—662; экспоненциальные веса 683; см. Авторегрессионные ряды, Временные ряды
Скотт (Scott E. L.), смещение, вызванное преобразованием 142
Слейтер (Slater P.), дискриминация данных по неврастении (пример 44.4) 452—453
Слуцкого—Юла эффект (Slutzky—Vule effect) 523; теорема о синусоидальном пределе (пример 47.6) 570—571
Случайности критерии (tests of randomness) 485—499; см. Знаков разностей, Рекордных значений, Экстремальных точек критерии, Ранги, Сериальные корреляции, Фазы
Случайные блоки (randomized blocks) 122, 191—193, (упражнения 38.3—38.4) 235; устойчивость к нормальной теории 201—204; при факторных экспериментах 226
Смешанные модели (mixed models) 118—121; с использованием информации между блоками 213—220
Смешивание (confounding) 226—229; в факторных экспериментах 228
Смещение (bias), вызванное преобразованием 142—143
Смит К. (Smith K.), планирование полиномиальной регрессии 234, (упражнение 38.19) 239
Смит С. (Smith C. A. B.), квадратичная дискриминация 448
СНБ, см. Сбалансированные неполные блоки
Снедекор (Snedecor G. W.), метод Уэйтса взвешенных квадратов от средних значений 50
Собственные числа (latent roots) матрицы рассеяния, «нулевое» распределение 361—366, 367, 368, (упражнение 41.14) 371, (пример 42.4) 394—395; эквивалентность проверки независимости и равенства с. ч. 409, (пример 43.3) 409; проверка равенства нулю 409; проверка равенства малых с. ч. 410; результаты для больших выборок 411—412; в дискриминации 453—454; см. Главных компонент анализ, Канонические корреляции, Факторный анализ
Соломон (Solomon H.), классификация по компонентам 468
Сопутствующие переменные (concomitant variables) 80, 301 (сноска)
Спектр (spectrum), спектральная плотность, спектральная функция 565; как производящая функция автокорреляций 567; для марковских рядов и рядов Юла (примеры 47.7—47.8) 575—577; для непрерывных рядов 580; влияние фильтра на спектр 583; спектральная теория 620—643; гармонический анализ 620—622; частота Найквиста и неразличимые частоты 622—625; влияние гармонической компоненты 625—628; влияние других периодичностей и тренда 628—629; критерии для ординаты спектра 629—630; сглаживание 632—633; оценивание плотностей 636—638; спектр и сезонные изменения 638; неравные временные интервалы 638—640; взаимные спектральные функции 669—676; когерентность 670, полиспектр 670; см. Временные ряды
Спенсера формула для 15 и 21 точек (примеры 46.3—46.4) 514—515
Спитволл (Spjøtvoll E.), несбалансированная модель II ДА 113, (упражнение 36.16) 128
Спротт (Sprott D. A.), планы СНБ 211
Сравнения (contrasts) 75; одновременные доверительные интервалы для них 74—78, (упражнения 35.11—35.14, 35.16—35.19) 85—87
Средние квадраты (СрК) (mean squares), их математические ожидания в модели II ДА 98, 106—107
Стабилизация дисперсии (stabilization of variance) 134—139
Стаддеи (Studden W. J.), оптимальные эксперименты 190
Стационарные временные ряды (stationary time-series) 556; см. Временные ряды
Стейн (Stein C.), использование информации между блоками 220
Стивенс (Stevens W. L.), неортогональная трехфакторная перекрестная классификация 62
Стоимость (cost), функция с. в выборочной теории 263—264; при расслоенном выборе 263—264; при многошаговом выборе 290; при подборе вероятностей 291—293; при двухэтапной выборке 322—324; уклонение от ответа (упражнение 40.7) 235—236
Ступенчатые процедуры (step-by-step procedures), в ДА 68—74, (упражнение 35.18) 87
Стьюарт (Stuart A.), случайное формирование слоев (упражнения 40.4—40.6) 336; критерий знаков разностей 494, 498; критерий рекордных значений 497, (упражнения 45.8—45.9) 503—504; критерий экстремальных точек 498, (упражнение 45.4) 502; критерии ранговой корреляции 498; критерии ранговой сериальной корреляции 498 (упражнение 45.5) 502
Стьюдент («Student», Gosset W. S.), критерий МЗР в ДА 70
Стьюдентизированный размах (studentized rang), критерий в ДА 71—74
Сугияма (Sugiyama T.), распределение собственных чисел 367
Сферичности критерий (sphericity test) 382—383, (пример 43.3) 409
Сэмпфорд (Sampford M. R.), выбор с неравными вероятностями (упражнение 39.2) 294; обратный выбор (упражнение 40.3) 334
Сэти (Setni V. K.), формирование слоев 268; несмещенные оценки-отношения 318
Сэттерсвэйт (Satterthwaite F. E.), приближенный F-критерий в ДА (упражнение 36.7) 126—127; случайные сбалансированные эксперименты 190—191
Сэшадри (Seshadri V.), информация между блоками 220
Сю (Hsu P. L.), собственные числа 365; многомерное бетта-распределение 368

Такэючи (Takeuchi K.), планы СНБ 211
Тамура (Tamura R.), многомерные свободные от распределения критерии о параметрах расположения 396
Тейлор (Taylor L. R.), таблицы преобразований 131
Тидвелл (Tidwell P. W.), преобразования 131, (упражнение 37.9) 170
Тин (Tin M.), оценки-отношения 310—311, (упражнения 40.13—40.14) 338
Тинтнер (Tintner G.), метод переменных разностей 538, 539
Томпсон В. (Thompson W. A. Jr.), методы множественных сравнений ранжированных сумм 74; неотрицательные оценки дисперсий в модели II ДА 110
Томпсон Д. (Thompson D. J.), выбор с неравными вероятностями 250
Точер (Tocher K. D.), потерянные наблюдения 166; другие «испорченные» эксперименты 168; блочные эксперименты 182, 206, (упражнения 38.2—38.3, 38.8—38.9) 235—236; информация между блоками 219, (упражнение 38.15) 238
Трайон (Tryon R. C.), классификация по компонентам 468
Тренд (trend) 482—484, 505; критерии отсутствия т. 490—491, 498; влияние выделения т. методом скользящих средних на другие компоненты (эффект Слуцкого — Юла) 519—531, 542—546, (упражнение 46.12) 554; см. Скользящие средние
Тьюки (Tukey J. W.), критерий проверки аддитивности 43; множественные сравнения 70—78; критерий стьюдентизированного размаха 71—74; одновременные доверительные интервалы для разностей и сравнений 74—78; (упражнение 35.13) 86; математическое ожидание средних квадратов в ДА 107; оценки в несбалансированной модели II ДА 113; модели в ДА 115, 118; моменты для оценок дисперсий в ДА (упражнение 36.10) 127; преобразования 131, 136, (упражнение 37.4) 169; анализ остатков 143; случайные сбалансированные эксперименты 191; спектральный анализ 636, 637, 638 (упражнение 49.5) 640
Тюрин Ю. Н., проверка нормальности многомерной выборки 115 (сноска)

Уайз (Wise J.), ряды авторегрессии 574, (упражнение 47.17) 588
Уайт (White J. S.), смещение в оценках сериальных корреляций 596; моменты сериальных корреляций в марковском случае 609
Уикс (Weeks D. L.), информация между блоками 220
Уилк (Wilk M. B.), модели в ДА 115; полная рандомизация 123—124; математическое ожидание СрК для латинских квадратов 202
Уилкинсон (Wilkinson G. N.), потерянные наблюдения 167—168
Уилкс (Wilks S. S.), критерий ОП независимости классов переменных (упражнения 41.10—41.11) 370, 382; критерий однородности 376, (пример 42.2) 383, (упражнение 42.7) 397
Уилльямс Е. (Williams E. J.), канонический анализ 427; дискриминация 454
Уилльямс У. (Williams W. H.), несмещенные регрессионные оценки 313, 318
Уилсон Е. (Wilson E. B.), преобразование 141
Уилсон К. (Wilson K. B.), эволюционная операция 230
Уинтерс (Winters P. R.), экспоненциальные веса 683
Уиттекер (Whittaker E. T.), периодограмма (упражнения 49.9—49.10) 641—642
Уиттл (Whittle P.), ряды авторегрессии (упражнение 47.16) 588; матрица автокорреляций (упражнение 47.18) 588; оценивание спектральных плотностей 637; оценивание для рядов скользящих средних (пример 50.1) 648—649; оценки и критерии для временных рядов 662
Уишарт (Wishart J.), распределение выборочных ковариаций для многомерного нормального случая 343—349, (упражнения 41.6—41.7) 369—370; нецентральное 367; корреляция между выборочными ковариациями (упражнение 41.4) 369; распределение ковариации (упражнение 41.13) 370—371; вид распределения матрицы рассеяния остатков 388
Уокер А. (Walker A. M.), схемы авторегрессии и скользящих средних 660
Уокер Г. (Walker G.), уравнения для рядов авторегрессии 574; критерий в гармоническом анализе 630
Уоллис (Wallis W. A.), критерии случайности 490, 492
Уорд (Ward D. H.), экспоненциальные веса 683
Уоркинг (Working H.), группы в марковских рядах (упражнение 47.15) 588
Устойчивость (robustness) процедур ДА 146—156, 163—164
Уэбстер (Webster J. T.), уменьшение смещения для оценок-отношений 308
Уэйгл (Wagle B.), распределение собственных чисел 367
Уэлч (Welch B. L.), устойчивость процедур ДА 154, 204
Уэйг (Wang Y. Y), оценки в модели II ДА 100

Фазы (phases), во временных рядах 486, 488—491, (упражнение 45.1) 501—502; в гармоническом анализе 621
Факторные эксперименты 225—226; смешивание 228; частичное дублирование 229
Факторный анализ (factor analysis) 427—433, (упражнения 43.11—43.13) 436; неопределенность системы уравнений 428; решение методом МП 429—431; критерий для числа факторов 431—432; обсуждение 432—433; см. Главных компонент анализ
Федоров В. В., оптимальные эксперименты 230 (сноска)
Фелледжи (Fellegi I.), выбор с неравными вероятностями 247, 251
Ферхаген (Verhagen A. M. W.), доказательство метода всех сравнений Шеффе (упражнение 35.19) 87
Филлипс (Phillips A. W.), многомерные временные ряды 666
Фильтры (filters) 582—583; см. Временные ряды
Финни (Finney D. J.), объединение критериев ДА 68; пробит и логит преобразования 142
Финч (Finch P.), схемы авторегрессии и скользящих средних 660

Фиск (Fisk P. R.), уравнения со стохастической зависимостью 681
Фишер (Fisher R. A. Sir Ronald), основатель ДА 13; критерий МЗР в ДА 71; преобразование r 139; критерий двойной симметрии 158; пропаганда рандомизации 177; статистические выводы из экспериментов 181; неравенство СНБ 208; смешивание 229; собственные числа 365; дискриминация (пример 44.1) 441—444, 471; критерий в гармоническом анализе 630
Фортье (Fortier J. J.), классификация по компонентам 468
Фостер (Foster F. G.), распределение собственных чисел 367, (пример 42.4) 394—395; критерий рекордных значений 497, (упражнения 45.8—45.9) 503—504
Фридмэн М. (Friedman M.), использование рангов в ДА (упражнение 37.14) 172
Фридмэн Х. (Friedman H. P.), методы формирования групп 469
Фримэн Г. (Freeman G. H.) неортогональная трехфакторная перекрестная классификация 62
Фримэн М. (Freeman M. F.), преобразования 136, (упражнение 37.4) 169

Хаавелмо (Haavelmo T.), системы уравнений 680
Халтквист (Hultquist R. A.), модель II ДА 90, 95—98
Хамис (Khamis S. H.), теория выборочных обследований 246, (упражнение 39.1) 294
Хан (Khan S.), расслоенный выбор 264
Ханани (Hanani H.), неравенство СНБ 208
Хансен (Hansen M. H.), выборочные обследования 241; выбор с пропорциональными объему вероятностями 281; подбор вероятностей извлечения 291; уклонение от ответа при обследованиях (упражнение 40.7) 335—336
Хантер (Hunter J. S.), случайные сбалансированные эксперименты 191; поверхность отклика 230
Ханурав (Hanurav T. V.), выбор с неравными вероятностями 251
Харвилл (Harville D. A.), оценки в несбалансированной модели II ДА 113, 124
Хармэн (Harman H. H.), факторный анализ 432
Хартер (Harter H. L.), методы множественных сравнений 74
Хартли (Hartley H. O), объединение критериев ДА 65—69; критерий Ньюмэна — Кейлса 74, (упражнение 35.10) 85; объединение в модели II ДА (пример 36.6) 105; метод МП в смешанной модели ДА 121; несмещенные оценки-отношения 303, 305, 307, 309; области изучения 327; случайное формирование слоев (упражнение 40.6) 335
Хатанака (Hatanaka M.), спектральный анализ 638, 670
Хейдер (Hader R. J.), ротатабельные планы 230
Хейланд (Hfyland A.), устойчивость оценок в ДА 163
Хекст (Hext G.), спектральный анализ 638
Хендерсон (Henderson C. R.), оценки в несбалансированной модели II ДА 113
Хеннан (Hannan E. J.), многомерные временные ряды 615 (сноска); спектр и сезонные изменения 638; регрессия с автокоррелированными ошибками 677
Хигхем (Higham J. A.), выравнивание тренда (упражнение 46.6) 552
Хикс Ч. (Hiks C. R.), принципы планирования эксперимента 230 (сноска)
Хили (Healy M. J. R.), таблицы преобразований 131; данные о представителях мужского пола (пример 42.4) 394—395
Хиллс (Hills M.), дискриминация 455, 458
Хилферти (Hilferty M. M.), преобразование 141
Химсворс (Himsworth F. R.), смешивание 229; эволюционная операция 230
Хинчин А. Я., эргодическая теорема 559, 562
Хогг (Hogg R. V.), вложенные гипотезы (упражнение 37.2) 169; проверка степени полиномиальной регрессии (упражнение 37.3) 169; мощность критериев 396
Ходжес (Hodges J. L.), выровненные ранговые критерии 161; устойчивость оценок в ДА 163—164; формирование слоев 267
Холландер (Hollander M.), проверка гипотез против альтернатив об упорядочениях 203
Холловэй (Holloway L. N.), устойчивость критерия T^2 396
Холт (Holt C. C.), экспоненциальные веса 682
Хорвитц (Horvitz D. G.), выбор с неравными вероятностями 250
Хорснелл (Horsnell G.), устойчивость процедур ДА 147
Хотеллинг (Hotelling H.), T^2 — распределение критерия ДА 121; стабилизация дисперсии 130; T^2 — распределение 355, 359, (упражнение 41.12) 370; геометрическая интерпретация 356—358; T^2 и R^2 367—358; T^2 и F 358; критерий T^2 для одной выборки 358; критерий T^2 для двух выборок 359; T^2 и D^2 для двух выборок 368; нецентральное T^2 367, 396; критерий T_0^2 для нескольких выборок 396; устойчивость и мощность критерия T^2 396, (упражнения 42.11, 42.13—42.14) 397—398; T^2 как критерий ОП (упражнение 42.10) 397; T^2 в обобщении критерия Шеффе (упражнение 42.12) 398; канонические корреляции 419, (пример 43.5) 423—424, 425, (упражнения 43.5—43.7) 434—435
Хоув (Howe W. G.), факторный анализ 431
Хоул (Hoel P. G.), планирование регрессии 233—235; распределение корреляционного определителя (упражнение 41.8) 370

Циклические (circular) сериальные корреляции 501; процессы 585; см. Временные ряды

Чако (Chacko V. J.), альтернативы об упорядочениях в ДА 79
Частично сбалансированные неполные блоки (ЧСНБ) (partially balanced incomplete blocks), 222—223
Чеймберс (Chambers J. M.), асимптотика многомерных распределений 368
Чернова Н. А., планирование экстремальных экспериментов 230 (сноска)
Чу (Chew V.), обзор по многомерным доверительным и прочим областям 373
Чупров (Tschuprow A. A.), расслоенный выбор 259

Шарма (Sharma D.), критерий Тьюки проверки аддитивности 43
Шатцов (Schatzoff M.), распределение критерия ОП 377, 393; сравнение критериев для вектора средних 396
Шах (Shah K. R.), информация между блоками 220
Шеффе Scheffe H.), критерий Тьюки проверки аддитивности 43; независимость равенства взаимодействий нулю от системы весов 45; двухфакторный перекрестный ДА с непропорциональными частотами 50; перекрестная классификация с пустыми клетками 52; трехфакторная иерархическая классификация 57; смешанная классификация 58; геометрическая интерпретация многофакторной перекрестной классификации 64; методы множественных сравнений 74; одновременные доверительные интервалы для всех сравнений 77—78, (упражнения 36.11—36.13) 85—86; ковариационный анализ 82; математические ожидания СрК в ДА 107; трехфакторная иерархическая классификация в модели II 109; доверительные интервалы в модели II ДА 110, (упражнение 36.15) 128; модели для ДА 115; смешанная модель 115, 118, 120—121; устойчивость процедур ДА 147; взаимодействия в латинских квадратах 197; устойчивость для случайных блоков и латинских квадратов 202—204; задача двух средних 396
Шискин (Shiskin J.), сезонные изменения 551
Шметтерер (Schmetterer L.), смещение, вызванное преобразованием 142
Шорак (Shorack G. R.), альтернативы об упорядочениях 79
Шривастава (Srivastava S. R.), процедуры объединения в модели II ДА (пример 36.6) 105
Шриканде (Shrikande S. S.), ошибочность предположения Эйлера о греко-латинских квадратах 199; ЧСНБ 223
Шулль (Schull W. J.) устойчивость критерия χ^2 396

Эволюционная операция (evolutionary operation) 229—230
Эйзенпресс (Eisenpress H.), сезонные изменения 551
Эйзенхарт (Eisenhart C.), модели ДА 88
Эйлер, латинские квадраты 196; неправильное предположение о греко-латинских квадратах 199
Эквивалентные выборки (equivalent samples) 245
Экман (Ekman G.), формирование слоев 267
Эксперименты (expirements), их планирование 176—235; сравнение с обследованиями 175—176, 262; принципы проведения, рандомизация 176—184; блочные эксперименты 182—191; матрица инциденций 183; линейная модель 184—190; ДА 188—190; планирование 190—191; два мешающих фактора 193—196; информация между блоками 214—220; с предпочтением 221; факторные 225—226; смешивание 226—229; последовательность э. 229—230; планирование регрессии 230—235; см. Сбалансированные неполные блоки, Латинские квадраты, Случайные блоки
Экстремальных точек критерий (turning-points test), 485—488, 489—490, (упражнение 45.4) 602
Эндрюс (Andrews F. C.), медианные критерии 162
Эйскомб (Anscombe F. J.), преобразования 136, (упражнения 37.4—37.5) 169—170; анализ остатков 143—145, (упражнения 37.15—37.16) 172—173; случайные сбалансированные эксперименты 191
Эргодический (ergodic) процесс 559, 562, см. Временные ряды
Эштон (Ashton E. H.), данные о представителях мужского пола (пример 42.4) 394—395

Юдэн (Youden W. J.), случайные сбалансированные эксперименты 191; квадраты Юдэна 221
Юл (Yule G. U.), эффект Слуцкого—Юла 523; уравнения для рядов авторегрессии 574; см. Юла ряды
Юла ряды (Yule series), коррелограмма и спектр (пример 47.8) 576, 577; предельный случай (пример 47.10) 578; непрерывный аналог (пример 47.11) 581; частные автокорреляции 584; дисперсия (упражнение 47.8) 587; автокорреляции остатков (упражнение 47.10) 587; стандартные ошибки сериальных корреляций (упражнения 48.2, 48.14) 615, 617; сериальные корреляции с ошибками наблюдений (упражнение 49.11) 642—643; критерий согласия (пример 50.3) 654—655; многомерные 672
Ютц (Youtz C.), таблицы преобразований 136, (упражнение 37.4) 169

Янг (Young D. H.), выборка по группам (упражнение 40.11) 337—338

С—Е—I—R — сборник статей по построению моделей 681