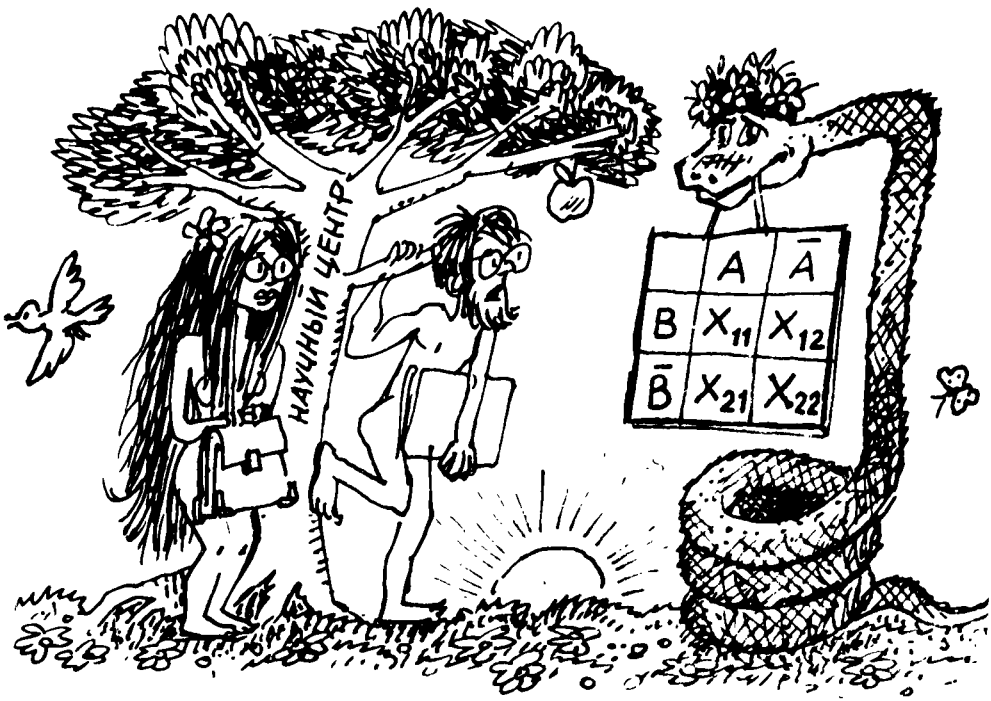
Дж. Флейс

Статистические методы для изучения таблиц долей и пропорций



Библиотечка иностранных книг для экономистов и статистиков

Statistical Methods for Rates and Proportions

Second Edition

JOSEPH L. FLEISS

Division of Biostatistics, School
of Public Health, Columbia University

JOHN WILEY & SONS
New York · Chichester
Brisbane · Toronto · Singapore

Дж. Флейс

Статистические методы для изучения таблиц долей и пропорций

Перевод с английского
И.Л. ЛЕГОСТАЕВОЙ
А.М. НИКИФОРОВА

Под редакцией и с предисловием
д-ра физ.-мат. наук
Ю.Н. БЛАГОВЕЩЕНСКОГО

Москва
"ФИНАНСЫ И СТАТИСТИКА"
1989

ББК 22.172
Ф71

БИБЛИОТЕЧКА ИНОСТРАННЫХ КНИГ
ДЛЯ ЭКОНОМИСТОВ И СТАТИСТИКОВ

Серия основана в 1968 году

Ф $\frac{0702000000-039}{010\ (01)-89}$ 112—89

ISBN 0-471-06428 (США)

ISBN 5-279-00249 (СССР)

Предисловие к русскому изданию

Книга Дж. Флейса представляет, на мой взгляд, весьма необычное явленне. О ней можно сказать коротко: она является одновременно и учебником, и рецептурным сборником, и монографией, и методологическим исследованием... Однако такая оценка книги была бы верной лишь с формальной точки зрения, в то время как она написана совсем не формально, а названные выше ее характеристики проявляют себя далеко не сразу.

Обосновать учебный характер книги незатруднительно, поскольку это наиболее очевидное ее свойство. В каждом разделе имеется несколько задач и в конце книги даются либо решения этих задач, либо пояснения к ним. С первых и до последних страниц книги весь материал иллюстрируется примерами, которые досчитываются до конца и интерпретируются на содержательном уровне.

Одновременно с примерами приводятся рецептурные формулы, причем рассматривается весьма богатое множество ситуаций и для каждой из них дается метод исследования вплоть до проверки гипотезы о выборе самой ситуации и предположений о структуре формирования данных или, точнее, о вероятностном механизме, ее порождающем.

И при всем том книга Флейса является монографией, тему которой можно представить со следующей, быть может, несколько непривычной точки зрения. Пусть мы изучаем серию объектов, извлеченных, по некоторому правилу G из общей совокупности. Пусть каждый объект описывается p свойствами $X^{(1)}, X^{(2)}, \ldots, X^{(p)}$, так что фиксация этих свойств для N извлекаемых и наблюдаемых в эксперименте объектов приводит к таблице данных:

$$X(N) = \begin{pmatrix} x_1^{(1)} x_2^{(1)} \ldots x_N^{(1)} \\ x_1^{(2)} x_2^{(2)} \ldots x_N^{(2)} \\ \ldots\ldots\ldots\ldots \\ x_1^{(p)} x_2^{(p)} \ldots x_N^{(p)} \end{pmatrix},$$

где $x_j^{(i)}$ — реализованное значение i-го свойства $X^{(i)}$, $i=1, 2, \ldots, p$, у j-го извлеченного объекта.

Одна из очень частых ситуаций на практике — наблюдение объектов со свойствами, о которых можно сказать лишь, что они либо есть, либо нет. Если таких свойств всего два ($p=2$) и если наличие свойства заносится в таблицу как 1, а его отсутствие — как 0, то вся таблица $X(N)$ будет представлять последовательность столбцов:

$$\binom{1}{1}, \binom{1}{0}, \binom{0}{1} \text{ и } \binom{0}{0}.$$

Таким образом, если номер наблюдения неинформативен, то вся таблица может быть заменена, по сути, таблицей из четырех чисел:

$$\hat{X}(N) = \begin{pmatrix} n_{00} & n_{01} \\ n_{10} & n_{11} \end{pmatrix},$$

где n_{00} — число столбцов вида $\binom{0}{0}$, а n_{01}, n_{10} и n_{11} — число столбцов вида $\binom{0}{1}$, $\binom{1}{0}$ и $\binom{1}{1}$, соответственно $n_{00}+n_{01}+n_{10}+n_{11}=N$.

Так вот, основная тема книги — изучение связей между свойствами по данным, представленным такими четырехклеточными таблицами с учетом того, как правило G извлечения объектов из общей совокупности влияет на выбор методов анализа и возможности получения тех или иных выводов.

Естественно, что во многих практических ситуациях характеризация объектов свойствами «наличия» или «отсутствия» некоторого фактора оказывается недостаточной. В книге рассмотрены отдельные случаи более сложной структуры описания объектов и даны определенные рекомендации по их анализу, но я бы рискнул утверждать, что главное достоинство книги не в этих обобщениях, а в том действительно широком многообразии вариантов взаимодействия между (а) правилами G извлечения объектов, (б) рабочими гипотезами о связях между свойствами в одной или нескольких таблицах и (в) методами анализа по данным таких таблиц. Насколько я могу судить, аналогичной книги среди доступных массовому «потребителю» в настоящее время нет.

А теперь — о методологических особенностях, присущих книге Флейса, отражающих его педагогические принципы. Автор использует одно и то же обозначение для вероятностей, для долей и для пропорций, сплошь и рядом одно и то же называя разными понятиями. Эта вариантность, воспринимаемая вначале как некая неряшливость, является педагогическим приемом, предназначенным для воспитания внимательности у читателя, без которой целый ряд важных методологических указаний не был бы акцентирован.

Целый спектр интересных проблем, чуть намеченных идей, эвристических приемов и отдельных мыслей, разбросанных по всей книге, дают очень неплохое представление о современном состоянии этого раздела математической статистики. Ненавязчивость материала, с одной стороны, позволяет справиться с текстом читателю, имеющему самые элементарные познания в теории вероятностей, а с другой стороны, предоставляет широкое поле для размышлений статистику-профессионалу.

Многослойность и насыщенность интерпретационно-методологическими рассуждениями является несомненным достоинством книги Флейса, однако является и основным источником трудностей перевода; особенно это относится к многозначной терминологии автора. Вполне возможно, что отдельные трудные места в книге, на которые наткнется читатель, привнесены переводом, поскольку далеко не во всех случаях удалось подыскать удачные аналоги в русском языке как для отдельных понятий, так и для некоторых интерпретационных высказываний.

В заключение заметим, что книга Флейса предназначена англоязычному читателю и предполагает его знакомство с определенными курсами по статистике, несколько отличными от курсов, читаемых в наших вузах. В целом же монография Флейса очень своевременна и в ближайшие годы, несомненно, будет незаменимым пособием для практиков при работе с четырехклеточными таблицами.

Главы 1—3 переведены И. Л. Легостаевой, предисловие и гл. с 4 по 14 — А. М. Никифоровым.

Ю. Н. Благовещенский

Предисловие

Через несколько лет после выхода в свет первого издания (1970 г.) необходимость подготовки пересмотренного варианта книги стала очевидной.

Рецензенты, исследователи, преподаватели и студенты указывали на некоторые важные темы, которые либо не затрагивались в книге, либо рассматривались чересчур кратко, либо формулировались в виде, не соответствующем современным требованиям. Тем временем продолжалось развитие прикладной статистики и появились новые результаты, заслуживающие обсуждения или хотя бы упоминания.

Среди вопросов, которых я не касался в первом издании, наиболее важным является вопрос о построении доверительных интервалов. В настоящей работе интервальному оцениванию уделено почти столько же места, сколько проверке гипотез. Тесная связь между этими задачами подчеркнута в новом разд. 1.4. Читатель найдет в нем, а также в разд. 5.6 (и ряде других мест книги) подтверждение замечанию о том, что правильно построить доверительный интервал часто сложнее, чем просто взять точечную оценку плюс-минус стандартную ошибку с некоторым множителем.

Другой важной темой, не раскрытой в первом издании, является планирование сравнительных исследований с неравными объемами выборок. Этой проблеме посвящен разд. 3.4. Кроме того, дополнительно в работу включено описание «точного» критерия Фишера — Ирвина для четырехклеточной таблицы, см. разд. 2.2. В разд. 5.7 и 6.4 изучается привносимый риск — важный показатель действия фактора риска. Метод построения выводов об отношении шансов по Корнфилду представлен в разд. 5.5 и 5.6.

Я надеюсь, что ряд вопросов, рассмотренных в первом издании поверхностно или недостаточно точно, обсужден теперь должным образом. Результаты анализа данных в исследовании по плану с перекрытием содержатся в разд. 7.2. Для исследования связанных пар в случае, когда результирующий фактор — порядковый, дан более подходящий метод, см. разд. 8.2. Разд. 8.4 пополнился описанием метода сравнения пропорций в связанных выборках, представляющих различные

градации количественного признака. Описание оказавшегося некорректным метода сравнения для данных из нескольких четырехклеточных таблиц перенесено в раздел, посвященный методам, которые не следует применять (теперь это разд. 10.7).

Новые статистические результаты, полученные после публикации первого издания, отражены во всех главах и в большинстве разделов. В разд. 3.2 по-новому обсуждается, как определять размер выборки, соответствующим образом пересмотрена табл. А.3 в приложении. Некоторые недавно предложенные альтернативы простой рандомизации в клинических исследованиях освещаются в разд. 4.3 и 7.3. В разд. 9.4 в свете последних исследований представлен ридит-анализ. Влияние и контроль ошибочной классификации по обоим факторам в четырехклеточной таблице рассмотрены в разд. 11.3 и 12.2. Новая глава 13, которая является расширением и развитием последнего раздела первого издания, содержит результат по измерению степени согласованности экспертов, участвующих в формировании категоризованных данных. В разд. 14.3 и 14.5 указаны некоторые новые достижения в области непрямой стандартизации.

Большое внимание уделено приложениям в медицине и здравоохранении, примеры берутся из этой области. Такой выбор иллюстративного материала обусловлен тем, что я знаком с этой областью лучше всего, хотя, конечно, ее можно считать и не самой важной.

Второе, как и первое, издание книги предназначено для научных работников и студентов, прослушавших по меньшей мере годовой курс прикладной статистики, включая критерий хи-квадрат и корреляционный анализ. Большая часть задач, помещенных в конце глав, были пересмотрены. Добавлены новые задачи.

Многие мои коллеги и некоторые из рецензентов советовали мне включить в книгу решения по крайней мере для некоторых расчетных задач. Я решил сопроводить решениями все такие задачи. Преподавателям, которые будут давать эти задачи для самостоятельного решения студентам, нужно будет лишь изменить некоторые числовые значения.

По-прежнему для применения описанных методов достаточно знать алгебру в объеме средней школы и уметь логарифмировать и извлекать квадратные корни. Все вычисления можно проводить с помощью микрокалькулятора[1]. Мощные,

[1] Большинство описанных в книге методов (или их точных аналогов) реализовано в пакете прикладных программ «Анализ данных», разработанном в лаборатории математического моделирования и вычислительной техники НИИ гигиены труда и профессиональных заболеваний АМН СССР.— *Прим. пер.*

но математически сложные методы логлинейного и логистического регрессионного анализа частотных таблиц высокого порядка в работе не рассматриваются. В книгах Cox D. R. The analysis of binary data, Methuen, London, 1970 и Bishop Y. M. M., Fienberg S. E. and Holland P. W. Discrete multivariate analysis: Theory and practice, M. I. T. Press, Cambridge, Mass., 1975 эти методы прекрасно были описаны на довольно высоком математическом уровне. В двух небольших монографиях, вышедших позднее (Everitt B. S. The analysis of contingency tables, Halsted Press, New York, 1977, Fienberg S. E. The analysis of cross-classified categorical data, M. I. T. Press, Cambridge, Mass., 1977), эти проблемы обсуждаются на менее высоком уровне.

Я весьма признателен профессорам Колумбийского государственного университета Агнес Бергер, Джону Фертигу, Брюсу Левину, Патрику Шрауту и профессору Нью-Йоркского государственного университета Гари Саймону, высказавшим много полезных замечаний по материалу книги, Беатрис Шьюб, редактору книги — за помощь и советы, Бланш Агдерн — за подготовку машинописного варианта работы, издателям — American Journal of Epidemiology, Biometrics, Journal of Chronic Diseases, предоставившим мне возможность использовать данные, опубликованные в этих изданиях, а также моей жене Изабелле, ободрявшей меня, когда дела шли плохо.

Самую большую помощь оказали мне студенты, которым я читал курс по анализу категоризованных данных и лекции по современным статистическим методам. Они стали подопытными объектами в моих «сравнительных испытаниях» различных подходов к подаче нового (и старого) материала.

Нью-Йорк, декабрь, 1980 г. *Дж. Флейс*

Глава 1

Введение в прикладную теорию вероятностей

Для правильного понимания и использования различных типов частот, возникающих в процессе научного исследования, необходимо рассмотреть некоторые исходные элементы прикладной теории вероятностей. Наиболее простым аналогом частоты, имеющим очевидную интерпретацию, является вероятность — частотная мера априорной осуществимости некоего конкретного события или того, что произвольно взятый член популяции обнаружит некие характерные свойства. Важнейшее применение вероятностей состоит в оценивании ожидаемого количества индивидуумов в выборке размера n, которым присущи определенные свойства. Если P — вероятность того что индивидуум обладает некоторой особенностью, то ожидаемое число индивидуумов в этой выборке, обладающих той же особенностью, примерно равно nP.

В разд. 1.1 представлены система обозначений и необходимые для дальнейшего изложения определения. Теоретические основы, содержащиеся в разд. 1.1, использованы в разд. 1.2 для количественной оценки критерия отбора и в разд. 1.3 для получения возможных ошибок в выводах, полученных при изучении подвыборок. Разд. 1.4 посвящен методам проверки гипотез о вероятностях или пропорциях и построению для них доверительных интервалов.

1.1. Обозначения и определения

В этой книге для терминов «вероятность», «относительная частота», «пропорция» и «доля» применяются одни и те же обозначения. Если событие A состоит в том, что случайно выбранный из популяции индивидуум имеет определенный признак (например, поражен атеросклерозом сосудов сердца), то $P(A)$ обозначает долю людей в популяции, которые имеют этот признак. В данном примере $P(A)$ — вероятность у случайно выбранного индивидуума обнаружить атеросклероз со-

судов сердца, или в терминах медицинской статистики — частота заболевания атеросклерозом сосудов сердца.

Однако только с общими (безусловными) частотами далеко не продвинешься: чаще встречаются и с большим успехом используются специфические частоты, т. е. частоты конкретного признака, специфичного для представителей определенных возрастных групп, расы, пола, рода занятий и т. д. Понятие, которое в эпидемиологии и в демографической статистике носит название специфическая частота, в теории вероятностей известно как условная вероятность. Она обозначается $P(A|B)$ и равна вероятности того, что случайно выбранный индивидуум имеет признак A, когда известно, что он имеет признак B (или, говоря другими словами, при условии, что он имеет признак B).

В нашем примере, если B означает событие, состоящее в том, что пациент принадлежит к возрастной группе 65—74 года, то $P(A|B)$ — условная вероятность того, что пациент поражен атеросклерозом сосудов сердца при условии, что ему 65—74 года. (Или, что эквивалентно, $P(A|B)$ есть частота поражения атеросклерозом сосудов сердца людей в возрасте 65—74 года.)

Пусть $P(B)$ обозначает долю людей, которые обладают признаком B, и пусть $P(A \text{ и } B)$ — доля всех людей, обладающих одновременно признаками A и B. Тогда по определению

$$P(A|B) = \frac{P(A \text{ и } B)}{P(B)}, \tag{1.1}$$

если только $P(B) \neq 0$.

Аналогично, если $P(A) \neq 0$, то

$$P(B|A) = \frac{P(A \text{ и } B)}{P(A)}. \tag{1.2}$$

Под связью двух признаков (свойств, факторов) подразумевается следующее: возможность проявления одного из признаков, присущих индивидууму, например B, зависит от проявления другого. Под независимостью или отсутствием связи двух признаков подразумевается тот факт, что индивидуум имеет один из признаков независимо от того, обладает ли он другим признаком. Таким образом, если A и B независимы, то частота $P(A|B)$, с которой признак A в действительности характерен для индивидуумов с признаком B, равна общей частоте $P(A)$, с которой наблюдается A. Из (1.1) вытекает теперь, что

$$\frac{P(A \text{ и } B)}{P(B)} = P(A),$$

или

$$P\ (A\ \text{и}\ B) = P\ (A) \cdot P\ (B). \qquad (1.3)$$

Равенство (1.3) часто используют как определение независимости признаков. Справедливо эквивалентное утверждение:

$$P(A \mid B) = P\ (A).$$

Эвристическое объяснение формулы (1.1) следующее. Пусть N обозначает общее число людей в популяции; N_A — число людей, обладающих признаком A; N_B — число людей, обладающих признаком B; N_{AB} — обладающих обоими признаками. Тогда ясно, что

$$P\ (A) = \frac{N_A}{N},$$

$$P\ (B) = \frac{N_B}{N}$$

и

$$P\ (A\ \text{и}\ B) = \frac{N_{AB}}{N}.$$

Под $P(A \mid B)$ подразумевается доля людей, имеющих признак A, среди тех, кто имеет признак B, так что и числитель, и знаменатель должны содержать B. Таким образом,

$$P(A \mid B) = \frac{N_{AB}}{N_B}. \qquad (1.4)$$

Если теперь числитель и знаменатель (1.4) разделить на N, то получим, что

$$P\ (A \mid B) = \frac{N_{AB}/N}{N_B/N} = \frac{P\ (A\ \text{и}\ B)}{P\ (B)}.$$

Подобным образом преобразуется уравнение (1.2).

$$P(B \mid A) = \frac{N_{AB}}{N_A} = \frac{N_{AB}/N}{N_A/N} = \frac{P\ (A\ \text{и}\ B)}{P\ (A)}.$$

Из теоремы Байеса, используя уравнения (1.1) и (1.2), получаем

$$P\ (B \mid A) = \frac{P\ (A \mid B) \cdot P\ (B)}{P\ (A)}. \qquad (1.5)$$

Соотношение (1.5) следует из определения $P\ (B \mid A)$ (формула (1.2)) и справедливости соотношения $P\ (A\ \text{и}\ B) = P\ (A \mid B)\ P\ (B)$, получаемого умножением обеих сторон уравнения (1.1) на $P\ (B)$.

1.2. Оценка теста отбора

Часто встречаемое применение теоремы Байеса состоит в оценивании диагностического теста, предназначенного для использования в программах отбора. Пусть B обозначает событие, состоящее в том, что индивидуум, о котором идет речь, имеет некоторую болезнь; $\overline{B}$ — событие, что у него этой болезни нет. Далее, пусть A означает, что отклик индивидуума на тест положителен, $\overline{A}$ — отклик отрицателен [1]. Предположим, что тест был применен для выявления событий B, т. е. для отбора людей с заболеванием, и B — для отбора людей без заболевания.

Результаты применения этого теста могут быть представлены двумя условными вероятностями: $P(A|B)$ и $P(A|\overline{B})$. Здесь $P(A|B)$ — условная вероятность положительного отклика среди людей, подверженных заболеванию; чем больше $P(A|B)$, тем более чувствителен тест. $P(A|B)$ — условная вероятность положительного отклика среди лиц, не подверженных данному заболеванию, чем меньше $P(A|\overline{B})$ (или, что эквивалентно, чем больше $P(\overline{A}|\overline{B})$), тем лучше выявляет тест альтернативное к B событие. Эти определения чувствительности и альтернативности критериев предложены Ерушелми [Yerushalmy, 1947]. Исследователя, однако больше заботят не сами чувствительность и альтернативность критерия, а оценки ошибок частот, возможных при реализации той или иной программы выбора.

Если положительный результат теста используется для фиксации присутствия заболевания, то ложная положительная частота скажем P_{F+}, определяется долей людей, которые фактически не подвержены заболеванию, но чей отклик на тест положителен, т. е. определяется вероятностью $P(\overline{B}|A)$. Из теоремы Байеса следует, что

$$P_{F+} = P(\overline{B}|A) = \frac{P(A|\overline{B})\ P(\overline{B})}{P(A)} = \frac{P(A|\overline{B})[1-P(B)]}{P(A)}, \qquad (1.6)$$

так как $P(B) = 1-P(B)$.

Ложная отрицательная частота, скажем P_{F-}, есть доля больных людей среди давших отрицательный отклик на тест, или $P(B|\overline{A})$. Снова по теореме Байеса

$$P_{F-} = P(B|\overline{A}) = \frac{P(\overline{A}|B)\,P(B)}{P(\overline{A})} = \frac{[1-P(A|B)]\ P(B)}{1-P(A)}, \qquad (1.7)$$

так как $P(\overline{A}|B) = 1-P(A|B)$ и $P(\overline{A}) = 1-P(A)$.

[1] Имеется в виду, что при отборе по тесту к группе с заболеванием относят тех людей, у которых отклик на тест положителен.— *Примеч. ред.*

Чтобы оценить эти две ложные частоты, нам по-прежнему требуются вероятности $P(A)$ и $P(B)$. На самом деле, нам необходима только одна величина — вероятность $P(B)$. Это видно из следующего соотношения:

$$P(A) = \frac{N_A}{N} = \frac{N_{AB} + N_{A\bar{B}}}{N} = \frac{N_{AB}}{N} + \frac{N_{A\bar{B}}}{N}. \qquad (1.8)$$

В формуле (1.8) N_{AB} обозначает число людей, подверженных болезни и давших положительный отклик на тест, и $N_{A\bar{B}}$ — число людей, не подверженных этой болезни, но также давших положительный отклик. Домножая и деля первое из двух выражений в правой части соотношения (1.8) на N_B, т. е. на число людей, имеющих болезнь, находим, что

$$\frac{N_{AB}}{N} = \frac{N_{AB}}{N_B} \cdot \frac{N_B}{N} = P(A|B)\ P(B). \qquad (1.9)$$

Аналогично, домножая и деля второй член на $N_{\bar{B}}$, число людей без болезни, находим

$$\frac{N_{A\bar{B}}}{N} = \frac{N_{A\bar{B}}}{N_{\bar{B}}} \quad \frac{N_{\bar{B}}}{N} = P(A|\bar{B})\ P(\bar{B}). \qquad (1.10)$$

Подставляя выражения (1.9) и (1.10) в формулу (1.8), получаем

$$P(A) = P(A|B)\ P(B) + P(A|\bar{B})\ P(\bar{B}). \qquad (1.11)$$

Это уравнение — специальный случай общего результата, говорящего о том, что безусловная частота $P(A)$ есть взвешенное среднее специфических частот — $P(A|B)$ и $P(A|\bar{B})$ — с весами, пропорциональными числу людей в специфических категориях или частотах $P(B)$ и $P(\bar{B})$.

Так как $P(\bar{B}) = 1 - P(B)$, то формула (1.11) преобразуется в следующее соотношение:

$$P(A) = P(A|B)\ P(B) + P(A|\bar{B})\ (1 - P(B)) =$$
$$= P(A|\bar{B}) + P(B)\,[P(A|B) - P(A|\bar{B})]. \qquad (1.12)$$

Подставляя (1.12) в (1.6), получаем выражение для ложной положительной частоты

$$P_{F+} = \frac{P(A|\bar{B})\ (1 - P(B))}{P(A|\bar{B}) + P(B)\ [P(A|B) - P(A|\bar{B})]}. \qquad (1.13)$$

Подставляя (1.12) в (1.7), получаем выражение для ложной отрицательной частоты

$$P_{F-} = \frac{(1 - P(A|B)) \cdot P(B)}{1 - P(A|\bar{B}) - P(B)\,[P(A|B) - P(A|\bar{B})]}. \qquad (1.14)$$

Анализ ложных положительных и отрицательных частот, связанных с критериями отбора и диагностики, был проведен Кохрэйном и Холлэндом [Cochrane, Holland, 1971], а также Галеном и Гамбино [Galen, Gambino, 1975] для различных заболеваний. Из (1.13) и (1.14) видно, что в целом обе ложные частоты являются функциями пропорций $P(A|B)$ и $P(A|\bar{B})$, которые могут быть оценены по результатам испытаний критерия отбора и априорной частоте $P(B)$, для сколь-нибудь точной оценки которой исследователь имеет достаточную информацию. Тем не менее значения вероятностей для ложных частот могут быть определены тем же способом, как и в рассмотренном ниже примере.

Предположим, что критерий был использован для диагностики 1000 человек, о которых заранее известно, что они имеют некоторое заболевание, и к выборке в 1000 человек, о которых неизвестно, имеют ли они это заболевание или нет. Результаты тестирования в частотах показаны в табл. 1.1.

Таблица 1.1

Результаты испытания критерия отбора

Наличие болезни	Результаты тестирования		Всего
	$+(A)$	$(\bar{A})$	
Присутствует (B)	950	50	1000
Отсутствует $(\bar{B})$	10	990	1000

По данным, приведенным в таблице, вычислим оценку чувствительности и оценку альтернативности использованного критерия:

$$P(A|B) = 950/1000 = 0{,}95;$$

$$P(\bar{A}|\bar{B}) = 990/1000 = 0{,}99.$$

Таким образом мы получили две величины, характеризующие тест с точки зрения его чувствительности и альтернативности к изучаемому заболеванию.

Подставляя эти две вероятности в (1.13), находим значение для ложной положительной частоты

$$P_{F+} = \frac{0{,}01\ (1 - P\ (B))}{0{,}01 + P\ (B)\ (0{,}95 - 0{,}01)} = \frac{0{,}01\ (1 - P\ (B))}{0{,}01 + 0{,}94\ P\ (B)} = \frac{1 - P(B)}{1 + 94\ P\ (B)}. \quad (1.15)$$

Окончательное выражение получено домножением числителя и знаменателя предыдущего выражения на 100. Подставляя их в (1.14), получаем значение для ложной отрицательной частоты

$$P_{F-} = \frac{(1-0{,}95)\ P\ (B)}{1-0{,}01-P\ (B)\ (0{,}95-0{,}01)} = \frac{0{,}05\ P\ (B)}{0{,}99-0{,}94\ P\ (B)} = \frac{5\ P\ (B)}{99-94\ P\ (B)}. \quad (1.16)$$

В табл. 1.2 приведены ложные частоты, соответствующие различным значениям $P\ (B)$, общим априорным частотам. Для реальных популяций частота $P\ (B)$ редко превышает 1%.

Таблица 1.2

Ложные частоты, связанные с критерием отбора

$P(B)$	Ложная положительная (P_{F+})	Ложная отрицательная (P_{F-})
1/1000000	0,9999	0
1/100000	0,9991	0
1/10000	0,9906	0,00001
1/1000	0,913	0,00005
1/500	0,840	0,00010
1/200	0,677	0,00025
1/100	0,510	0,00051

Если заболевание не слишком распространенное — если оно затрагивает, скажем, менее 1% населения — ложная отрицательная частота будет совсем маленькой, но ложная положительная частота будет весомой. С одной точки зрения тест удачен: так как P_{F-} меньше, чем 5/10000, то из каждых 10000 человек с отрицательным откликом на тест и, таким образом, предположительно отнесенных к здоровым, не более пяти будут в действительности больными. С другой точки зрения тест неудачен: поскольку P_{F+} больше, чем 50/100, то, следовательно, из каждых 100 человек, давших положительный отклик и отнесенных, таким образом, к больным или, по крайней мере, к тем, кому рекомендуется продолжать обследование, более чем 50 человек будут в действительности здоровыми.

Окончательное решение о том, каким образом использовать тест, будет зависеть от серьезности заболевания и от стоимо-

сти дальнейшего тестирования или лечения. Поскольку ложная положительная частота значительна, используемый тест, хотя его трудно считать приемлемым в любых ситуациях, вполне оправдывает себя в случае тяжелых заболеваний. Частоты $P(A|B)$ и $P(\overline{A}|\overline{B})$, рассмотренные в этом примере, больше значений тех же частот, вычисленных при использовании тестов отбора, поэтому вызывает тревогу тот факт, что ложные положительные частоты других тестов, возможно, будут превышать 50 из 100 — частоту, определенную здесь в качестве нижнего предела (P_{F+}).

Метод уменьшения ложных частот, как положительных, так и отрицательных, состоит в повторении теста для каждого индивидуума некоторое число раз, что увеличивает его стоимость. Окончательный результат тестирования для данного индивидуума считается положительным, если индивидуум откликается положительно на каждое воздействие теста или если он откликается положительно в большинстве воздействий. В пунктах б и в задачи 1.2 показано, как улучшаются характеристики процедуры отбора, когда тест применяется дважды. Сандифер, Флейс и Грин [Sandifer, Fleiss, Green, 1968] показали, что для некоторых заболеваний лучшие результаты достигаются при трехкратном применении теста. Окончательный результат признавался положительным, если объект имел положительный отклик не менее чем в двух из трех испытаний, причем тест применялся в третий раз только к тем пациентам, которые имели и положительный, и отрицательный отклик после первых двух его применений.

Пациентам, дававшим положительный отклик оба раза в первых двух испытаниях, устанавливался положительный результат, а тем, кто оба раза давал отрицательный отклик, — отрицательный.

Когда серьезность болезни делает необходимым выявление всех больных, а не вероятностное определение доли имеющих это заболевание, как в предыдущем случае, возможна более аккуратная, но и более сложная, чем данная выше, оценка рабочих характеристик процедуры отбора. Соответствующий анализ первоначально был проведен Нейманом [Neyman, 1947] и позднее осуществлен Гринхаусом и Мантелом [Greenhous, Mantel, 1950] и Ниссен-Мейером [Nissen-Meyer, 1964]. Советуем читателям обратить внимание на статью Мак-Нейла, Килера и Аделстейна [Mc Neil, Keeler, Adelstein, 1975], также посвященную этому вопросу.

Надо заметить, что универсальных, принятых всеми исследователями определений для ложных положительной и отрицательной частот не существует. Гален и Гамбино [Galen, Gambino, 1975] и Роган и Гладен [Rogan, Gladen, 1978], например, ис-

пользуют определения, приведенные в данной работе. Другие применяют эти термины просто как дополнения к характеристикам специфичности и чувствительности теста, что весьма расточительно для совершенной терминологии. Когда обсуждают, какой смысл имеет здесь термин «ложная положительная частота», они ссылаются на «дополнительность предсказанного значения результата положительного тестирования». Аналогично «ложная отрицательная частота» в их терминологии есть «дополнение предсказанного значения отрицательного результата тестирования». Эти формулировки, мягко говоря, неуклюжи.

1.3. Смещение в результате изучения выборок из подпопуляции

Первые представления относительно связи между болезнью и предшествующей жизнью могут быть получены при изучении отдельных выборок, например, госпитализированных больных. Смещение может возникнуть тогда, когда полученная в отдельных выборках связь распространяется на всю популяцию (хотя не для всех пациентов в выборке госпитализированных больных итог лечения одинаков).

Классический пример смещения этого рода встречается в работе Пирла [Pearl, 1929]. В результате анализа большого числа вскрытий была установлена отрицательная связь между двумя болезнями — раком и туберкулезом, т. е. при вскрытиях было обнаружено, что эти два заболевания значительно реже встречаются одновременно, чем порознь. Пирл сделал вывод, что эта отрицательная связь имеет место не только для умерших, но и для живущих. Основываясь на этом выводе, он лечил безнадежных онкологических больных туберкулином (протеином туберкулезной бациллы) в надежде, что развитие рака будет приостановлено. Пирл проигнорировал тот факт, что если не все умершие с равной вероятностью могут быть подвергнуты вскрытию, то распространение выводов, построенных по результатам вскрытий, на живущих больных — неверно. Между тем скорее всего отмеченной связи для живых пациентов нет, так как строгая связь была установлена при случайном попадании умерших на вскрытие.

Такой род смещения, названный ошибкой Берксона по имени исследователя, впервые изучившего ее детально [Berkson, 1946, 1955], возможен всякий раз, когда с помощью «перемены темы разговора» в рассматриваемую выборку вносятся изменения. Ошибка Берксона также изучалась Мэйнлендом [Mainland, 1953, 1963, с. 117—124], Уайтом [White, 1953],

Мантелом и Ханзелом [Mantel, Haenszel, 1959] и Брауном [Brown, 1976]. Любопытно, что наличие ошибки не было продемонстрировано эмпирически до тех пор, пока не появилось сообщение Робертса, Спитцера, Делмора и Сэкетта [Roberts, Spitzer, Delmore, Sackett, 1978]. Это сообщение проиллюстрировано данными, взятыми из работы Сэкетта [Sackett, 1979]:

Из 2784 индивидуумов, случайно выбранных в ходе индивидуального собеседования из популяции, 257 человек были госпитализированы в течение первых шести месяцев. В табл. 1.3 для лиц, помещенных в больницу, приведены данные о наличии или отсутствии у них респираторного заболевания и болезней суставов или органов движения (называемых локомоторными заболеваниями).

Таблица 1.3

Связь между локомоторным и респираторным заболеваниями в подвыборке госпитализированных больных

Респираторное заболевание	Локомоторное заболевание		Всего	Доля больных с локомоторным заболеванием
	Наличие	Отсутствие		
Наличие	5	15	20	$0{,}25=p_1$
Отсутствие	18	219	237	$0{,}08=p_2$
Всего	23	234	257	$0{,}09=p$

Ясна связь между госпитализированными по тем или иным причинам больными, которым поставлен или не поставлен диагноз респираторное заболевание, и госпитализированными больными, имеющими или не имеющими локомоторных заболеваний: доля больных с локомоторным заболеванием среди имеющих респираторное заболевание равна $p_1=0{,}25$, что в три раза выше доли пациентов с локомоторным заболеванием среди не имеющих респираторных заболеваний, $p_2=0{,}08$. Однако возникает вопрос: насколько корректен вывод о том, что эти два вида заболеваний сопутствуют друг другу?

Вывод не верен, сопутствие не обязательно. В действительности эти две характеристики — респираторные и локомоторные заболевания — независимы в людской популяции. Как показано в табл. 1.4, частоты локомоторных заболеваний практически одинаковы для людей, больных респираторными заболеваниями и не подверженных им.

Источником наблюдаемого парадокса являются вариации среди четырех возможных групп людей (присутствуют оба заболевания, присутствует только респираторное заболевание,

Таблица 1.4

Связь локомоторного и респираторного заболеваний в генеральной совокупности

Респираторное заболевание	Локомоторное заболевание		Всего	Доля лиц с локомоторным заболеванием
	Наличие	Отсутствие		
Наличие	17	207	224	0,08
Отсутствие	184	2376	2560	0,07
Всего	201	2583	2784	0,07

присутствует только локомоторное заболевание, оба заболевания отсутствуют) в вероятности их госпитализации. Для сравнения могут быть выбраны частоты в соответствующих ячейках табл. 1.3 и 1.4. Сравнение показывает, что частота госпитализации больных с обоими заболеваниями (25%) была более чем в три раза выше каждой из частот для других групп людей, которые составляют 7—10%.

Ошибка Берксона всегда возможна в тех случаях, когда исходные частоты для людей с различными комбинациями факторов меняются или могут быть неосознанно изменены, в то время как болезнь или болезни, о которых идет речь, требуют особой осторожности в построении выводов (как, например, при лейкемии и некоторых других онкологических заболеваниях). Основа этой ошибки следующая. Пусть B обозначает событие, что объект имеет один из двух признаков, которые мы изучаем (в нашем случае, респираторное заболевание), и $\overline{B}$ — событие, состоящее в том, что объект этим признаком не обладает. Пусть $P(B)$ обозначает долю всех людей в сообществе, которые имеют признак B, и $P(\overline{B}) = 1 - P(B)$ — долю всех людей, не имеющих этого признака.

Пусть A обозначает событие, что объект проявляет при изучении другой признак (в нашем случае, локомоторное заболевание), и $\overline{A}$ — событие, что объект не проявляет этого признака, и пусть $P(A)$ и $P(\overline{A}) = 1 - P(A)$ обозначают соответствующие доли людей в популяции. Пусть $P(A$ и $B)$ — доля всех людей в популяции, имеющих оба признака, и предположим, что эти два признака независимы. Таким образом, из формулы (1.3) получаем, что $P(A \text{ и } B) = P(A)\,P(B)$.

Пусть H обозначает событие, состоящее в том, что больной госпитализирован с одним из этих признаков. Определим $P(H \mid B \text{ и } A)$ — как долю от всех людей, имеющих оба признака и госпитализированных, $P(H \mid B \text{ и } \overline{A})$ — как долю от всех лю-

дей, которые имели один признак, B, но не имели другого признака, A, и были при этом госпитализированы; $P(H|\overline{B}$ и $A)$ и $P(H|\overline{B}$ и $\overline{A})$ определяются аналогично. Проблема состоит в том, чтобы оценить, в терминах вероятностей,

$$p_1 = P(A|B \text{ и } H),$$

т. е. долю людей в изучаемой популяции, которые имеют признак B и госпитализированы и которые при этом имеют признак A, и

$$p_2 = P(A|\overline{B} \text{ и } H),$$

т. е. долю всех госпитализированных людей, не имеющих признака B, но которые также имеют признак A.

Воспользуемся следующим вариантом определения условной вероятности

$$p_1 = P(A|B \text{ и } H) = \frac{P(B \text{ и } A|H)}{P(B|H)}. \qquad (1.17)$$

Уравнение (1.17) отличается от (1.2) только одним, однако, важным моментом; второе условие H (для госпитализированных больных) остается условием, определяющим все вероятности.

Числитель (1.17) есть согласно теореме Байеса, см. (1.5),

$$P(B \text{ и } A|H) = \frac{P(H|B \text{ и } A)\,P(B \text{ и } A)}{P(H)} =$$

$$= \frac{P(H|B \text{ и } A)\,P(B)\,P(A)}{P(H)}, \qquad (1.18)$$

так как предполагается независимость событий A и B.

Знаменатель (1.17), по теореме Байеса можно записать в виде

$$P(B|H) = \frac{P(H|B)\,P(B)}{P(H)}. \qquad (1.19)$$

Чтобы найти $P(H|B)$, мы используем (1.11), а именно то, что общая частота есть взвешенное среднее специфических (условных) частот. Дополнительно потребуем, чтобы вероятность события B оставалась одной и той же. Таким образом,

$$P(H|B) = P(H|B \text{ и } A)\,P(A|B) + P(H|B \text{ и } \overline{A})\,P(\overline{A}|B) =$$

$$= P(H|B \text{ и } A)\,P(A) + P(H|B \text{ и } \overline{A})\,P(\overline{A}),$$

поскольку предположение о независимости A и B влечет $P(A|B) = P(A)$ и $P(\overline{A}|B) = P(\overline{A})$. Следовательно, (1.19) приобретает вид

$$P(B|H) = \frac{P(H)\,[P(H|B \text{ и } A)\,P(A) + P(H|B \text{ и } \overline{A})\,P(\overline{A})]}{P(H)}. \qquad (1.20)$$

Подставляя (1.18) в числитель формулы (1.17) и значение (1.20) — в знаменатель, получим

$$p_1 = \frac{P(H \mid B \text{ и } A)\, P(A)}{P(H \mid B \text{ и } A)\, P(A) + P(H \mid B \text{ и } \bar{A})\, P(\bar{A})}. \quad (1.21)$$

Аналогично

$$p_2 = \frac{P(H \mid \bar{B} \text{ и } A)\, P(A)}{P(H \mid \bar{B} \text{ и } A)\, P(A) + P(H \mid \bar{B} \text{ и } \bar{A})\, P(\bar{A})}. \quad (1.22)$$

Эти две вероятности не равны, если частоты госпитализаций не равны, даже если равны соответствующие вероятности в популяции, $P(A \mid B)$ и $P(A \mid \bar{B})$.

В нашем примере для события A, обозначающего локомоторное заболевание, $P(A) = 7\%$. Частоты госпитализаций различны:

$$P(H \mid B \text{ и } A) = 29\%,$$
$$P(H \mid B \text{ и } \bar{A}) = 7\%,$$
$$P(H \mid \bar{B} \text{ и } A) = 10\%,$$

и

$$P(H \mid \bar{B} \text{ и } \bar{A}) = 9\%.$$

Подставляя эти величины в (1.21) и в (1.22), находим значения

$$p_1 = \frac{0{,}29 \cdot 0{,}07}{0{,}29 \cdot 0{,}07 + 0{,}07 \cdot 0{,}93} = \frac{0{,}020}{0{,}085} = 0{,}24;$$

$$p_2 = \frac{0{,}10 \cdot 0{,}07}{0{,}10 \cdot 0{,}07 + 0{,}09 \cdot 0{,}93} = \frac{0{,}007}{0{,}091} = 0{,}08,$$

близкие к приведенным в табл. 1.3.

Вывод, который можно сделать из этого примера, ясен. Надо относиться с изрядной долей скептицизма к любому обобщению на всех людей популяции зависимостей, выведенных лишь для госпитализированных пациентов или по результатам вскрытий. Это же предостережение очевидным образом относится и к обобщениям данных, полученных из отчетов исследований, объектами которых выступали добровольцы.

1.4. Выводы, основанные на одной пропорции

Интерес к частотам в большинстве случаев обусловлен возможностью сравнения двух или более пропорций. Иногда, однако, частота вызывает интерес сама по себе. Этот раздел представляет собой краткий обзор соответствующих методов.

Пусть n обозначает число людей в изучаемой выборке, p — доля людей в выборке, обладающих некоторым признаком, и P — истинная, но априори неизвестная доля лиц в популяции, обладающих этим признаком. Заключение относительно величины P может быть получено с помощью биномиального распределения вероятностей, которое описано во всех хороших учебниках по статистике. Когда n велико (в том смысле, что $nP \geqslant 5$ и $nQ \geqslant 5$, где $Q = 1 - P$), следующие процедуры (алгоритмы), основанные на нормальном распределении, дают идеальные аппроксимации соответствующих точных биномиальных процедур.

Пусть P — истинная вероятность, когда выборочная частота p аппроксимируется нормальным распределением со средним P и стандартной ошибкой

$$\text{s. e. } (p) = \sqrt{\frac{PQ}{n}}. \tag{1.23}$$

Для того чтобы проверить гипотезу, что P равно некоторому ранее установленному P_0 против альтернативной гипотезы, что $P \neq P_0$, надо вычислить критическое отношение

$$z = \frac{|p - P_0| - 1/(2n)}{\sqrt{\frac{P_0 Q_0}{n}}}, \tag{1.24}$$

где $Q_0 = 1 - P_0$. Гипотеза отвергается, если z превосходит критическое значение нормальной кривой для заданного двустороннего уровня значимости. Величина $1/(2n)$, подставленная в числитель, является поправкой на непрерывность, компенсирующей ошибку, вносимую нормальной вероятностной кривой в точные значения с биномиальными вероятностями. Она применяется только тогда, когда ее величина меньше, чем $|p - P_0|$.

Рассмотрим данные табл. 1.3. Размер выборки $n = 257$, и доля больных с локомоторным заболеванием равна $p = 23/257 = 0{,}09$. Предположим (для иллюстрации результатов), что частота локомоторного заболевания в сравниваемой госпитальной выборке есть $P_0 = 0{,}05$ Будет ли представленная выборка типичной или атипичной по отношению к этому условию? Значение критической величины, данное в (1.24), есть

$$z = \frac{\left|0{,}09 - 0{,}05\right| - \frac{1}{2 \cdot 257}}{\sqrt{\frac{0{,}05 \cdot 0{,}95}{257}}} = 2{,}80.$$

Оно указывает на имеющееся различие при уровне значимости 0,01 между вероятностями заболевания локомоторной бо-

лезнью в подвыборке госпитализированных больных и общей выборке.

Может возникнуть желание построить для истинной вероятности доверительный интервал. Необходимо, чтобы исследователи отдавали себе отчет в том, что доверительные интервалы по своей сути более связаны с критериями значимости, чем с простыми точечными оценками. В нашем случае $100(1-\alpha)\%$-ный доверительный интервал для истинной пропорции (обычные значения для α равны 0,01 и 0,05) состоит из всех таких значений P, которые не могут быть отвергнуты двусторонним критерием на уровне значимости α. Если критерий основан на отношении, данном в (1.24), и если $c_{\alpha/2}$ обозначает величину, отсекающую площадь $\alpha/2$ в верхней части стандартного нормального распределения, аппроксимационный $100(1-\alpha)\%$-ный доверительный интервал состоит из всех таких значений P, которые удовлетворяют неравенству

$$\frac{|p-P|-1/(2n)}{\sqrt{\frac{PQ}{n}}} \leqslant c_{\alpha/2}. \tag{1.25}$$

Пределы этого интервала задаются двумя корнями квадратичного уравнения, полученного возведением в квадрат левой части неравенства (1.25), приравненной к $c^2_{\alpha/2}$. Обозначим $q=1-P$. Нижний и верхний пределы задаются точными формулами (см. задачу 1.5):

$$P_L = \frac{(2np + c^2_{\alpha/2} - 1) - c_{\alpha/2}\sqrt{c^2_{\alpha/2} - \left(2 + \frac{1}{n}\right) + 4p(nq+1)}}{2(n + c^2_{\alpha/2})} \tag{1.26}$$

и

$$P_U = \frac{(2np + c^2_{\alpha/2} + 1) + c_{\alpha/2}\sqrt{c^2_{\alpha/2} + \left(2 - \frac{1}{n}\right) + 4p(nq-1)}}{2(n + c^2_{\alpha/2})}. \tag{1.27}$$

Для анализируемых данных $n=257$, $p=0{,}09$ и $q=0{,}91$. Если построить 95%-ный доверительный интервал для P, то $c_{\alpha/2}=c_{0{,}025}=1{,}96$. Можно проверить, что $P_L=0{,}059$ и $P_U=0{,}133$, так что

$$0{,}059 \leqslant P \leqslant 0{,}133 \tag{1.28}$$

есть аппроксимационный 95%-ный доверительный интервал для P. Кроме того, можно проверить, что если $P_0=0{,}059$ или $P_0=0{,}133$ — гипотетические значения P, то результирующая величина z в (1.24) в точности равна 1,96.

Предыдущая, несколько усложненная процедура нахождения доверительного интервала для P предпочтительна, когда p

близко к нулю или к единице. Если же p имеет умеренные значения (скажем, $0{,}3 \leqslant p \leqslant 0{,}7$), может быть применена другая, более простая и более привычная процедура.

Вследствие того что значение выражения $\sqrt{x(1-x)}$ остается достаточно постоянным для $0{,}3 \leqslant x \leqslant 0{,}7$, величина $\sqrt{PQ}$ в знаменателе (1.25) может быть заменена на $\sqrt{pq}$. Получаем интервал

$$p - c_{\alpha/2}\sqrt{\frac{pq}{n} - \frac{1}{2n}} \leqslant P \leqslant p + c_{\alpha/2}\sqrt{\frac{pq}{n} + \frac{1}{2n}}, \qquad (1.29)$$

который можно использовать в качестве аппроксимационного $100(1-\alpha)\%$-ного доверительного интервала для P. Для рассматриваемых данных результирующий интервал есть

$$0{,}053 \leqslant P \leqslant 0{,}126. \qquad (1.30$$

Он сдвинут влево по отношению к формально более обоснованному интервалу, данному в (1.28).

Когда p очень мало, сдвиг влево интервала (1.29) по отношению к интервалу, задаваемому (1.26) и (1.27), может оказаться опасным. Задача 1.4 иллюстрирует возможные аномалии в выводах в такой ситуации.

Задачи

1.1. Признаки A и B независимы, если $P(A \text{ и } B) = P(A)P(B)$. Покажите, что если это так, то $P(A \text{ и } \overline{B}) = P(A)P(\overline{B})$; $P(\overline{A} \text{ и } B) = P(\overline{A})P(B)$ и $P(\overline{A} \text{ и } \overline{B}) = P(\overline{A})P(\overline{B})$. (Указание. $P(A) = P(A \text{ и } \overline{B}) + P(A \text{ и } B)$, так что $P(A \text{ и } \overline{B}) = P(A) - P(A \text{ и } B)$. Используйте данное соотношение, $P(A \text{ и } B) = P(A)P(B)$ и тот факт, что $P(\overline{B}) = 1 - P(B)$.)

1.2. Допустим, что случайная частота некоторой болезни, $P(B)$, характеризуется одним случаем из 1000 и что тест отбора больных, пораженных этой болезнью, находится в стадии изучения.

а) Предположим, что тест применен в данный момент к выборке людей, пораженных каким-то заболеванием, из которых 99% дают на него положительный отклик. Предположим, что этот тест также применен к выборке людей, не имеющих этого заболевания, из них 1% дает положительный отклик. Как в данном случае определить ложную положительную и ложную отрицательную частоты? Считаете ли Вы эти тесты хорошими?

б) Предположим теперь, что тест должен быть применен дважды с установлением окончательного положительного результата, если он оба раза дал положительный результат. Предположим далее, что согласно проверяемому определению 98% больных из выборки дают положительный отклик, но только один из 10 000 людей, не имеющих этого заболевания, дает положительный отклик. Как определяются ложная отрицательная и ложная положительная частоты? Стоит ли применять этот критерий отбора при таких проверяемых условиях?

в) Заметим, что не ко всем людям тест применяется дважды. Тест применяется второй раз, если первый результат тестирования был положительным; окончательный результат будет принят положительным, только если оба испытания дадут положительный исход. Важно оценить долю всех людей, по отношению к которым тест будет применяться вторично, т. е. тех, кто даст положительный результат при первом тестировании. Что это за до-

ля? Из каждых 100 000 людей, тестированных один раз, к скольким людям потребуется применить тест второй раз?

1.3. Может появиться тип смещения, противоположный тому, который был рассмотрен в разд. 1.3. Итак, два признака могут быть зависимы в популяции, но могут быть и независимы, когда изучается выборка госпитализированных больных.

Пусть A — событие, состоящее в том, что человек живет один, $\bar{A}$ — живет с семьей, B — имеет невроз, $\bar{B}$ — не имеет невроза. Предположим, что $P(A|B)=0{,}40$ и $P(A|\bar{B})=0{,}20$. Таким образом, живут одиноко 40% больных неврозом и 20% не больных неврозом. Предположим, что в большой популяции 100 000 человек имеют невроз и 1 000 000 не имеют этого заболевания.

а) Рассмотрим сначала 100 000 больных неврозом. (1). Какая доля их живет одиноко? (2). Если среднегодовая частота госпитализации больных неврозом, живущих одиноко, равна 5/1000, то определите, сколько таких людей будет госпитализировано? Заметим, что это есть число госпитализированных больных неврозом, живущих одиноко, т. е. числитель p_1. (3). Сколько человек из 100 000 больных неврозом живут со своей семьей? (4). Если среднегодовая частота госпитализации для больных неврозом, живущих со своей семьей, равна 6/1000, сколько таких людей будет госпитализировано? Заметим, что сумма чисел, найденных в (2) и (4), есть общее число госпитализированных больных неврозом, т. е. знаменатель p_1. (5). Чему равна величина p_1, доля госпитализированных больных неврозом, живущих одиноко? (6). Как сравнить p_1 с $P(A|B)$, долей больных неврозом в популяции, живущих одиноко?

б) Рассмотрим теперь миллион людей, не больных неврозом. (1). Сколько из них живут одиноко? (2). Если ежегодная частота госпитализации для тех из них, кто живет одиноко, равна 5/1000, сколько таких людей будет госпитализировано? Заметим, что это есть число больных, госпитализированных не по поводу невроза, найденных из числа живущих одиноко, т. е. числитель p_2. (3). Сколько из 1 000 000 не больных неврозом живет со своей семьей? (4). Если ежегодная частота госпитализации не больных неврозом, живущих со своими семьями, есть 225/100 000, сколько таких людей будет госпитализировано? Заметим, что сумма чисел, найденных в пунктах (б2) и (б4), есть общее число госпитализированных людей, не больных неврозом, т. е. знаменатель p_2. (5). Чему равна величина доли госпитализированных больных, не имеющих невроза, живущих одиноко? (6). Как сравнить p_2 с $P(A|\bar{B})$, долей людей популяции, не больных неврозом, живущих одиноко?

в) Какой вывод может быть сделан на основании сравнения p_1 и p_2? Как сравнить p_2 с $P(A|B)$, долей людей популяции, не больных неврозом, с $P(A|\bar{B})$?

1.4. Предположим, что в выборке из $n=100$ объектов доля $p=0{,}05$ имеет специфическую характеристику.

а) С точностью до двух десятичных знаков найдите нижнюю и верхнюю 99%-ные доверительные границы для P, воспользовавшись (1.26) и (1.27). Используйте $c_{0{,}005}=2{,}576$.

б) С точностью до двух десятичных знаков найдите нижнюю и верхнюю 99%-ные доверительные границы для P, используя (1.29).

в) Как сравнить два интервала? Произойдет ли какое-нибудь улучшение (см. п. (б)), если игнорировать поправку на непрерывность?

1.5. Докажите, что (1.26) и (1.27) являются двумя решениями уравнения

$$\frac{n(|p-P|-1/(2n))^2}{PQ}=c^2_{\alpha/2}.$$

(Указание. Исходите из того, что целевое назначение поправки на непрерывность состоит во внесении различия между p и P, изначально близкого к нулю. Установленные нижний и верхний пределы, следовательно, «работают» с величинами в числителе, равными соответственно $p-P-\frac{1}{2n}$ и $p-P+\frac{1}{2n}$.)

1.6. Докажите, что P_L из (1.26) никогда не бывает меньше нуля, а P_U из (1.27) — больше единицы.

ЛИТЕРАТУРА

Berkson, J. (1946). Limitations of the application of fourfold table analysis to hospital data. *Biom. Bull.* (now *Biometrics*), **2**, 47–53.

Berkson, J. (1955). The statistical study of association between smoking and lung cancer. *Proc. Staff Meet. Mayo Clin.*, **30**, 319–348.

Brown, G. W. (1976). Berkson fallacy revisited: Spurious conclusions from patient surveys. *Am. J. Dis. Child.*, **130**, 56–60.

Cochrane, A. L. and Holland, W. W. (1971). Validation of screening procedures. *Br. Med. Bull.*, **27**, 3–8.

Galen, R. S. and Gambino, S. R. (1975). *Beyond normality: The predictive value and efficiency of medical diagnoses.* New York: Wiley.

Greenhouse, S. W. and Mantel, N. (1950). The evaluation of diagnostic tests. *Biometrics*, **6**, 399–412.

Mainland, D. (1953). The risk of fallacious conclusions from autopsy data on the incidence of diseases with applications to heart disease. *Am. Heart J.*, **45**, 644–654.

Mainland, D. (1963). *Elementary medical statistics*, 2nd ed. Philadelphia: W. W. Saunders.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.*, **22**, 719–748.

McNeil, B. J., Keeler, E., and Adelstein, S. J. (1975). Primer on certain elements of medical decision making. *New Engl. J. Med.*, **293**, 211–215.

Neyman, J. (1947). Outline of statistical treatment of the problem of diagnosis. *Public Health Rep.*, **62**, 1449–1456.

Nissen-Meyer, S. (1964). Evaluation of screening tests in medical diagnosis. *Biometrics*, **20**, 730–755.

Pearl, R. (1929). Cancer and tuberculosis. *Am. J. Hyg.* (now *Am. J. Epidemiol.*), **9**, 97–159.

Roberts, R. S., Spitzer, W. O., Delmore, T., and Sackett, D. L. (1978). An empirical demonstration of Berkson's bias. *J. Chronic Dis.*, **31**, 119–128.

Rogan, W. J. and Gladen, B. (1978). Estimating prevalence from the results of a screening test. *Am. J. Epidemiol.*, **107**, 71–76.

Sackett, D. L. (1979). Bias in analytic research. *J. Chronic Dis.*, **32**, 51–63.

Sandifer, M. G., Fleiss, J. L., and Green, L. M. (1968). Sample selection by diagnosis in clinical drug evaluations. *Psychopharmacologia*, **13**, 118–128.

White, C. (1953). Sampling in medical research. *Br. Med. J.*, **2**, 1284–1288.

Yerushalmy, J. (1947). Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. *Public Health Rep.*, **62**, 1432–1449.

Глава 2

Проверка значимости по данным четырехклеточных таблиц сопряженности

Четырехклеточная таблица сопряженности (см. табл. 2.1) была и, вероятно, остается наиболее часто используемым способом представления статистических данных. Проверка значимости зависимости между признаками по таким таблицам чаще всего осуществляется — и это самый простой метод — с помощью классического критерия хи-квадрат. Этот критерий основан на величине статистики

$$\chi^2 = \frac{n_{..}\,(|\,n_{11}\,n_{22} - n_{12}\,n_{21}\,| - \frac{1}{2}\,n_{..})^2}{n_{1.}\,n_{2.}\,n_{.1}\,n_{.2}}. \qquad (2.1)$$

Вычисленное значение χ^2 сравнивается с табличными значениями функции распределения хи-квадрат с одной степенью свободы (см. табл. А.1 в приложении). Если величина статистики χ^2 превышает табличное значение с заданным уровнем значимости, то считается, что между признаками A и B имеется зависимость. Интересный графический способ проверки значимости предложен в [Zubin, 1939].

Вычислим, например, значение χ^2 для гипотетических данных, приведенных в табл. 2.2, по формуле (2.1):

$$\chi^2 = \frac{200\,(|\,15\cdot 40 - 135\cdot 10\,| - \frac{1}{2}\,200)^2}{150\cdot 50\cdot 25\cdot 175} = 2{,}58. \qquad (2.2)$$

Для того чтобы можно было утверждать на уровне значимости 0,05, что существует зависимость между признаками, значение χ^2 должно было бы превысить 3,84. Таким образом, на основе вычисленных данных такой вывод сделать нельзя.

В том, что статистика хи-квадрат (2.1) крайне проста для расчета, имеется и негативная сторона. Во-первых, ее вычисление не требует от исследователя нахождения истинных пропорций для всей изучаемой генеральной совокупности, так как в формуле используются лишь частоты. Во-вторых, исследова-

Таблица 2.1

Модель четырехклеточной таблицы сопряженности

Признак A	Признак B		
	Наличие	Отсутствие	Всего
Наличие	n_{11}	n_{12}	$n_{1.}$
Отсутствие	n_{21}	n_{22}	$n_{2.}$
Всего	$n_{.1}$	$n_{.2}$	$n_{..}$

Таблица 2.2

Гипотетическая четырехклеточная таблица сопряженности

Признак A	Признак B		
	Наличие	Отсутствие	Всего
Наличие	15	135	150
Отсутствие	10	40	50
Всего	25	175	200

тель может не обратить внимания на то, каким образом были получены данные, так как формула этого не отражает, и сделать неверный вывод, основываясь на величине статистики χ^2. Эти соображения развиваются в разд. 2.1, где рассмотрены три метода получения данных для четырехклеточной таблицы сопряженности, и для каждого метода указана проверяемая статистическая гипотеза.

«Точный» критерий Фишера—Ирвина [Fisher, 1935], [Irwin, 1935] представлен в разд. 2.2; необходимость включения в формулы для вычисления статистик хи-квадрат и критического отношения (critical ratio) поправок Иэйтса [Yates, 1934] на непрерывность обсуждается в разд. 2.3. Некоторые рекомендации, касающиеся выбора между односторонним и двусторонним критериями значимости, предложены в разд. 2.4. Разд. 2.5 посвящен построению доверительных границ для разности двух независимых пропорций. В разд. 2.6 рассматривается отличающийся от классического критерий критического отношения, который тесно связан с построением доверительных интервалов.

2.1. Методы получения четырехклеточной таблицы

Обычно на практике применяются лишь следующие три метода выбора объектов с целью получения частотных четырехклеточных таблиц (см. [Barnard, 1947], где эта тема обсуждается более подробно).

Метод выбора 1. Первый метод выбора, называемый перекрестным (cross-sectional) отбором, состоит в том, что из некоторой совокупности выбирается $n_{..}$ объектов и для каждого объекта устанавливается, присутствуют или отсутствуют у него признаки A и B. До сбора данных назначается лишь размер выборки $n_{...}$.

Многие выборочные обследования проводятся именно таким образом. Приведем несколько примеров применения метода выбора 1. При изучении качества медицинского обслуживания в клинике всех пациентов, впервые участвовавших в некотором обследовании, можно разделить на группы в зависимости от пола и прохождения или непрохождения какого-то определенного осмотра во время обследования. При изучении распространения некоторой болезни в популяции можно случайным образом выбрать людей для обследования и классифицировать их в зависимости от расы и присутствия или отсутствия некоторого симптома болезни. При изучении связи между весом новорожденного и возрастом матери все роды в родильном доме можно классифицировать в зависимости от веса ребенка и возраста матери.

Целью исследования, основанного на использовании метода выбора 1, является выяснение, зависимы признаки A и B или нет. Пропорции (неизвестные, конечно, исследователю) для генеральной совокупности, из которой была извлечена выборка, представлены в табл. 2.3.

Таблица 2.3

Совместные пропорции признаков A и B в генеральной совокупности

Признак A	Признак B		
	Наличие	Отсутствие	Всего
Наличие	P_{11}	P_{12}	$P_{1.}$
Отсутствие	P_{21}	P_{22}	$P_{2.}$
Всего	$P_{.1}$	$P_{.2}$	1

По определению (см. разд. 1.1) признаки A и B независимы тогда и только тогда, когда каждая совместная пропорция (например P_{12}) является произведением двух соответствующих безусловных или маргинальных пропорций (в этом примере $P_1 P_2$.). Имеют ли пропорции это свойство в действительности, можно определить лишь по тому, насколько близки совместные выборочные пропорции к соответствующим произведениям маргинальных. Выборочная таблица классификации должна, следовательно, быть аналогом табл. 2.3. Она получается в результате деления каждой частоты в табл. 2.1 на $n_{..}$. Результат деления представлен в виде табл. 2.4.

Таблица 2.4

Совместные пропорции для A и B в выборке

Признак A	Признак B		
	Наличие	Отсутствие	Всего
Наличие	p_{11}	p_{12}	$p_{1.}$
Отсутствие	p_{21}	p_{22}	$p_{2.}$
Всего	$p_{.1}$	$p_{.2}$	1

Степень доверия к гипотезе о независимости признаков A и B зависит от величин четырех разностей $p_{ij} - P_{i.} P_{.j}$, где i и j равны 1 или 2. Чем меньше эти разности, тем лучше данные согласуются с гипотезой. Чем они больше, тем более сомнительной становится гипотеза. (Фактически достаточно вычислить только одну из этих четырех разностей, так как остальные равны ей с точностью до знака — см. задачу 2.1.)

Пирсон [Pearson, 1900] предложил критерий для проверки значимости этих разностей. Статистикой этого критерия, включающей поправку на непрерывность, является

$$\chi^2 = n_{..} \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(|p_{ij} - p_{i.} p_{.j}| - 1/(2 n_{..}))^2}{p_{i.} p_{.j}} . \qquad (2.3)$$

Задача 2.2 посвящена доказательству того, что выражения (2.1) и (2.3) идентичны. По таблице распределения хи-квадрат с одной степенью свободы можно определить, будет ли вычисленная величина статистики достаточно большой, чтобы считаться значимой. Если значение χ^2 значимо, то исследователь может сделать вывод, что признаки A и B зависимы, и перейти к описанию степени их зависимости. Глава 5 посвя-

щена методам описания зависимости, установленной по данным, полученным с помощью метода выбора I.

Предположим, что табл. 2.2 содержит результаты исследования, в котором использовался метод выбора I. После их нормировки в табл. 2.5 представлены данные. Отметим, например, что $p_{22}=0{,}20$, в то время как если бы A и B были независимы, то ожидаемое значение пропорции было бы равно $p_{2.}p_{.2}=0{,}25\cdot 0{,}875=0{,}21875$. Кроме того, заметим, что каждая из четырех разностей, участвующих в формуле (2.3), равна $|\pm 0{,}01875|-0{,}0025=0{,}01625$.

Таблица 2.5

Совместные пропорции для гипотетических даиных из табл. 2.2

Признак A	Признак B		
	Наличие	Отсутствие	Всего
Наличие	0,075	0,675	0,75
Отсутствие	0,050	0,200	0,25
Всего	0,125	0,875	1

Применительно к данным табл. 2.5, формула (2.3) дает значение

$$\chi^2 = 200\left(\frac{0{,}01625^2}{0{,}09375}+\frac{0{,}01625^2}{0{,}65625}+\frac{0{,}01625^2}{0{,}03125}+\frac{0{,}01625^2}{0{,}21875}\right)=2{,}58\,, \quad (2.4)$$

которое было ранее получено по формуле (2.2).

Метод выбора II. Второй метод выбора, иногда называемый целевым (purposive) отбором, состоит в том, что для анализа отбирается заранее установленное число, $n_{1.}$, объектов, которые имеют признак A и заранее установленное число, $n_{2.}$ объектов, не имеющих признака A. На основе этого метода проводятся сравнительные проспективные и ретроспективные обследования. При проспективном обследовании изучаются две группы — первая, состоящая из $n_{1.}$ объектов, у которых присутствует исследуемый фактор риска, и вторая, состоящая из $n_{2.}$ объектов, у которых он отсутствует. Определяется, сколько случаев развития болезни будет зафиксировано в каждой из групп. При ретроспективном обследовании двух групп, состоящих соответственно из $n_{1.}$ здоровых людей и из $n_{2.}$ больных, выясняется, какое количество индивидуумов в каждой из групп ранее имело исследуемый фактор риска.

Целью исследования при применении метода выбора II является определение, равны или нет пропорции, скажем P_1 и P_2, в генеральных совокупностях, откуда были извлечены выборки. Для этого необходимо, чтобы данные были представлены в таком виде, при котором легко получить интересующую информацию о пропорциях. Подходящий способ представления данных показан в табл. 2.6.

Таблица 2.6

Пропорции для выделенного признака в двух независимых выборках

	Размер выборки	Пропорция
Выборка 1	$n_{1.}$	$p_1\ (=n_{11}/n_{1.})$
Выборка 2	$n_{2.}$	$p_2\ (=n_{21}/n_{2.})$
Объединенная	$n_{..}$	$\bar{p}\ (=n_{.1}/n_{..})$

Статистическая значимость разности между p_1 и p_2 оценивается с помощью статистики

$$z = \frac{|p_2 - p_1| - \frac{1}{2}(1/n_{1.} + 1/n_{2.})}{\sqrt{\bar{p}\,\bar{q}\,(1/n_{1.} + 1/n_{2.})}}, \qquad (2.5)$$

где $\bar{q} = 1 - \bar{p}$. Для проверки гипотезы, что P_1 и P_2 равны, можно воспользоваться таблицей стандартного нормального распределения. Если вычисленное значение z превышает табличное для заданного уровня значимости, то делается вывод, что P_1 и P_2 не равны. По определению квадрат величины, имеющей стандартное нормальное распределение, будет иметь распределение хи-квадрат с одной степенью свободы; тогда вычисленное значение может быть сопоставлено с табличным значением функции распределения хи-квадрат с одной степенью свободы. В задаче 2.3 требуется доказать, что z^2 эквивалентно (2.1).

Если будет установлено на требуемом уровне значимости, что P_1 и P_2 не равны, то дальнейший анализ проводится методами, рассматриваемыми в гл. 6, которые отличны от методов, используемых в случае применения метода выбора I.

Предположим для иллюстрации, что данные табл. 2.2 были получены при изучении выборок, сформированных в результате целевого отбора 150 объектов, имеющих признак A, и

ло объектов, его не имеющих. Табл. 2.7 содержит эти данные и их характеристики.

Таблица 2.7

Пропорции объектов с признаком *B* среди объектов, имеющих или не имеющих признак *A*, для гипотетических данных табл. 2.2

	Размер выборки	Пропорция объектов с признаком *B*
Наличие *A*	150	$0,10=p_1$
Отсутствие *A*	50	$0,20=p_2$
Всего	200	$0,125=\bar{p}$

По формуле (2.5)

$$z=\frac{|0,20-0,10|-\frac{1}{2}\left(\frac{1}{150}+\frac{1}{50}\right)}{\sqrt{0,125\cdot 0,875\left(\frac{1}{150}+\frac{1}{50}\right)}}=1,60, \qquad (2.6)$$

что меньше табличного значения 1,96, соответствующего уровню значимости 0,05. Квадрат полученной величины z равен 2,56, что совпадает, с точностью до ошибок округления, с величиной χ^2 в (2.2).

Метод выбора III. Третий метод выбора похож на второй в том, что противопоставляются две выборки заранее определенного размера. Однако в отличие от метода выбора II метод выбора III требует чтобы эти выборки формировались случайно. Этот метод лежит в основе контроля эффективности способа клинического лечения: из общего числа $n_{..}$ больных, n_1 отбирают случайно и лечат обычным способом, остальных же n_2 больных лечат тем способом, эффективность которого исследуется.

Важно найти пропорции больных в каждой из двух групп, которым лечение принесло желаемый результат (например, ремиссию симптомов). Значимость различия этих пропорций проверяется при помощи той же статистики (2.5), что и при использовании метода II. Однако последующая обработка данных, полученных по методу выбора III, рассматриваемая в гл. 7, отличается от соответствующих обработок данных, полученных по методам выбора I и II.

2.2. «Точный» анализ четырехклеточной таблицы

Приведем еще один вариант статистики χ^2. Он известен почти так же хорошо, как и вариант, данный в (2.1):

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\left(\left|n_{ij} - N_{ij}\right| - \frac{1}{2}\right)^2}{N_{ij}}, \qquad (2.7)$$

где

$$N_{ij} = \frac{n_{i.}\, n_{.j}}{n_{..}} \qquad (2.8)$$

— частота, которая должна была бы находиться в i-й строке таблицы на j-м месте, если бы была верна гипотеза независимости (для метода выбора I, см. задачу 2.4) или в случае справедливости гипотезы о равенстве истинных вероятностей (для методов выбора II и III, см. задачу 2.5). Если маргинальные частоты малы в том смысле, что одно или несколько значений N_{ij} меньше 5, то проверять значимость с помощью хи-квадрат критерия (или эквивалентного ему критерия, использующего нормальное распределение) не рекомендуется.

Альтернативная процедура, предложенная Фишером [Fisher, 1934] и Ирвином [Irwin, 1935], основана на рассмотрении лишь тех четырехклеточных таблиц, в которых маргинальные частоты $n_{1.}$, $n_{2.}$, $n_{.1}$ и $n_{.2}$ фиксированы и имеют наблюденные значения. При этом условии точные вероятности получения в ячейках таблицы разных наборов частот n_{11}, n_{12}, n_{21} и n_{22} могут быть вычислены по формуле гипергеометрического распределения

$$\Pr\{n_{11}, n_{12}, n_{21}, n_{22}\} = \frac{n_{1.}!\, n_{2.}!\, n_{.1}!\, n_{.2}!}{n_{..}!\, n_{11}!\, n_{12}!\, n_{21}!\, n_{22}!}, \qquad (2.9)$$

где $n! = n(n-1)\ldots 3\cdot 2\cdot 1$. По соглашению $0! = 1$.

«Точный» тест Фишера—Ирвина состоит в определении вероятности для фактически наблюденной четырехклеточной таблицы, скажем $P_{\text{набл}}$, по формуле (2.9), также как и для всех других таблиц, имеющих такие же маргинальные частоты. При этом важны лишь те вероятности, которые меньше или равны $P_{\text{набл}}$.

Если сумма всех таких вероятностей не больше заданного уровня значимости, гипотеза (см. выше) отвергается; в противном случае — не отвергается.

Рассмотрим гипотетические данные табл. 2.8. Точная ве-

роятность, вычисленная по таблице с использованием формулы (2.9), равна

$$P_{\text{набл}} = \frac{5!\ 4!\ 6!\ 3!}{9!\ 2!\ 3!\ 4!\ 0!} = 0{,}1190. \qquad (2.10)$$

Таблица 2.8

Гипотетические данные, представляющие малые маргинальные частоты

	B	$\overline{B}$	Всего
A	2	3	5
$\overline{A}$	4	0	4
Всего	6	3	9

Три другие возможные таблицы с такими же маргинальными частотами, как в табл. 2.8, вместе с соответствующими вероятностями представлены в табл. 2.9.

Таблица 2.9

Другие возможные четырехклеточные таблицы с такими же маргинальными частотами, как в табл. 2.8

Таблица		Соответствующие вероятности реализации таблиц
3	2	0,4762
3	1	
4	1	0,3571
2	2	
5	0	0,0476
1	3	

Только для последней из этих таблиц значение вероятности не превышает $P_{\text{набл}}=0{,}1190$; таким образом, точный уровень значимости, соответствующий наблюденным данным, равен $0{,}1190+0{,}0476=0{,}1666$. Статистика χ^2 для данных табл. 2.8 имеет значение 1,41. Вероятность получить такое же или большее значение равна 0,23, что несколько отличается от точной вероятности 0,17.

Для функции распределения гипергеометрических вероятностей существуют достаточно подробные таблицы. Одной из наиболее доступных является табл. 38 из «Биометрических таблиц» Пирсона и Хартли [Pearson, Hartley, 1970]. Более обширные таблицы составлены Либерманом и Оуэном [Lieberman, Owen, 1961] и Финни, Латча, Беннета и Су [Finney, Latscha, Bennet, Hsu, 1963].

В дальнейших иллюстрационных примерах книги маргинальные частоты будут считаться достаточно большими и обеспечивающими обоснованное использование критериев хи-квадрат и критического отношения.

2.3. Поправка Иэйтса на непрерывность

Иэйтс [Yates, 1934] предложил включить в выражение (2.1) для χ^2 поправку

$$C_1 = -\frac{1}{2} n_{..} \tag{2.11}$$

и поправку

$$c_2 = -\frac{1}{2}\left(\frac{1}{n_{1.}} + \frac{1}{n_{2.}}\right) \tag{2.12}$$

в выражение (2.5) для z. Эти поправки вводятся ввиду того, что непрерывные распределения (хи-квадрат и нормальное соответственно) использованы для представления дискретного распределения выборочных частот.

Изучение эффекта введения поправки на непрерывность проводили Пирсон [Pearson, 1947], Маут, Пейвит и Андерсон [Mote, Pavate, Anderson, 1958] и Плакетт [Plackett, 1964]. Основываясь на этих работах и полученных ими результатах, Гризл [Grizzle, 1967] и Коновер [Conover, 1968, 1974] рекомендовали не использовать поправку на непрерывность. В качестве обоснования такой рекомендации они указали на очевидное занижение фактического уровня значимости при использовании коррекции. Заниженный уровень значимости приводит к уменьшению мощности, т. е. вероятность обнаружения существующей зависимости или существующего различия в частотах убывает.

Мантел и Гринхаус [Mantel, Greenhouse, 1968] указали на недостатки в анализе Гризла, а заодно и Коновера и опровергли их аргументацию против использования поправки. Мы не ставим перед собой цель подробно изложить опровержение, проведенное Мантелом и Гринхаусом, приведем только краткое изложение их доводов.

При применении метода выбора I исследователь предполагает, что верна гипотеза отсутствия зависимости между факторами A и B, которая означает, что все четыре вероятности в ячейках таблицы есть функции маргинальных пропорций $P_{1.}$, $P_{2.}$, $P_{.1}$ и $P_{.2}$ (см. табл. 2.3). Так как исследователь почти никогда не знает истинных величин этих пропорций, он должен осуществить их оценку с помощью полученных маргинальных частот.

При применении методов выбора II и III исследователь, кроме того, предполагает, что различия между двумя независимыми пропорциями, P_1 и P_2, отсутствуют. Так как почти никогда невозможно определить, какова истинная величина общей пропорции, ему приходится использовать полученные маргинальные частоты для ее оценки.

Следовательно, при применении любого из методов выбора исследователь при анализе данных ограничен тем, что его выводы могут основываться лишь на знании выборочных маргинальных пропорций, а не на истинных пропорциях в генеральной совокупности. Это ограничение эквивалентно признанию четырех полученных маргинальных частот (см. табл. 2.1) фиксированными величинами. Как было отмечено в разд. 2.2, точные вероятности, соответствующие набору наблюденных частот в ячейках, при условии, что маргинальные частоты фиксированы, можно вычислить по формуле гипергеометрического распределения. Так как включение в формулы поправки на непрерывность приводит к большему соответствию между вероятностями, чем в случае, когда поправка не используется, то ее рекомендуется применять всегда [1].

2.4. Выбор: односторонний или двусторонний критерий

Рассмотренные до сих пор критерий хи-квадрат и эквивалентный ему тест, использующий нормальное распределение, являются примерами двусторонних критериев. Подробнее говоря, значимое различие устанавливается, если верно хотя бы одно соотношение или p_2 существенно больше, чем p_1, или p_2 существенно меньше, чем p_1. Предположим, что исследователя в действительности интересует проверка другой гипотезы, утверждающей существование различия только в одном направлении, например, что P_2, истинная пропорция в группе 2, больше, чем P_1, истинная пропорция в группе 1. Мощность такого сравнения можно увеличить, применив односторонний критерий.

[1] Дополнительные аргументы к вводу поправок даны в книге: Тьюки Дж., Мостеллер Ф. Анализ данных и регрессия. Вып. 1.— М.: Финансы и статистика, 1982 (см. например, разд. 5.9).— *Примеч. ред.*

Исследователь после применения одностороннего теста может сделать один из двух выводов: или p_2 существенно больше, чем p_1, или нет; возможная при последнем выводе ситуация, что p_1 значительно больше, чем p_2, исключается как неинтересующая.

Одностороннее тестирование начинается с изучения данных с целью проверки, согласуются ли они с предположением односторонней гипотезы. Если это не так, т. е. $p_1 > p_2$, а исследователь был заинтересован только в обнаружении обратного неравенства, то больше вычислений не производится и делается вывод, что P_2 не может быть больше, чем P_1. Если же данные не противоречат непосредственно проверяемой альтернативной гипотезе, то исследователь переходит либо к вычислению статистики χ^2 (2.1), либо статистики z (2.5).

Величина χ^2 проверяется на значимость следующим образом. Если исследователь желает иметь уровень значимости α, он находит в табл. А.1 (см. приложение) значение в колонке для 2α. Если вычисленная величина χ^2 превышает табличное критическое значение, исследователь делает вывод, что истинные вероятности отличаются в направлении, соответствующем утверждению альтернативной гипотезы (в нашем примере, что $P_2 > P_1$). Если же нет, он заключает, что истинные пропорции не различаются в этом направлении. Значимость величины z проверяется точно таким же образом: когда желаемый уровень значимости равен α, надо использовать табличное значение, соответствующее 2α.

Из табл. А.1 и А.2 (см. приложение) видно, что критические величины для уровня значимости 2α меньше, чем критические значения для уровня значимости α. Полученное значение для тестовой статистики (χ^2 или z) недостаточное, чтобы превысить критическое значение для доверительного уровня α может, тем не менее, превысить критическую величину для доверительного уровня 2α. С помощью одностороннего критерия легче отвергнуть гипотезу отсутствия различия, чем с помощью двустороннего, когда пропорции отличаются в утверждаемом односторонней гипотезой направлении; поэтому первый тест более мощен, чем второй.

Из приведенных рассуждений следует, что односторонний тест предусмотрен только для тех ситуаций, когда исследователя не интересует различие в направлении, противоположном тому, что утверждает гипотеза. Например, если гипотеза состоит в том, что $P_2 > P_1$, тогда неразличимы возможности $P_2 = P_1$ или $P_2 < P_1$. Такие случаи встречаются редко. Примером такой ситуации, когда используется односторонний критерий, может быть сравнение частоты случаев результативности нового способа лечения (p_2) с частотой случаев результатив-

ности стандартного способа лечения (p_1), и новый способ на практике заменит стандартный, если только p_2 значимо больше, чем p_1. При этом не имеет значения, равны ли способы лечения по эффективности или новое лечение в действительности хуже, чем стандартное; в том и другом случае исследователь будет придерживаться стандартного лечения.

Если, однако, исследователь намерен сообщить результаты коллегам по профессии, он из этических соображений обязан провести проверки по двустороннему критерию. Если результаты указывают, что новое лечение в действительности хуже, чем стандартное, — вывод, который возможен только при использовании двустороннего критерия, — исследователь обязан сообщить об этом с целью предупреждения тех, кто планирует изучение нового способа лечения.

В подавляющем большинстве исследований применяется именно двусторонний критерий. Даже если теория или большое количество опубликованных данных наводит на мысль, что интересующее различие должно бы быть в одном направлении, а не в другом, исследователь тем не менее должен подстраховаться от возможных неожиданных результатов, выполнив двусторонний тест. Научная важность различия, обнаруженного в неожиданном направлении, может быть значительно выше, чем еще одно подтверждение различия, имеющегося в предполагаемом направлении.

2.5. Простой доверительный интервал для разности между двумя независимыми пропорциями

Когда допускается, что пропорции P_1 и P_2 могут быть неравными, хорошей оценкой стандартной ошибки разности $p_2 - p_1$ является следующая:

$$\text{s. e. } (p_2 - p_1) = \sqrt{\frac{p_1 q_1}{n_{1.}} + \frac{p_2 q_2}{n_{2.}}}, \qquad (2.13)$$

где $q_1 = 1 - p_1$ и $q_2 = 1 - p_2$. Предположим, что и $n_{1.}$, и $n_{2.}$ велики в том смысле, что $n_{i.} \cdot p_i \geqslant 5$ и $n_{i.} q_i \geqslant 5$ для $i = 1, 2$ и что требуется построить 100 $(1-\alpha)$%-ный доверительный интервал для разности $P_2 - P_1$. Пусть $c_{\alpha/2}$ обозначает величину, отсекающую долю вероятности $\alpha/2$ от верхнего хвоста стандартной нормальной кривой ($1 - \alpha/2$ — квантиль). Интервал

$$(p_2 - p_1) - c_{\alpha/2} \sqrt{\frac{p_1 q_1}{n_{1.}} + \frac{p_2 q_2}{n_{2.}}} - \frac{1}{2}\left(\frac{1}{n_{1.}} + \frac{1}{n_{2.}}\right) \leqslant$$

$$\leqslant P_2 - P_1 \leqslant (p_2 - p_1) + c_{\alpha/2} \sqrt{\frac{p_1 q_1}{n_{1.}} + \frac{p_2 q_2}{n_{2.}}} + \\ + \frac{1}{2}\left(\frac{1}{n_{1.}} + \frac{1}{n_{2.}}\right) \qquad (2.14)$$

таков, что будет включать истинную разность пропорций приблизительно в 100 $(1-\alpha)$ % случаев.

Рассмотрим, например, данные из табл. 2.7. Выборочная разность есть $p_2 - p_1 = 0{,}10$, и оценка ее стандартной ошибки

$$\text{s. e. } (p_2 - p_1) = \sqrt{\frac{0{,}10 \cdot 0{,}90}{150} + \frac{0{,}20 \cdot 0{,}80}{50}} = 0{,}062.$$

Приближенный 95%-ный доверительный интервал для истинной разности равен.

$$0{,}10 - 1{,}96 \cdot 0{,}062 - 0{,}013 \leqslant P_2 - P_1 \leqslant 0{,}10 + \\ + 1{,}96 \cdot 0{,}062 + 0{,}013$$

или

$$-0{,}035 \leqslant P_2 - P_1 \leqslant 0{,}235.$$

Интервал включает значение 0, что согласуется с полученным ранее отрицательным результатом (см. уравнение (2.6)) при попытке найти значимое различие между p_1 и p_2.

2.6. Альтернатива тесту критического отношения

В некоторых случаях только что отмеченной согласованности теста значимости различия между p_1 и p_2 и доверительного интервала для $P_2 - P_1$ не будет. Например, тест критического отношения (2.5) может не отвергнуть гипотезу $P_1 = P_2$, а доверительный интервал (2.14) тем не менее не будет содержать 0. Частично для того, чтобы исключить возможность несогласованности. Эберхардт и Флигнер [Eberhardt, Fligner, 1977] и Роббинс [Robbins, 1977] рассмотрели альтернативный тест критического отношения, в котором знаменатель статистики (2.5) заменен на (2.13). Тогда статистику критерия получаем в виде

$$z' = \frac{|p_2 - p_1| - \frac{1}{2}(1/n_{1.} + 1/n_{2.})}{\sqrt{\frac{p_1 q_1}{n_{1.}} + \frac{p_2 q_2}{n_{2.}}}}. \qquad (2.15)$$

Эберхардт и Флигнер (1977) сравнили критерий, основанный на критическом отношении в (2.5), и критерий, основан-

ный на статистике (2.15), хотя они не включали поправку на непрерывность. При $n_{1.}=n_{2.}$ они нашли, что тест, основанный на z', всегда более мощен, чем тест, основанный на z, однако доля случаев, когда он отвергает верную гипотезу, больше, чем номинальная пропорция, равная α. Когда $n_{1.}\neq n_{2.}$, существуют такие пары пропорций P_1 и P_2, для которых тест, основанный на z', имеет большую мощность, и другие пары, для которых большую мощность имеет тест, основанный на z.

Рассмотрим снова данные табл. 2.7. Значение z' равно:

$$z'=\frac{|0{,}20-0{,}10|-\frac{1}{2}\left(\frac{1}{150}+\frac{1}{50}\right)}{\sqrt{\frac{0{,}10\cdot 0{,}90}{150}+\frac{0{,}20\cdot 0{,}80}{50}}}=1{,}41,$$

что меньше, чем величина z в (2.6). Для этих гипотетических данных оказывается, что классический тест, использующий z (см. (2.5)), дает больше оснований к отвержению гипотезы, чем тест, основанный на z' (см. 2.15)).

Дальнейший анализ показывает, что в тех случаях, когда тест, основанный на z', более мощен, чем тест, основанный на z, увеличение мощности незначительно, исключая ситуацию, когда разность между P_1 и P_2 велика (в том смысле, что отношения шансов больше, чем 10; (см. уравнение 3.1 для определения отношения шансов). Следовательно, нет убедительных причин для замены хорошо известного теста, основанного на z, и эквивалентного ему классического теста хи-квадрат, на тест, основанный на z'.

Задачи

2.1. Рассмотрите совместные пропорции в табл. 2.4. Докажите, что $p_{12}=p_{1.}p_{.2}=-(p_{11}-p_{1.}p_{.1})$, что $p_{21}-(p_{2.}p_{.1})=-(p_{11}-p_{1.}p_{.1})$ и что $p_{22}-p_{2.}p_{.2}=p_{11}-p_{1.}p_{.1}$. (Указание. Так как $p_{11}+p_{12}=p_{1.}$, следовательно, $p_{12}=p_{1.}-p_{11}$. Используйте тот факт, что $1-p_{.2}=p_{.1}$.)

2.2. Докажите, что формулы (2.3) и (2.1) для χ^2 эквивалентны. (Указание. Сначала используйте результат задачи 2.1 для множителя $(|p_{11}-p_{1.}p_{.1}|-1/(2n_{..}))^2$, а затем просуммируйте (формула (2,3).) Приведите четыре оставшихся члена, $1/(p_{i.}p_{.j})$, к общему знаменателю и покажите, что числитель результирующего выражения равен единице при $p_{1.}+p_{2.}=p_{.1}+p_{.2}=1$. Наконец, замените каждую пропорцию соответствующим отношением частот.)

2.3. Докажите, что квадрат z (см. 2.5) равен выражению χ^2, заданному формулой (2.1).

2.4. Покажите, что ожидаемое значение в i-й строке на j-м месте четырехклеточной таблицы для метода выбора I задается выражением (2.8) при гипотезе независимости. (Указание. Ожидаемое значение равно $n_{..}P_{ij}$. Что является оценкой P_{ij} при гипотезе независимости?)

2.5. Покажите, что ожидаемое значение в i-й строке на j-м месте четырехклеточной таблицы, порожденное методами выбора II или III, задается

выражением (2.8) при гипотезе о равенстве истинных вероятностей. (Указание. При проверке гипотезы $P_1=P_2$; обозначим $P_1=P_2=P$. Ожидаемые значения равны $N_{i1}=n_{i.}P$ и $N_{i2}=n_{i.}Q$, где $Q=1-P$. Что будет являться оценками для P и Q при заданной гипотезе?)

2.6. Когда размеры двух выборок равны, знаменатель (2.5) включает число $2\bar{p}\bar{q}$, в то время как знаменатель (2.15) — число $p_1q_1+p_2q_2$. Докажите, что $p_1q_1+p_2q_2 \leqslant 2\bar{p}\bar{q}$, когда $n_{1.}=n_{2.}$, причем равенство выполняется тогда и только тогда, когда $p_1=p_2$.

ЛИТЕРАТУРА

Barnard, G. A. (1947). Significance tests for 2 × 2 tables. *Biometrika*, **34**, 123–138.

Conover, W. J. (1968). Uses and abuses of the continuity correction. *Biometrics*, **24**, 1028.

Conover, W. J. (1974). Some reasons for not using the Yates continuity correction on 2 × 2 contingency tables. (With comments). *J. Am. Stat. Assoc.*, **69**, 374–382.

Eberhardt, K. R. and Fligner, M. A. (1977). A comparison of two tests for equality of two proportions. *Am. Stat.*, **31**, 151–155.

Finney, D. J., Latscha, R., Bennett, B. M., and Hsu, P. (1963). *Tables for testing significance in a 2 × 2 contingency table*. Cambridge, England: Cambridge University Press.

Fisher, R. A. (1934). *Statistical methods for research workers*, 5th ed. Edinburgh: Oliver and Boyd.

Русский перевод: Фишер Р.
Статистические методы для исследователей. – М.: Госстатгиз, 1958.

Grizzle, J. E. (1967). Continuity correction in the χ^2-test for 2 × 2 tables. *Am. Stat.*, 21 (October), 28–32.

Irwin, J. O. (1935). Tests of significance for differences between percentages based on small numbers. *Metron*, **12**, 83–94.

Lieberman, G. J. and Owen, D. B. (1961). *Tables of the hypergeometric probability distribution*. Stanford: Stanford University Press.

Mantel, N. and Greenhouse, S. W. (1968). What is the continuity correction? *Am. Stat.*, 22 (December), 27–30.

Mote, V. L., Pavate, M. V., and Anderson, R. L. (1958). Some studies in the analysis of categorical data. *Biometrics*, **14**, 572–573.

Pearson, E. S. (1947). The choice of statistical tests illustrated on the interpretation of data classed in a 2 × 2 table. *Biometrika*, **34**, 139–167.

Pearson, E. S. and Hartley, H. O. (Eds.) (1970). *Biometrika tables for statisticians*, Vol. 1, 3rd ed. Cambridge, England: Cambridge University Press.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos. Mag.*, **50**(5), 157–175.

Plackett, R. L. (1964). The continuity correction in 2 × 2 tables. *Biometrika*, **51**, 327–337.

Robbins, H. (1977). A fundamental question of practical statistics. *Am. Stat.*, **31**, 97.

Yates, F. (1934). Contingency tables involving small numbers and the χ^2 test. *J. R. Stat. Soc. Suppl.*, **1**, 217–235.

Zubin, J. (1939). Nomographs for determining the significance of the differences between the frequencies of events in two contrasted series or groups. *J. Am. Stat. Assoc.*, **34** 539–544.

Глава 3

Определение размера выборки, необходимого для обнаружения различия между двумя пропорциями

Существуют два типа ошибок, против которых следует применять меры предосторожности при планировании сравнительных исследований. Хотя такие ошибки могут встретиться в любой статистической процедуре, здесь их обсуждение ограничено случаем, когда проводится сопоставление пропорций для двух независимых выборок, т. е. выборок, полученных методами выбора II и III. Два типа ошибок при использовании метода выбора I рассматриваются в книге Коэна [Cohen, 1977, гл. 7].

Первая ошибка, называемая ошибкой первого рода, состоит в установлении различия между изучаемыми пропорциями, когда на самом деле их разность равна нулю. Этому типу ошибок уделяется больше внимания в элементарных учебниках статистики и, следовательно, в реальных исследованиях. Обычно их избегают просто за счет выбора достаточно малого уровня значимости α применяемого статистического критерия, например, равного 0,01 или 0,05.

Этот способ контроля не всегда является подходящим, потому что ошибка первого рода в чистом виде на практике почти никогда не встречается. Это связано с тем, что две популяции, из которых были извлечены выборки, неизбежно несколько отличаются друг от друга, хотя, может быть, и очень незначительно. Такова ситуация и в случае сравнения эффективности двух методов лечения на основе доли больных, почувствовавших улучшение, и в случае определения влияния возможного фактора риска на вероятность заболевания. В задаче 3.2 показано, что не имеет значения, сколь мала разность между двумя истинными пропорциями — при условии, что она не равна нулю, так как при помощи выборки достаточно большого размера всегда можно выявить различие на требуемом уровне значимости. В предположении, что исследователь желает найти практически важное различие, а не только установить факт существования сколько угодно малого разли-

чия, он должен дополнительно позаботиться о том, чтобы размер выборки не был бы больше, чем ему требуется. Это нужно, чтобы избежать ошибки второго рода.

Вторая ошибка, называемая ошибкой второго рода, заключается в утверждении равенства двух пропорций, в то время как в действительности они различны. Как только что было отмечено, такая ошибка может не приниматься во внимание, когда пропорции различаются пренебрежимо мало. Она важна только тогда, когда пропорции существенно различны с практической точки зрения. Контроль ошибок второго рода должен, следовательно, включать указание определяемого исследователем уровня различия, который в действительности важно обнаружить, и задание им желаемой вероятности фактического обнаружения данного различия. Эта вероятность, обозначаемая $1-\beta$, называется мощностью критерия; величина β есть вероятность потерпеть неудачу при попытке установить, что разность статистически значима.

Некоторые способы установления, является ли разность между пропорциями существенной, даны в разд. 3.1. Указав величины α, $1-\beta$ и минимальную разность пропорций, которая еще считается важной, исследователь может использовать результаты разд. 3.2 или данные табл. А.3, приведенной в приложении (о том, как пользоваться этой таблицей, сказано в разд. 3.3), чтобы найти размеры выборок, гарантирующие выполнение двух условий. Во-первых, уменьшение размера выборки повлечет уменьшение $1-\beta$, вероятности обнаружения указанного различия, и, во-вторых, любое достаточно большое увеличение размера выборки может привести к тому, что вероятность признания значимой несущественно малой разности будет намного превышать α.

Часто исследователь вынужден работать с выборкой, размер которой продиктован финансовыми или временными ограничениями. Он и в этом случае может использовать табл. А.3, так как с ее помощью определяется величина той разности, которую еще можно обнаружить при разумных значениях допустимой ошибки и, таким образом, получить реалистическую оценку шансов на успех.

Разд. 3.4 посвящен случаю, когда заранее запланированы неравные размеры выборок. Некоторые дополнительные вопросы рассматриваются в разд. 3.5.

3.1. Назначение значимого порога разности

Исследователю, как правило, необходимо иметь некоторые априорные представления о величине пропорций, которые он изучает. Это знание можно получить в результате предшест-

вующего исследования, накопления клинического опыта, из предварительной экспериментальной работы на малой выборке или из статистических отчетов. Если имеется хоть какая-нибудь информация, исследователь может, используя свою интуицию и опыт, получить оценку разности между двумя пропорциями, которая важна как в научном, так и в практическом отношении. При отсутствии информации исследователь не имеет базы для обоснованного планирования эксперимента.

В этой главе рассматриваются только два способа (из многих существующих) определения минимальной значимой величины разности пропорций. Каждый из этих способов проиллюстрирован двумя примерами. Пусть P_1 обозначает пропорцию членов первой группы, которые обладают изучаемым свойством или имеют интересующий исследователя результат. Вообще говоря, разделение группы на две подгруппы произвольно. Здесь, однако, мы выделяем в качестве первой группу, которую можно рассматривать как стандартную или типичную, так как о ней может быть известно больше, чем о другой группе. Наша задача состоит в том, чтобы определить величину пропорции P_2 для второй группы, которая, если удалось бы найти ее истинное значение, могла бы считаться столь существенно отличающейся с практической точки зрения от P_1, что гарантировала бы вывод о различии групп.

Пример 1. Проводится сравнение двух способов клинического лечения. Первая группа объединяет больных, к которым применяется обычная терапия. Пропорция P_1 относится к наблюдаемому отклику (эффекту лечения), например ремиссии, которая наступает через определенный промежуток времени после начала лечения. Вторую группу могут представлять пациенты, к которым применялся непроверенный пока (альтернативный) метод лечения. Важная в клиническом отношении пропорция P_2, связанная с альтернативным лечением, может быть определена следующим образом.

Предположим, что все пациенты, дающие положительный результат при стандартном лечении, будут таким же образом реагировать на новую терапию. Далее предположим, что если по крайней мере заданная исследователем добавочная часть f больных без положительного эффекта при стандартном лечении дает положительный результат при новом лечении, то исследователь может посчитать правдоподобным вывод, что новое лечение лучше старого. Так как пропорция неоткликнувшихся на стандартное лечение равна $1-P_1$, величина P_2 есть, соответственно, $P_1+f(1-P_1)$.

Допустим, например, что частота ремиссии, связанная со стандартным лечением, равна $P_1=0{,}60$. Если исследователь будет считать, что альтернативное лечение лучше стандарт-

ного, если оно сможет уменьшить симптомы по крайней мере у одной четверти тех пациентов, у которых при стандартном лечении не проявляется ремиссии, т. е. $f=0{,}25$, тогда он определяет значение $P_2=0{,}60+0{,}25\cdot(1-0{,}60)=0{,}70$, т. е. как величину, которая с практической точки зрения существенно отличается от $P_1=0{,}60$.

В этом примере пропорции P_1 и P_2 относятся к благополучному исходу, а именно к ремиссии симптомов. Подобные рассуждения могут быть применены и к исследованиям, в которых рассматривается неблагополучный исход (т. е. заболевание или смерть).

Пример 2. Предположим, что частота преждевременных родов среди женщин определенного возраста и расы, проходивших профилактические осмотры во время беременности в специальной клинике, есть P_1. Интенсивный курс предродового поведения и подготовки к родам женщин, не посещавших клинику, разумно проводить только в том случае, если частота преждевременных родов P_2 среди будущих матерей, не посещавших клинику, но во всем остальном похожих на посещавших ее, на некую величину больше, чем P_1.

Резонно ожидать, что мать, находившаяся под наблюдением врача клиники, у которой ребенок появился преждевременно, родила бы до срока и в том случае, если бы она не посещала клинику. Дополнительный риск преждевременных родов, связанных с непосещением клиники, может, таким образом, существовать только для тех матерей, которые не родили ребенка до срока и находились под наблюдением клиники. Если f обозначает практически существенную долю дополнительного риска, то тогда ожидаемая значимой величина P_2 равна $P_1+f(1-P_1)$.

Пусть, например, частота преждевременных родов среди посещавших клинику матерей есть $P_1=0{,}25$. Предположим далее, что курс обучения проводится только в том случае, если из женщин, которые посещали клинику и которые не родили раньше времени, по крайней мере 20% имели бы преждевременные роды в случае непосещения клиники. Тогда f равняется 0,20, и ожидаемая значимой величина $P_2=0{,}25+0{,}20\cdot(1-0{,}25)=0{,}40$.

Метод сравнения двух пропорций, проиллюстрированный приведенными примерами, был рекомендован Шепс [Sheps, 1958, 1959, 1961]. Он снова будет рассмотрен в гл. 7, где более детально изучается относительная разность $f=(P_2-P_1)/(1-P_1)$.

Пример 3. Часто исследователь предпринимает изучение для того, чтобы повторить (или опровергнуть) выводы, полученные другими исследователями, или чтобы посмотреть, сох-

раняются ли его собственные ранее полученные результаты в новых условиях. Однако необходима аккуратность анализа, так как в сравниваемых группах при новых условиях могут быть другие частоты. Это фактически исключает возможность получения в новом исследовании найденного ранее значения для простой разности пропорций.

Для примера предположим, что частота депрессии среди женщин в возрасте 20—49 лет, находящихся в психиатрической лечебнице определенного региона, на 40% выше, чем частота поражения депрессией мужчин того же возраста. Такую же разность пропорций в психиатрических лечебницах других регионов нельзя будет обнаружить, если, например, частота депрессии в них для мужчин возраста 20—49 лет равна 70%, и, следовательно, разность в 40% влечет невозможную пропорцию в 110% для женщин того же возраста.

Следовательно, требуется такая мера степени неравенства между двумя частотами, которая бы сохранялась при изменении уровня частот в новых условиях. Мерой, часто удовлетворяющей данному требованию на практике, является так называемое отношение шансов (odds ratio), которое обозначается ω. Отношение шансов обсуждается более детально в гл. 5 и 6. Здесь мы дадим только его определение.

Если P_1 — частота, с которой происходит событие в первой популяции, то соответствующий шанс события в первой популяции равен $\Omega_1 = P_1/Q_1$, где $Q_1 = 1 - P_1$. Аналогично шанс события во второй популяции есть $\Omega_2 = P_2/Q_2$. Отношение шансов есть просто отношение этих двух величин

$$\omega = \frac{\Omega_2}{\Omega_1} = \frac{P_2\ Q_1}{P_1\ Q_2}. \qquad (3.1)$$

Отношение шансов называют также перекрестным отношением (cross-product) [Fisher, 1962] и приближенным относительным риском [Cornfield, 1951]. Если $P_2 = P_1$, то $\omega = 1$. Если $P_2 < P_1$, то $\omega < 1$, если $P_2 > P_1$, то $\omega > 1$.

Предположим, что исследование проводится с целью повторить результаты предыдущего исследования, в котором отношение шансов было равно ω. Если в популяции, в которой проводится новое исследование, частота появления события в первой группе есть P_1 и если величина ω для отношения шансов предполагается такой же, как и в ранее исследованной популяции, то величина P_2 есть

$$P_2 = \frac{\omega\ P_1}{\omega\ P_1 + Q_1}. \qquad (3.2)$$

Допустим, что величина $\omega = 2{,}5$ была найдена ранее как отношение шанса иметь депрессию для женщин — пациентов психиатрической клиники возраста 20—49 лет к шансу для

мужчин — пациентов той же клиники и того же возраста. Если предполагается та же величина для отношения шансов и в психиатрической клинике нового региона, а частота депрессии среди пациентов-мужчин возраста 20—49 лет предполагается приближенно равной $P_1=0{,}70$, то частота среди пациентов-женщин того же возраста будет равна:

$$P_2=\frac{2{,}5\cdot 0{,}70}{2{,}5\cdot 0{,}70+0{,}30}=0{,}85.$$

Важное свойство отношения шансов, которое продемонстрировано в гл. 5 и 6, состоит в том, что в результате как проспективного, так и ретроспективного исследования должна получаться одна и та же величина. Этот факт может быть полезным, если исследователь хочет повторить эксперимент, но при этом изменить его план, скажем поменять ретроспективное изучение на проспективное.

Пример 4. В некотором школьном округе было проведено ретроспективное исследование случаев психологических отклонений среди школьников. Школьники с эмоциональными нарушениями (изучаемая группа), требующие внимания психологов, сравнивались с нормальными детьми (контрольная группа) по некоторым признакам, возможно, лежащим в основе данных нарушений. Пусть в ходе исследования было установлено, что четвертая часть детей изучаемой группы потеряли (в результате смерти, развода или раздельной жизни супругов) по крайней мере одного родителя в возрасте до 5 лет, в то время как в контрольной группе неполные семьи составляют только 1/10 часть. Отношение шансов, как это следует из (3.1), тогда равно:

$$\omega=\frac{0{,}25\cdot 0{,}90}{0{,}10\cdot 0{,}75}=3{,}0.$$

Запланируем изучение этой зависимости с помощью проведения проспективного обследования в другом школьном округе. Будем наблюдать на протяжении школьных лет за выборкой детей, начавших учебу в условиях, когда оба родителя живы и живут вместе (группа 1), и за выборкой детей, у которых при поступлении в школу была неполная семья (группа 2), и сравним затем для этих групп частоты развития эмоциональных нарушений.

Изучая школьную документацию в новом округе, исследователь может установить, что P_1, доля детей, начавших обучение, живя с обоими родителями, и у которых в дальнейшем развились эмоциональные нарушения, есть $P_1=0{,}05$. Если $\omega=3{,}0$, найденное при ретроспективном изучении, будет при-

мещено в новом школьном округе, исследователь может предположить, что (см. уравнение (3,2))

$$P_2 = \frac{3{,}0 \cdot 0{,}05}{3{,}0 \cdot 0{,}05 + 0{,}95} = 0{,}136.$$

Следовательно, приблизительно 15% детей, обнаруживших эмоциональные нарушения, в течение периода школьного обучения потеряли по крайней мере одного родителя в возрасте до 5 лет.

Только что продемонстрированные методы могут быть использованы для генерации гипотез в исследованиях, проводимых в относительно короткие сроки, но, вероятно, они не годятся в долгосрочных сравнительных исследованиях. Гальперин, Рогот, Гуриан и Эдерер [Halperin, Rogot, Gurian, Ederer, 1968] предложили модель и некоторые числовые результаты, когда сравнивались два терапевтических метода, рассчитанных на длительный срок, и ожидалось полное отсутствие или незначительное количество пропущенных наблюдений.

Если возможны пропуски, для выдвижения гипотез может быть применена модель Шорка и Ремингтона [Schork, Remington, 1967]. Если требуется сравнить больше двух методов лечения или если результат измеряется с помощью шкалы с более чем двумя категориями, следует воспользоваться результатами Лачина [Lachin, 1977].

3.2. Математическое определение размера выборки

В этом и следующем разделе мы предполагаем, что размеры выборок из двух сравниваемых совокупностей, n_1 и n_2, совпадают, скажем с n. Мы найдем такую величину для общего размера выборок n, что, во-первых, если в действительности нет различия между двумя истинными пропорциями, то величина ошибки признать их различными равна α, и, во-вторых, если в действительности пропорции есть P_1 и $P_2 \neq P_1$, то вероятность правильного вывода о различии двух выборок есть $1-\beta$. Так как в этом разделе устанавливаются математические результаты, на которых основаны только значения табл. А.3 (о том, как с ней работать, см. в разд. 3.3), то он не является необходимым для чтения следующих разделов.

Мы начнем с определения требуемого размера выборки в обеих группах, скажем n', при условии, что в статистику не включается поправка на непрерывность. С помощью n' как первого приближения мы затем получим формулу для нужного размера выборок, $\hat{n}$, соответствующего случаю включения в формулу статистики критерия поправки на непрерывность.

Предположим, что пропорции, найденные в двух выборках, равны p_1 и p_2. Статистика, которая используется для проверки значимости их разности при условии, что мы временно не применяем поправку на непрерывность, есть

$$z = \frac{p_2 - p_1}{\sqrt{2\bar{p}\,\bar{q}/n'}}, \tag{3.3}$$

где

$$\bar{p} = \frac{1}{2}(p_1 + p_2)$$

и

$$\bar{q} = 1 - \bar{p}.$$

Для того чтобы гарантировать, что вероятность ошибки первого рода есть α, разность между p_1 и p_2 должна считаться значимой, только если

$$|z| > c_{\alpha/2}, \tag{3.4}$$

где $c_{\alpha/2}$ обозначает величину, отсекающую долю $\alpha/2$ от верхнего хвоста стандартной нормальной кривой, и $|z|$ — абсолютная величина z, которая всегда неотрицательна. Например, если $\alpha = 0{,}05$, то $c_{0{,}05/2} = c_{0{,}025} = 1{,}96$, и разность признается значимой, если $z > 1{,}96$ или $z < -1{,}96$.

В том случае, когда разность между истинными пропорциями фактически равна $P_2 - P_1$, мы хотим, чтобы вероятность отклонения гипотезы о равенстве пропорций, имеющем место при выполнении неравенства (3.4), была равна $1-\beta$. Таким образом, мы должны найти величину n', такую, что когда $P_2 - P_1$ есть истинная разность между двумя пропорциями,

$$\Pr\left\{\frac{|p_2 - p_1|}{\sqrt{2\bar{p}\,\bar{q}/n'}} > c_{\alpha/2}\right\} = 1 - \beta. \tag{3.5}$$

Вероятность в (3.5) есть сумма двух вероятностей

$$1 - \beta = \Pr\left\{\frac{p_2 - p_1}{\sqrt{2\bar{p}\,\bar{q}/n'}} > c_{\alpha/2}\right\} + \Pr\left\{\frac{p_2 - p_1}{\sqrt{2\bar{p}\,\bar{q}/n'}} > -c_{\alpha/2}\right\}. \tag{3.6}$$

Если предполагается, что P_2 больше чем P_1, тогда второе слагаемое в правой стороне соотношения (3.6) (представляющее вероятность события, что p_2 заметно меньше, чем p_1) близка

к нулю (см. задачу 3.1). Таким образом, нам нужно только найти величину n', такую, что

$$1-\beta = \Pr\left\{ \frac{p_2 - p_1}{\sqrt{2\bar{p}\,\bar{q}/n'}} > c_{\alpha/2} \right\}, \tag{3.7}$$

где истинная разность есть $P_2 - P_1$.

Вероятность в (3.7) не может пока быть вычислена, потому что среднее и стандартная ошибка величины $p_2 - p_1$, оценивающей истинную разность $p_2 - p_1$, еще не принимались в расчет. Среднее для $p_2 - p_1$ есть $P_2 - P_1$ и стандартная ошибка разности есть

$$\text{s. e. } (p_2 - p_1) = \sqrt{(P_1 Q_1 + P_2 Q_2)/n'}, \tag{3.8}$$

где $Q_1 = 1 - P_1$ и $Q_2 = 1 - P_2$.

Следующий шаг состоит в выполнении простых алгебраических преобразований в формуле (3.7):

$$\begin{aligned} 1-\beta &= \Pr\left\{ (p_2 - p_1) > c_{\alpha/2}\sqrt{2\bar{p}\,\bar{q}/n'} \right\} = \\ &= \Pr\left\{ (p_2 - p_1) - (P_2 - P_1) > c_{\alpha/2}\sqrt{2\bar{p}\,\bar{q}/n'} - (P_2 - P_1) \right\} = \\ &= \Pr\left\{ \frac{(p_2 - p_1) - (P_2 - P_1)}{\sqrt{(P_1 Q_1 + P_2 Q_2)/n'}} > \frac{c_{\alpha/2}\sqrt{2\bar{p}\,\bar{q}/n'} - (P_2 - P_1)}{\sqrt{(P_1 Q_1 + P_2 Q_2)/n'}} \right\}. \end{aligned} \tag{3.9}$$

Окончательная вероятность в (3.9) может быть вычислена с использованием таблиц нормального распределения, потому что, когда истинные пропорции есть P_2 и P_1, распределение величины

$$Z = \frac{(p_2 - p_1) - (P_2 - P_1)}{\sqrt{(P_1 Q_1 + P_2 Q_2)/n'}}, \tag{3.10}$$

если n' велико, хорошо аппроксимируется стандартным нормальным распределением.

Пусть $c_{1-\beta}$ обозначает величину, отсекающую долю $1-\beta$ от верхнего хвоста и долю β — от нижнего хвоста стандартной нормальной кривой. Тогда по определению

$$1-\beta = \Pr\left\{ Z > c_{1-\beta} \right\}. \tag{3.11}$$

Сравнивая (3.11) с последней строчкой в выражении (3.9), находим, что величина n', которую мы ищем, удовлетворяет равенству

$$c_{1-\beta} = \frac{c_{\alpha/2}\sqrt{2\bar{p}\,\bar{q}/n'} - (P_2 - P_1)}{\sqrt{(P_1 Q_1 + P_2 Q_2)/n'}} =$$

$$= \frac{c_{\alpha/2}\sqrt{2\bar{p}\bar{q}} - (P_2 - P_1)\sqrt{n'}}{\sqrt{P_1 Q_1 + P_2 Q_2}}. \qquad (3.12)$$

Прежде чем получить окончательное выражение для n^1, заметим, что (3.12) является функцией не только от P_1 и P_2, которые могут быть установлены исследователем при задании гипотезы, но также и от $\bar{p}\bar{q}$, которое становится известным лишь после окончания исследования. Однако если n' достаточно велико, $\bar{p}$ будет близко к

$$\bar{P} = \frac{P_1 + P_2}{2} \qquad (3.13)$$

и, что более важно, $\bar{p}\bar{q}$ будет близко к $\bar{P}\bar{Q}$, где $\bar{Q} = 1 - \bar{P}$. Следовательно, заменяя $\sqrt{2\bar{p}\bar{q}}$ в (3.12) на $\sqrt{2\bar{P}\bar{Q}}$ и решая уравнение относительно n', находим

$$n' = \frac{\left(c_{\alpha/2}\sqrt{2\bar{P}\bar{Q}} - c_{1-\beta}\sqrt{P_1 Q_1 + P_2 Q_2}\right)^2}{(P_2 - P_1)^2}, \qquad (3.14)$$

т. е. требуемый размер выборки для каждой из двух сравниваемых популяций в случае, когда поправка на непрерывность не применяется.

Отрицательный знак перед $c_{1-\beta}$ в числителе (3.14) — не типографская ошибка. Вспомните определение c_p, данное ранее, а именно: c_p есть решение уравнения $\Pr(z > c_p) = p$, где z имеет стандартное нормальное распределение. Когда p превосходит 0,5, значение c_p будет отрицательным.

Хейзман [Haseman, 1978] установил, что уравнение (3.14) дает слишком малую величину размера выборки в том смысле, что мощность критерия, основанного на выборках размера $n_1 = n_2 = n'$ меньше, чем $1-\beta$, когда P_1 и P_2 — истинные вероятности. Крамер и Гринхаус [Kramer, Greenhouse, 1959] предложили улучшенный вариант формулы (3.14), при котором поправка на непрерывность применяется дважды: сначала в статистике (3.3), а затем в статистике (3.10). Их модификация, которая была табулирована в первом издании этой книги, как было обнаружено Касаграндом, Пайком и Смитом [Casagrande, Pike and Smith, 1978b], приводит к неоправданно большому влиянию коррекции.

Включая поправку на непрерывность только в статистику критерия (3.3), эти авторы предложили значение

$$n = \frac{n'}{4}\left[1 + \sqrt{1 + \frac{4}{n' \mid P_2 - P_1 \mid}}\right]^2 \qquad (3.15)$$

в качестве размера выборки, который обеспечивает вполне достаточную степень приближения для желаемых значений уровня значимости и мощности. Размеры выборок, представленные в табл. А.3, которая отличается от табл. А.3 в первом издании, основаны на этой формуле. Значения в этой таблице очень хорошо совпадают с теми, какие были табулированы Касаграндом, Пайком и Смитом [1978a], а также Хейсманом [1978]. В работе [Ury and Fleiss, 1980] проведены сравнения с некоторыми другими формулами.

С высокой степенью точности (особенно, когда n' и $|P_2-P_1|$ таковы, что $n'|P_2-P_1| \geqslant 4$)

$$n = n' + \frac{3}{|P_2 - P_1|}. \tag{3.16}$$

Этот результат, полученный Флейсом, Тайтоном и Ури [Fleiss, Tytun and Ury, 1980], полезен как для быстрой оценки требуемого размера выборки, так и для оценки мощности в исследовании, где размер выборки заранее фиксирован. Предположим, что в некоторой ситуации можно изучать не более чем $2n$ объектов. Если уровень значимости равен α и если две истинные пропорции, которые исследователь хочет сравнить, есть P_1 и P_2, можно разрешить (3.16) и (3.14) относительно $c_{1-\beta}$, получив

$$c_{1-\beta} = \frac{c_{\alpha/2}\sqrt{2\bar{P}\bar{Q}} - |P_2 - P_1|\sqrt{n - \frac{2}{|P_2-P_1|}}}{\sqrt{P_1 Q_1 + P_2 Q_2}}. \tag{3.17}$$

Это выражение можно рассматривать как уравнение, определяющее абсциссу нормальной кривой, соответствующую мощности для заданного размера выборок. Затем при помощи табл. А.2 можно найти значение и мощности.

3.3. Использование таблиц для определения размера выборки

Таблица А.3 содержит требуемые размеры для двух равных по объему выборок из двух групп при различных значениях предполагаемых пропорций P_1 и P_2, для различных уровней значимости (α=0,01, 0,02, 0,05, 0,10, 0,20) и для различных значений мощности [1—β=0,50, 0,65(0,05) 0,95, 0,99]. Значение 1—β=0,50 включено не столько потому, что исследователь станет намеренно проводить исследование, для которого вероятность успеха будет только 50:50, а сколько для того, чтобы помочь ему определить нижнюю границу для необходимого минимального размера выборки.

В первую очередь обычно задается вероятность α ошибки первого рода. Если на основе утверждения о значимости различия двух пропорций принимается решение проводить дальнейшие (возможно, дорогие) исследования или заменить традиционный вид лечения на новый, то ошибка первого рода играет важную роль, и значение α должно быть взято достаточно малым (скажем, 0,01 или 0,02). Если изучение направлено только на увеличение публикуемой информации, имеющей отношение к некоторой теории, то ошибка первого рода менее серьезна, и значение α может быть увеличено до 0,05 или 0,10 (чем больше опубликованных работ указывает на различие пропорций, тем большее значение α можно использовать без особого риска).

Задав значение α, исследователю нужно затем указать вероятность $1-\beta$ обнаружения различия пропорций, если считать, что в соответствующих совокупностях пропорции равны P_1 и P_2. Для этой цели предлагается применить правило, предложенное Коэном [Cohen, 1977, p. 56]. Оно предполагает, что типичным является случай, когда ошибка первого рода примерно в 4 раза более опасна, чем ошибка второго рода. Это означает, что следует задать β, вероятность ошибки второго рода, приближенно равной 4α, так что мощность получается приближенно равной $1-\beta=1-4\alpha$. Таким образом, когда $\alpha=0{,}01$, $1-\beta$ можно положить равным 0,95; для $\alpha=0{,}02$ положить $1-\beta=0{,}90$ и для $\alpha=0{,}05$ положить $1-\beta=0{,}80$. Когда α больше, чем 0,05, кажется безопасным взять $1-\beta=0{,}75$ или меньше. Использование табл. А.3 будет проиллюстрировано на каждом из примеров разд. 3.1.

Пример 1. Исследователь предполагает, что частоты ремиссии следующие: $P_1=0{,}60$ для стандартного метода и $P_2=0{,}70$ для нового метода лечения. Пусть уровень значимости α равен 0,01, а мощность $1-\beta$ равна 0,95. В этом случае необходимо изучить 827 пациентов, получивших стандартное лечение, и еще 827 — получивших новое лечение, причем распределение пациентов по группам должно быть случайным (рандомизированным) для того, чтобы обеспечить установленные уровни значимости и мощность.

При уменьшении вероятности обнаружения различия $(1-\beta)$ до 0,75 без возрастания уровня значимости необходимо изучить по 499 пациентов для каждого метода лечения. Если исследователь может позволить себе изучить не более чем 600 пациентов, так что каждое лечение будет применено не более чем к 300 пациентам, вероятность обнаружения предполагаемого различия становится меньше, чем 50 : 50.

Пример 2. Исследователь выдвигает гипотезу о частотах преждевременных родов: $P_1=0{,}25$ для наблюдавшихся в кли-

нике и $P_2=0{,}40$ для ненаблюдавшихся. Уровень значимости α задается равным 0,01, а мощность — $1-\beta=0{,}95$. При этих условиях необходимо изучить по 357 матерей из каждой группы, причем возраст всех этих женщин должен удовлетворять заданным ограничениям. Если уровень значимости повысить до $\alpha=0{,}02$, а мощность уменьшить до $1-\beta=0{,}95$, то необходимо будет обследовать в каждой группе по 265 матерей.

Пример 3. Исследователь предполагает, что частота депрессий среди пациентов-мужчин возраста 20—49 лет в некоторой психиатрической клинике $P_1=0{,}70$, а частота среди пациенток-женщин того же возраста — $P_2=0{,}85$. Пусть уровень значимости $\alpha=0{,}05$ и мощность $1-\beta=0{,}80$. Тогда необходимо изучить 134 пациента каждого пола. Если бы исследователь планировал изучить по 250 пациентов каждого пола, то вероятность обнаружить предполагаемое различие была бы выше 95%, что более, чем требуется.

Пример 4. Проверяемая гипотеза состоит в следующем: пропорция развития эмоциональных нарушений среди детей, начинающих обучение в школе, имея полную семью, есть $P_1=0{,}05$, пропорция детей, начинающих обучение в школе при отсутствии хотя бы одного родителя, P_2, приближенно равна 0,15. Уровень значимости α назначен равным 0,05, а мощность — $1-\beta=0{,}80$. Необходимо изучить по 160 детей каждой группы, контролируя при этом сходство двух групп по полу и расе. Если исследователь может включить в изучаемые группы не более чем по 120 детей, и если он желает увеличить вероятность появления ошибки первого рода до $\alpha=0{,}10$, тогда вероятность обнаружения различия по-прежнему будет превышать 75% при условии, что предполагаемые значения для P_1 и P_2 верны.

3.4. Неодинаковые размеры выборок

Предположим, что соображения стоимости проведения исследования, желание получить более точные оценки для одной из групп или какие-нибудь другие причины (см. статью Вальтера [Walter, 1977]) приводят к использованию выборок разного размера для изучаемых совокупностей. Обозначим требуемый размер выборки для первой совокупности через m и для второй совокупности — через rm $(0<r<\infty)$ при заданном заранее r. Размер общей выборки есть, скажем, $N=(r+1)m$.

Если p_1 и p_2 — две полученные выборочные пропорции, то статистикой критерия является

$$z=\frac{|p_2-p_1|-\frac{1}{2m}\left(\frac{r+1}{r}\right)}{\sqrt{\frac{\bar{p}\,\bar{q}\,(r+1)}{m\,r}}},$$

где $p=(p_1+rp_2)/(r+1)$ и $\bar{q}=1-\bar{p}$. Если нужный уровень значимости есть α и если желаемая мощность равна $1-\beta$ при альтернативной гипотезе $P_1 \neq P_2$ с заданными P_1 и P_2, то таким же образом, как и в разд. 3.2, приходим к значению

$$m=\frac{m'}{4}\left[1+\sqrt{\frac{2(r+1)}{1+m'r|P_2-P_1|}}\right]^2, \qquad (3.18)$$

определяющему требуемый размер выборки из первой совокупности и rm — из второй. В формуле (3.18)

$$m'=\frac{\left[c_{\alpha/2}\sqrt{(r+1)\bar{P}\bar{Q}}-c_{1-\beta}\sqrt{rP_1Q_1+P_2Q_2}\right]^2}{r(P_2-P_1)^2}, \qquad (3.19)$$

где $\bar{P}=(P_1+rP_2)/(r+1)$ и $\bar{Q}=1-\bar{P}$.

Как получено Флейсом, Тайтоном и Ури [Fleiss, Tytun and Ury, 1980], m приближенно равно:

$$m=m'+\frac{r+1}{r|P_2-P_1|}. \qquad (3.20)$$

Заметим, что уравнения (3.14)—(3.16) являются частным случаем приведенных выше уравнений, когда размеры двух выборок равны (т. е. когда $r=1$).

Рассмотрим снова пример сравнения частот преждевременных родов среди наблюдавшихся в клинике (P_1 предполагается равной 0,25) и среди непосещавших ее (P_2 по-прежнему равно 0,40). Пусть уровень значимости α снова есть 0,01, а мощность — $1-\beta=0{,}95$. Предположим, что получить данные о наблюдавшихся в клинике легче, чем о ненаблюдавшихся, и что поэтому исследователь решает изучать вдвое меньшее количество ненаблюдавшихся женщин по сравнению с количеством наблюдавшихся; таким образом, $r=0{,}5$. Найдем значение m' по формуле (3.19):

$$m'=\frac{\left[2{,}576\sqrt{1{,}5\cdot 0{,}30\cdot 0{,}70}-(-1{,}645)\sqrt{0{,}5\cdot 0{,}25\cdot 0{,}75+0{,}40\cdot 0{,}60}\right]^2}{0{,}5\,(0{,}15)^2}=$$

$$=510{,}34.$$

Таким образом, следуя (3.18), требуемый размер выборки из совокупности наблюдавшихся в клинике равен:

$$m=\frac{510{,}34}{4}\left(1+\sqrt{1+\frac{2(1{,}5)}{510{,}34\cdot 0{,}5\cdot 0{,}15}}\right)^2=530.$$

Уравнение (3.20) дает то же значение с точностью округления до ближайшего целого числа.

Требуемые размеры двух выборок есть, следовательно, $n_1=m=530$ и $n_2=0{,}5m=265$. Общее число женщин $N=795$,

что примерно на 80 больше, чем требовалось в случае выборок одинакового размера (см. пример 2 разд. 3.3).

Задача 3.5 посвящена различным применениям результатов этого раздела.

3.5. Некоторые дополнительные комментарии

Коэн [1977, гл. 6] приводит набор таблиц, а Фейгл [Feigl, 1978] предлагает несколько графических методов для определения размеров выборок, когда задаются те же параметры, какие рассматривались до сих пор. Так как критерий значимости, который они использовали, отличен от стандартного и не включает поправку на непрерывность, их размеры выборок слегка отличаются от приведенных в табл. А.3. Вообще говоря, таблицы Коэна и номограммы Фейгла следует использовать, если исследователь может выдвинуть некоторые гипотезы лишь о порядке величины разности между P_1 и P_2, но не об их значениях. Если же априорно удается оценить значения P_1 и P_2 по отдельности, предпочтение лучше отдать приведенным в книге таблицам.

До сих пор мы исходим из того, что сравнение двух пропорций будет проводиться с помощью двустороннего критерия (см. разд. 2.4). Если исследователь хочет применить односторонний тест, он может по-прежнему пользоваться табл. А.3, но при этом удваивать уровень значимости. Так, для одностороннего уровня значимости 0,01 нужно использовать табличное значение, соответствующее $\alpha=0{,}02$; для одностороннего уровня значимости 0,05 используется $\alpha=0{,}10$. При этом нет необходимости изменять величину $1-\beta$.

Задачи

3.1. Предположим, что $P_2>P_1$ и что n' — размер изучаемой выборки в каждой из двух групп. Пусть Z обозначает случайную величину, имеющую стандартное нормальное распределение.

а) Покажите, что вероятность события p_2 значительно меньше, чем p_1, величина

$$\Pr\left\{\frac{p_2-p_1}{\sqrt{2\,\bar{p}\,\bar{q}/n'}} > -c_{\alpha/2}\right\},$$

приближенно равна:

$$\Pi=\Pr\left\{Z<\frac{-c_{\alpha/2}\sqrt{2\bar{P}\bar{Q}}-(P_2-P_1)\sqrt{n'}}{\sqrt{P_1Q_1+P_2Q_2}}\right\}.$$

б) Если $P_2=P_1$, то $\Pi=\alpha/2$. В случае если $P_2>P_1$, покажите, почему $\Pi<\alpha/2$. (Указание. Докажите, что $\sqrt{P_1Q_1+P_2Q_2}<\sqrt{2\bar{P}\bar{Q}}$, если $P_2\neq P_1$. Следовательно, если $P_2>P_1$, то

$$\frac{-c_{\alpha/2}\sqrt{2\bar{P}\bar{Q}}-(P_2-P_1)\sqrt{n'}}{\sqrt{P_1Q_1+P_2Q_2}}<-c_{\alpha/2}-\frac{(P_2-P_1)\sqrt{n'}}{\sqrt{P_1Q_1+P_2Q_2}}<-c_{\alpha/2}\ .)$$

в) Π мало, даже если P_2 только немного больше, чем P_1, и даже если n' мало. Найдите величину Π, когда $P_1=0{,}10$, $P_2=0{,}11$, $n'=9$ и $\alpha=0{,}05$. Заметим, что вероятность, указаиная в табл. А.2, должиа быть умеиьшена вдвое.

3.2. При использовании обозначеннй и предположений из задачи 3.1 мощиость критерия, сравнивающего p_1 и p_2, приближенно равна:

$$1-\beta=\Pr\left\{Z>\frac{c_{\alpha/2}\sqrt{2\bar{P}\bar{Q}}-(P_2-P_1)\sqrt{n'}}{\sqrt{P_1Q_1+P_2Q_2}}\right\}.$$

а) Покажите, что $1-\beta$ приближается к единице, когда n' стремится к бесконечности, а α остается фиксированным. (Указание. Чему равна вероятиость события, состоящего в том, что стандартная нормальная случайная величина превышает -1? -2? -3? К какой величине приближается значение выражения, стоящего справа от знака равенства в формуле, приведеиной выше, когда n' увеличивается?)

б) Покажите, что $1-\beta$ уменьшается с уменьшением α при фиксированном n'.

3.3. Покажите, что когда статистика критерия включает поправку иа иепрерывность, уравнение (3.15) дает размер выборки, требуемый в каждой группе для того, чтобы достичь мощности $1-\beta$, когда истиниые вероятности есть P_1 и P_2. (Указание. Гипотеза, утверждающая, что $P_1=P_2$, отвергается, если

$$\frac{|p_2-p_1|-\dfrac{1}{n}}{\sqrt{2\bar{p}\bar{q}/n}}>c_{\alpha/2}\ .$$

Допустим, что $P_2>P_1$, и применим те же алгебраические преобразования, как и в разд. 3.2. В результате получим уравнение

$$\sqrt{n}-\frac{1}{\sqrt{n(P_2-P_1)}}=\sqrt{n'}$$

для n, где n' определено в (3.14). Решите приведенное выше квадратичное уравнение относительно $\sqrt{n}$ и затем найдите n, заметив, что один из двух корией отрицателеи и, следовательио, иеприемлем.)

3.4. Исследователь предполагает, что частота улучшений, связанных с плацебо (безвредным препаратом, прописываемым для успокоения больного), есть $P_1=0{,}45$ и что частота улучшений, связаиных с активным лекарством, есть $P_2=0{,}65$. Планируется осуществить односторонний тест.

а) Если уровень значимости $\alpha=0{,}01$ и мощиость $1-\beta=0{,}95$ заданы, сколь большие выборки равного объема он должеи изучить для каждого лекарства?

б) Каков должен быть размер выборок, если уровень значимости α снизить до 0,05, а мощность $1-\beta$ — до 0,80?

в) Чему равна мощность одностороннего критерия в случае, если исследователь может изучить только по 52 пациента из каждой группы и если нужный ему уровень значимости равен 0,05? (Указание. Так как мы рассматриваем односторонний тест, следует заменить $c_{\alpha/2}$ в уравнении (3.17) на 1.645.)

3.5. Сравнение двух популяций с предполагаемыми вероятностями $P_1=0{,}25$ и $P_2=0{,}40$ встречалось уже несколько раз в тексте.

а) Продолжая изучение этого примера, возьмите $\alpha=0{,}01$ и $1-\beta=0{,}95$ и заполните пропуски во втором, третьем и четвертом столбце следующей таблицы.

Отношение размеров выборок ($n_2/n_1=r$)	n_1	n_2	Требуемый общий размер двух выборок	Полная стоимость, доллары
0,5	530	265	795	8480
0,6	—	—	—	—
0,7	—	—	—	—
0,8	—	—	—	—
0,9	—	—	—	—
1	357	357	714	7854

б) Будем исходить из того, что средняя стоимость изучения членов группы 1 и 2 равна соответственно 10 и 12 дол. Найдите общую стоимость, соответствующую каждой величине r, приведенной в таблице. Для какого из этих отношений размеров выборок полная стоимость изучения минимальна?

в) Предположим, что исследователь может позволить себе потратить на изучение только 6240 дол. и решает использовать значение r, найденное в пункте (б). Чему равна величина m? Каково соответствующее значение m', когда $P_1=0{,}25$ и $P_2=0{,}40$. (Решите уравнение (3.20) для нахождения m'.) Чему равно соответствующее значение $c_{1-\beta}$ при $\alpha=0{,}01$? (Решите уравнение (3.19) для определения $c_{1-\beta}$.) Какова мощность критерия?

ЛИТЕРАТУРА

Casagrande, J. T., Pike, M. C., and Smith, P. G. (1978a). The power function of the "exact" test for comparing two binomial distributions. *Appl. Stat.*, 27, 176–180.

Casagrande, J. T., Pike, M. C., and Smith, P. G. (1978b). An improved approximate formula for calculating sample sizes for comparing two binomial distributions. *Biometrics*, 34, 483–486.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*, revised ed. New York: Academic Press.

Cornfield, J. (1951). A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast and cervix. *J. Natl. Cancer Inst.*, **11**, 1269–1275.

Feigl, P. (1978). A graphical aid for determining sample size when comparing two independent proportions. *Biometrics*, **34**, 111–122.

Fisher, R. A. (1962). Confidence limits for a cross-product ratio. *Aust. J. Stat.*, **4**, 41.

Fleiss, J. L., Tytun, A., and Ury, H. K. (1980). A simple approximation for calculating sample sizes for comparing independent proportions. *Biometrics*, **36**, 343–346.

Halperin, M., Rogot, E., Gurian, J., and Ederer, F. (1968). Sample sizes for medical trials with special reference to long-term therapy. *J. Chronic Dis*, **21**, 13–24.

Haseman, J. K. (1978). Exact sample sizes for use with the Fisher-Irwin test for 2 × 2 tables. *Biometrics*, **34**, 106–109.

Kramer, M. and Greenhouse, S. W. (1959). Determination of sample size and selection of cases. Pp. 356–371 in National Academy of Sciences–National Research Council Publication 583, *Psychopharmacology: Problems in evaluation*. Washington, D. C.

Lachin, J. M. (1977). Sample size determinations for $r \times c$ comparative trials. *Biometrics*, **33**, 315–324.

Schork, M. A. and Remington, R. D. (1967). The determination of sample size in treatmentcontrol comparisons for chronic disease studies in which drop out or nonadherence is a problem. *J. Chronic Dis.*, **20**, 233–239.

Sheps, M. C. (1958). Shall we count the living or the dead? *New Engl. J. Med.*, **259**, 1210–1214.

Sheps, M. C. (1959). An examination of some methods of comparing several rates or proportions. *Biometrics*, **15**, 87–97.

Sheps, M. C. (1961). Marriage and mortality. *Am. J. Public Health*, **51**, 547–555.

Ury, H. K. and Fleiss, J. L. (1980). On approximate sample sizes for comparing two independent proportions with the use of Yates' correction. *Biometrics*, **36**, 347–351.

Walter, S. D. (1977). Determination of significant relative risks and optimal sampling procedures in prospective and retrospective comparative studies of various sizes. *Am. J. Epidemiol.*, **105**, 387–397.

Глава 4
Как проводить рандомизацию

В предыдущих главах несколько раз упоминалось понятие «рандомизация». Мы считали, что объекты выборки «случайно отобраны» из большей группы объектов, или что объекты «случайно назначены» в группы, различающиеся обработкой (лечением), или выборка сформирована как-то иначе. В этой главе описаны некоторые методы рандомизации отбора объектов и назначения обработок (необходимые при извлечении выборок соответственно методами I, II, см. разд. 4.1 и методом III, см. разд. 4.2 и 4.3).

Важно понимать, что термин «рандомизация» здесь относится не к выборке, а к способу ее генерирования. Говоря, что группа данного размера является *простой случайной выборкой* из большей группы, мы подразумеваем, что все возможные выборки этого размера извлекаются с равными вероятностями. Говоря, что обработка назначается объектам *случайно*, мы подразумеваем, что вероятность назначения каждого вида обработки одинакова для всех объектов.

На необходимость рандомизации в планируемых экспериментах впервые было указано Фишером [Fisher, 1935]. О целях, которые преследуются при случайном назначении обработок объектам в сравнительных испытаниях, Хилл [Hill, 1962] писал: «[Рандомизация] обеспечивает три вещи: она гарантирует, что наши наклонности и предпочтения не повлияют на формирование групп с различными обработками; она предотвращает опасность, связанную с выбором на основе личных суждений, — считая, что наши суждения могут быть пристрастными, мы стараемся учесть и устранить пристрастность и при этом можем перестараться, ударяясь в другую крайность и приходя к противоположным смещениям; наконец, при случайном распределении обработок самый строгий критик не сможет сказать, что группы рассматривались по-разному вследствие наших предпочтений или нашей глупости».

Несмотря на широкое применение рандомизации в сравнительных испытаниях, споры о ее универсальности продолжаются. Эта тема будет рассмотрена в разд. 7.3.

В табл. А.4 приведены 20 000 случайных цифр. На каждой странице таблицы — 10 столбцов, в каждом столбце — 50 пятизначных целых чисел. Ниже даны примеры использования этой таблицы.

4.1. Формирование простой случайной выборки

Предположим, что мы намереваемся детально обследовать состояние здоровья 100 из 250 сотрудников, работающих в фирме. Простую случайную выборку объема 100 из большей группы объема 250 можно сформировать следующим образом.

Будем последовательно просматривать по табл. А.4 трехзначные числа, игнорируя 000 и числа от 251 до 999. Первые различные 100 из полученных чисел будут обозначать сотрудников, которых и нужно обследовать. Пронумеровав сотрудников по списку от 1 до 250 (порядок нумерации несущественен), мы получим требуемую выборку сотрудников в соответствии с набором из 100 случайных чисел.

Начнем, например, со второго столбца второй страницы таблицы. Будем учитывать первые три цифры пятизначных чисел. Первые 5 чисел, которые мы получаем (670, 716, 367, 988, 283), больше 250. Шестое число, 142, меньше 251 и больше 000 и, следовательно, означает выбор одного из сотрудников. Из второго столбца будут выбраны числа 021, 166, 127, 060, 098, 219, 161, 042, 043, 157, 113, 234, 024, 028 и 128.

После просмотра второго столбца нам остается получить 84 числа, и мы переходим к третьему столбцу. Просмотрев его, мы выберем 29 различных чисел между 1 и 250 (приведем их в порядке возрастания):

001	028	052	107	142	166
014	034	059	113	146	219
021	042	060	121	157	234
024	043	080	127	160	244
026	047	098	128	161	

Чтобы получить недостающие числа, аналогично просмотрим следующие столбцы. Если встречается число между 1 и 250, уже выбранное ранее (например, 244 встречается еще раз в столбце 4), оно игнорируется.

4.2. Рандомизация в клинических испытаниях

Предположим, необходимо провести клинические испытания лекарственного препарата, чтобы установить его эффективность. Для этого, например, 50 больным назначают лекар-

ство, а другим 50 больным назначают нейтральный препарат («пустышку»). Предположим еще, что больные поступают на испытания сериями, в течение некоторого времени, а не одновременно.

Рассмотрим два метода рандомизации. В первом методе требуется выбрать 50 различных чисел между 1 и 100, активное лекарство должно быть назначено тем из 100 больных, чьи номера попали в этот набор. Остальные 50 пациентов будут получать нейтральный препарат.

Этот метод имеет два недостатка. Во-первых, если придется преждевременно завершить исследование, то общее число пациентов, принимавших активный препарат, с большой вероятностью не будет равно числу пациентов, принимавших нейтральный препарат. Между тем статистические методы сравнения теряют чувствительность, если размеры выборок различаются. Во-вторых, если клиническое состояние пациентов, включающихся в испытание в один момент времени, отличается от состояния пациентов, включающихся в другой момент, или меняются правила приема препаратов, то, несмотря на рандомизацию, две группы, возможно, будут отличаться по типу пациентов или по правилам приема лекарств (см. [Cutler et al., 1966, p. 865]).

Второй возможный метод рандомизации лишен недостатков, присущих первому. С помощью этого метода проводится независимая последовательная рандомизация пациентов, поступающих в течение коротких промежутков времени, по группам лечений.

Предположим, что ежемесячно в испытаниях начинают участвовать десять больных. Разумно случайно назначать пяти пациентам лечение одного вида, а остальным пяти пациентам — другого, повторяя случайное назначение каждый месяц, по мере поступления новых партий больных.

Реализацию этой процедуры можно осуществить, например, с помощью табл. А.4. Будем вести просмотр по десяти цифрам от 0 до 9, поскольку выбор ведется из 10 больных. Нулем обозначим десятого больного. Если мы начнем с пятого столбца, то первыми пятью различными цифрами окажутся 2, 5, 4, 8, 6. Значит, из десяти больных второму, пятому, четвертому, восьмому и шестому будет назначен активный, а остальным — нейтральный препарат.

Продолжая просматривать таблицу, увидим, что из следующих десяти больных первый, третий, пятый, восьмой и десятый будут принимать активный, а остальные — нейтральный препарат. Использовав первые цифры в столбце, можно продолжать просмотр по вторым цифрам этого столбца.

Для каждой следующей группы больных следует получать новый набор случайных чисел, чтобы избежать смещений, которые могут появиться вследствие скрытой периодичности типа больных или ввиду того, что сотрудникам клиники вскоре будет ясен вид лекарства (он должен быть неизвестен сотрудникам, контактирующим с пациентами).

Частный случай этого метода — испытания на парах пациентов, когда один из двух пациентов получает активный, а другой — нейтральный препарат. В этом случае рандомизацию проводить очень просто.

Сначала каким-либо образом, например по алфавитному порядку фамилий, выделяют одного из двух больных как первого. Этот выбор надо сделать до проведения рандомизации. Затем, начиная с любого удобного места, просматривают однозначные числа в табл. А.4. Если цифра нечетная — 1, 3, 5, 7 или 9, то первый больной принимает активный, а второй — нейтральный препарат. Если цифра четная — 0, 2, 4, 6 или 8, активное лекарство назначают второму больному.

Рассмотрим пример. Начнем с первого столбца третьей страницы таблицы. Первой будет четная цифра 2. Это означает, что в первой паре активное лекарство назначат второму больному, а нейтральное — первому. Вторая цифра — 8, т. е. назначение во второй паре будет проведено аналогично. Шестая, седьмая, восьмая цифры — 3, 9 и 1. Значит, в шестой, седьмой и восьмой парах активный препарат назначается первому, а нейтральный — второму больному.

Исследователи, которым не хватит 20 000 случайных чисел в табл. А.4, могут обратиться к обширным таблицам из [Rand Corporation, 1955]. Для тех, чьи научные интересы требуют применения к каждой выборке объектов каждой из (более чем двух) обработок, незаменимы таблицы [Moses and Oakford, 1963]. Эти таблицы пригодны как для первого, так и для второго метода рандомизации, описанных выше.

4.3. Некоторые варианты простой рандомизации

Описанные методы рандомизации приводят к назначению каждому пациенту одного из двух видов лечения с шансами 50 на 50. Все эти методы за исключением тех из них, в которых пары пациентов подбираются по признакам, взаимодействующим с изучаемым фактором, сопряжены с риском дисбаланса между группами больных (в которых проводится различное лечение) в распределении возраста, пола, начальной тяжести заболевания или других прогностически важных факторов. Кроме метода подбора пар по прогностическим факторам,

можно использовать другое решение: провести стратификацию, т. е. разделить пациентов на определенные слои, группы (например, выделить мужчин в возрасте от 20 до 29 лет, женщин 20—29 лет, мужчин 30—39 лет и т.д.), а затем применить независимо и раздельно внутри каждого слоя один из методов простой рандомизации.

Простая рандомизация с расслоением уменьшает, но не устраняет полностью риск дисбаланса, особенно если испытания проводятся на пациентах, поступающих в разное время, и окончательное число пациентов в каждой группе неизвестно до конца набора пациентов, участвующих в испытаниях. Средством дальнейшего уменьшения возможности дисбаланса является концепция «несимметричной монеты» (biased coin), которую предложил Эфрон [Efron, 1971] и развили Покок и Саймон [Pocock and Simon, 1975]. Предположим, что поступающий пациент относится к группе, в которой большему числу больных назначено лечение одного вида, а меньшему числу — лечение другого вида. Тогда, согласно схеме несимметричной монеты, новому больному с некоторой вероятностью $p>0{,}5$ назначают лечение, которое получила на текущий момент меньшая часть больных. Вероятность назначения ему лечения, которое получила большая часть больных, будет равна $1-p<0{,}5$. При равном числе больных, получивших лечение того и другого вида, новому пациенту вид лечения назначается с вероятностью 0,5.

Мы продемонстрируем схему Эфрона с предложенным им (и в общем случае удачным) значением $p=2/3$. Пусть пациент относится к группе, в которой два вида лечения были назначены различному числу пациентов. Из табл. А.4 извлечем одно из целых чисел от 1 до 9. Если оно делится на 3, (3, 6 или 9), то больному назначают лечение, полученное на этот момент большим числом пациентов рассматриваемой группы, а если нет (1, 2, 4, 5, 7 или 8), то назначают лечение другого вида.

Вероятность назначения определенного лечения, отличная от 1/2, требуется также в так называемых адаптивных клинических испытаниях. Здесь проводят лечение так, чтобы к концу испытаний большее число пациентов прошло курс лечения, считающегося лучшим, а меньшее число — курс лечения, которое считается менее эффективным (обсуждение адаптивных клинических испытаний и ссылки на соответствующие работы даны в гл. 7). Пусть задана вероятность $p>0{,}5$ того, что пациенту назначают лечение, которое к этому моменту считается лучше другого (если ни одному из видов лечения не отдается предпочтение, то полагают $p=0{,}5$). В табл. А.4 просматривают двузначные числа: если выбранное целое число лежит в ин-

тервале от единицы до $100p$, то поступившему пациенту назначают лечение лучшего вида, если же число больше $100p$ (нуль принимается за 100), то назначают лечение второго вида. Например, при $p=0.60$ любое число от 01 до 60 будет соответствовать назначению первого вида, от 61 до 00 — второго.

ЛИТЕРАТУРА

Cutler, S. J., Greenhouse, S. W., Cornfield, J., and Schneiderman, M. A. (1966). The role of hypothesis testing in clinical trials: Biometrics seminar. *J. Chronic Dis.*, 19, 857–882.

Efron, B. (1971). Forcing a sequéntial experiment to be balanced. *Biometrika*, 58, 403–417.

Fisher, R. A. (1935). *The design of experiments.* Edinburgh: Oliver and Boyd.

Hill, A. B. (1962). *Statistical methods in clinical and preventive medicine.* New York: Oxford University Press.

Moses, L. E. and Oakford, R. V. (1963). *Tables of random permutations.* Stanford: Stanford University Press.

Pocock, S. J. and Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*, 31, 103–115.

Rand Corporation (1955). *A million random digits with 100,000 normal deviates.* New York: The Free Press.

Глава 5

Метод выбора I. Перекрестные исследования

Данная глава посвящена методу извлечения выборок, обозначенному в разд. 2.1 как метод 1. В этом методе, называемом *перекрестным*, или *мультиномиальным* (cross-sectional, multinomial, naturalistic), не фиксируются заранее какие-либо частоты (кроме общего числа объектов в таблице).

Мы рассматриваем только такой случай, когда данные расположены в таблице 2×2. В большинстве работ (например, [Dixon and Massey, 1969] и [Everitt, 1977]) критерий хи-квадрат для проверки связи описан для частотных таблиц с числом строк и столбцов, большим двух. Точность критерия хи-квадрат для общей таблицы сопряженности при малых объемах выборок изучалась в [Craddock and Flood, 1970]. Методы оценивания связи в таких таблицах данных приведены в [Goodman and Kruskal, 1954, 1959; Goodman, 1964; Altham, 1970а, 1970в].

В разд. 5.1 представлены данные, которые в этой и следующей главах используются для иллюстрации. В разд. 5.2 мы обсудим подход к оцениванию степени связи между двумя признаками (факторами), основанный на критерии χ^2, а в разд. 5.3 основанный на других мерах. Некоторые свойства отношения шансов и его логарифма описаны в разд. 5.4.

Методы проверки гипотез и построение доверительных интервалов для отношения шансов описываются соответственно в разд. 5.5 и 5.6. Разд. 5.7 посвящен построению выводов о привносимом риске.

5.1. Иллюстративные данные

Предположим, мы изучаем связь (или ее наличие) между весом новорожденного и возрастом матери. Символ A обозначает, что возраст роженицы — 20 лет или менее, $\bar{A}$ — более 20 лет, B — вес новорожденного 2500 г и менее, $\bar{B}$ — более 2500 г. Поскольку эта связь может меняться в зависимости от

расы и социального положения матери, будем исследовать только чернокожих женщин, рожавших в одной клинике, и принадлежащих к одной социальной группе.

Допустим, на каждого новорожденного заведена карта, где зарегистрированы его вес и возраст, цвет кожи и социальное положение матери. Предположим, что число подходящих карт слишком велико, а определение факта события A или $\overline{A}$ и B или $\overline{B}$ не очень просто. Например, вместо возраста матери в карте может быть записана дата ее рождения. Разумным решением в этой ситуации будет проведение анализа по выборке, состоящей только из 200 записей.

В идеале следует сформировать простую случайную выборку (см. гл. 4), но есть и другие возможности. Пусть имеется 1000 карт с записями о подходящих по цвету кожи и социальному положению матерях. Отбирая каждую пятую карту, начав с карты, случайно выбранной из первых пяти, можно сформировать *систематическую случайную выборку* объема 200. Можно вести отбор другим способом, к примеру отбирать карты с последней цифрой порядкового номера, равной 3 или 7.

Таблица 5.1

Зависимость веса новорожденного от возраста матери. Перекрестное исследование

Возраст матери	Вес новорожденного		Сумма
	B	$\overline{B}$	
A	10	40	50
$\overline{A}$	15	135	150
Сумма	25	175	200

Предположим, что выборка объема 200 получена и данные представлены в табл. 5.1. Как указано в гл. 2, данные, полученные методом выбора I, более адекватно представлять в виде табл. 5.2.

По данным, полученным методом выбора I, можно оценивать все вероятности. Скажем, для новорожденных с весом 2500 г или менее, чьи матери не старше 20 лет, $p(A \text{ и } B) = p_{11} = 0{,}05$. Пропорция новорожденных с весом 2500 г и менее оценивается величиной $p(B) = p_{\cdot 1} = 0{,}125$ и т. д.

Значимость зависимости между весом новорожденного и возрастом матери (это первый, но отнюдь не самый важный

Таблица 5.2

Совместные пропорции для данных табл. 5.1

Возраст матери	Вес новорожденного		Сумма
	B	$\bar{B}$	
A	0,050 ($=p_{11}$)	0,200 ($=p_{12}$)	0,25 ($=p_{1.}$)
$\bar{A}$	0,075 ($=p_{21}$)	0,675 ($=p_{22}$)	0,75 ($=p_{2.}$)
Сумма	0,125 ($=p_{.1}$)	0,875 ($=p_{.2}$)	1

вопрос) можно установить с помощью обычного критерия хи-квадрат. Значение статистики критерия

$$\chi^2 = \frac{200\,(|10\cdot 135 - 40\cdot 15| - 200/2)^2}{50\cdot 150\cdot 25\cdot 175} = 2{,}58 \qquad (5.1)$$

указывает, что зависимость (связь) статистически незначима.

5.2. Меры связи, основанные на χ^2

Отсутствие значимости статистической связи может иногда служить сигналом к завершению анализа. Однако в задачах сравнения часто представляет интерес *степень* связи между двумя признаками, поэтому мы перейдем к оцениванию степени связи. Начнем с мер, основанных на статистике χ^2.

Распространенной ошибкой является использование непосредственно значения χ^2 как меры связи. Даже если χ^2 — прекрасная мера *значимости* связи, она совершенно бесполезна в качестве меры *степени* связи. Объясняется это тем, что χ^2 является функцией как пропорций в различных клетках, так и суммарного числа наблюдений, в то время как мера степени связи должна быть функцией только пропорций. Число объектов в выборке играет роль в определении значимости связи, если связь существует, но не может учитываться при определении степени связи (см., например, [Fisher, 1954, p. 89—90]).

Предположим, что другой специалист, проводя в той же клинике аналогичное исследование по выборке объема 400, получает в результате табл. 5.3. Здесь χ^2 принимает значение

$$\chi^2 = \frac{400\,(|20\cdot 270 - 80\cdot 30| - 400/2)^2}{100\cdot 300\cdot 50\cdot 350} = 5{,}97,$$

что указывает на наличие связи при уровне значимости 0,05.

Таблица 5.3

Зависимость веса новорожденного от возраста матери. Перекрестное исследоваание (400 наблюдений)

Возраст матери	Вес новорожденного		Сумма
	B	$\overline{B}$	
A	20	80	100
$\overline{A}$	30	270	300
Сумма	50	350	400

Выводы, основанные на данных табл. 5.1 и 5.3, различны: в первой таблице связь незначима, во второй — значима. Единственной же причиной различия явилось простое удвоение величин в табл. 5.1. В то же время пропорции (см. табл. 5.2) в этих таблицах идентичны, что соответствует одинаковой степени зависимости между весом новорожденного и возрастом матери. По сути, большее значение χ^2 для второй таблицы (табл. 5.3) является следствием большего объема выборки, а не усиления связи (удвоение всех частот привело к более чем двукратному увеличению хи-квадрат).

Мерой степени связи между A и B, которая основана на χ^2, но не зависит от объема выборки $n_{..}$, является *фи-коэффициент:*

$$\varphi = \sqrt{\frac{\chi_u^2}{n_{..}}}, \tag{5.2}$$

где χ_u^2 — статистика хи-квадрат без поправок:

$$\chi_u^2 = \frac{n_{..}\,(n_{11}\,n_{22} - n_{12}\,n_{21})^2}{n_{1.}\,n_{2.}\,n_{.1}\,n_{.2}}. \tag{5.3}$$

Фи-коэффициент часто используют как меру связи в науках о поведении. Его можно интерпретировать как коэффициент корреляции. Вычислить значение φ можно, присвоив категориям A и $\overline{A}$ два любых различных числовых значения (например, 0 и 1 для простоты), затем присвоив B и $\overline{B}$ два любых различных значения и вычисляя по получаемым данным (по суммам взаимных произведений и квадратов) обычный коэффициент корреляции.

Значение φ, близкое к нулю, означает слабую связь или отсутствие связи. Если значение φ близко к 1, то связь сильная, и по принадлежности объекта к A или $\overline{A}$ можно с боль-

ной точностью предсказать его принадлежность к одной из категорий B и $\bar{B}$. Максимальное значение φ — единица. Если маргинальные частоты не равны, максимальное значение меньше 1. В качестве значений, соответствующих слабой связи, можно ориентировочно указать величины менее 0,30—0,35.

Для табл. 5.1

$$\chi^2_u = \frac{200\,(10\cdot 135 - 40\cdot 15)^2}{50\cdot 150\cdot 25\cdot 175} = 3{,}43,$$

так что

$$\varphi = \sqrt{\frac{3{,}43}{200}} = 0{,}13. \qquad (5.4)$$

Едва ли это значение можно считать высоким. Значение φ для табл. 5.3, очевидно, также равно 0,13.

Фи-коэффициент используется преимущественно при анализе ответов в психологических тестах, тестах, применяемых в педагогике [Lord and Novick, 1968] и в факторном анализе дихотомических признаков типа «да—нет». Общее описание факторного анализа дано в [Harman, 1960] и [Nunnally, 1978]. Применимость фи-коэффициента в факторном анализе обсуждается в [Lord and Novick, 1968]. Бергер [Berger, 1961] описывает метод сравнения двух фи-коэффициентов, полученных по двум независимым выборкам.

Однако фи-коэффициенту присущи серьезные недостатки. Как показано в гл. 6, значения φ, полученные при проспективном и ретроспективном изучении связи между A и B, не согласуются как друг с другом, так и со значением, полученным в перекрестном исследовании. Кэролл [Carroll, 1961] показал, что, если один или оба непрерывных признака дихотомизированы с помощью разбиения непрерывного распределения на две части, то значение φ сильно зависит от точки (точек) дихотомизации.

Отсутствие инвариантности фи-коэффициента и других мер, основанных на χ^2, указано в [Goodman and Kruskal, 1954]. На с. 740 Гудмен и Краскел пишут, что они «не смогли найти опубликованное доказательное выступление в пользу применения статистик, основанных на χ^2, в качестве мер связи». Хотя это высказывание игнорирует полезность фи-коэффициента в психометрии, в то же время оно справедливо указывает на необходимость отказаться от использования φ и других мер связи, основанных на χ^2, в тех исследованиях, где нужна сравнимость результатов для различных методов проведения исследования.

5.3. Другие меры связи. Отношение шансов

Гудмен и Краскел [Goodman and Kruskal, 1954, 1959] описали большое число мер связи, которые не являются функциями χ^2, и их статистические свойства в двух более поздних работах (1963, 1972). Мы сконцентрируем свое внимание на одной из этих мер — на *отношении шансов* (odds ratio).

Часто один из двух признаков (факторов) является исходным по отношению к другому. В разд. 5.1 возраст роженицы — исходный фактор по отношению к весу новорожденного. Мера риска появления изучаемого результирующего фактора в присутствии исходного фактора A определяется следующим образом:

$$\Omega_A = \frac{P(B \mid A)}{P(\bar{B} \mid A)} \tag{5.5}$$

(определение условной вероятности дано в разд. 1.1). Ω_A — шансы появления B при выполнении A. Поскольку $P(B|A)$ можно оценить как

$$p(B \mid A) = \frac{p_{11}}{p_{1.}},$$

а $P(\bar{B}|A)$ — как

$$p(\bar{B} \mid A) = \frac{p_{12}}{p_{1.}},$$

то в качестве оценки для Ω_A можно взять

$$O_A = \frac{p_{11}/p_{1.}}{p_{12}/p_{1.}} = \frac{p_{11}}{p_{12}}. \tag{5.6}$$

В нашем примере такая оценка для матерей не старше 20 лет иметь новорожденного с весом до 2500 г равна (см. табл. 5.2)

$$O_A = \frac{0{,}05}{0{,}20} = 1/4 = 0{,}25. \tag{5.7}$$

Значит, у матерей не старше 20 лет на каждых четырех младенцев, родившихся с весом более 2500 г, приходится один новорожденный с весом не более 2500 г.

Эта оценка содержит ту же информацию, что и доля новорожденных с низким весом, наблюдаемая у молодых матерей,

$$p(B \mid A) = \frac{0{,}05}{0{,}25} = 1/5 = 0{,}20,$$

но акценты различны. Представим, что мы проводим консультацию в группе будущих матерей не старше 20 лет. Слова

«Ожидается, что каждая пятая из вас родит ребенка с низким весом» и слова «Ожидается, что на четыре ребенка с нормальным весом, рожденных вами, придется один ребенок с низким весом» произведут, по-видимому, разное впечатление на будущих матерей.

Когда A не происходит, шансы, что происходит B, равны

$$\Omega_{\bar{A}} = \frac{P(B \mid \bar{A})}{P(\bar{B} \mid \bar{A})}, \tag{5.8}$$

что можно оценить величиной

$$O_{\bar{A}} = \frac{p_{21}/p_{2.}}{p_{22}/p_{2.}} = \frac{p_{21}}{p_{22}}. \tag{5.9}$$

В нашем примере оценка шансов для матерей старше 20 лет иметь новорожденного с весом 2500 г или меньше равняется:

$$O_{\bar{A}} = \frac{0{,}075}{0{,}675} = 1/9. \tag{5.10}$$

Значит, у матерей старше 20 лет на девять детей с нормальным весом приходится один ребенок с весом не более 2500 г (в отличие от шансов для молодых матерей, равных 4 к 1).

Сопоставляя различными способами шансы Ω_A (5.5) и $\Omega_{\bar{A}}$ (5.8), можно получить различные меры связи. Одной из них является мера, предложенная Юлом [Yule, 1900]:

$$Q = \frac{\Omega_A - \Omega_{\bar{A}}}{\Omega_A + \Omega_{\bar{A}}}. \tag{5.11}$$

Другая мера, также предложенная Юлом [Yule, 1912], определяется выражением

$$Y = \frac{\sqrt{\Omega_A} - \sqrt{\Omega_{\bar{A}}}}{\sqrt{\Omega_A} + \sqrt{\Omega_{\bar{A}}}}. \tag{5.12}$$

Наиболее популярной в настоящее время мерой связи, основанной на Ω_A и $\Omega_{\bar{A}}$, является просто их отношение

$$\omega = \frac{\Omega_A}{\Omega_{\bar{A}}}, \tag{5.13}$$

которое можно оценить выборочным отношением шансов

$$o = \frac{O_A}{O_{\bar{A}}} = \frac{p_{11}/p_{12}}{p_{21}/p_{22}} = \frac{p_{11}\ p_{22}}{p_{12}\ p_{21}}. \tag{5.14}$$

Если вероятности $P(B|A)$ и $P(B|\bar{A})$ равны, что указывает на независимость признаков (отсутствие связи), то шансы Ω_A

и $\Omega_{\bar{A}}$ также равны (см. задачу 5.1), и отношение шансов $\omega=1$. Если $P(B|A)>P(B|\bar{A})$, то $\Omega_A>\Omega_{\bar{A}}$ и $\omega>1$ (см. задачу 5.2). Если $P(B|A)<P(B|\bar{A})$, то $\Omega_A<\Omega_{\bar{A}}$ и $\omega<1$.

Для наших данных оценка отношения шансов

$$o=\frac{0{,}05\cdot 0{,}675}{0{,}20\cdot 0{,}075}=2{,}25, \tag{5.15}$$

что указывает, что у молодой матери шансы родить ребенка с низким весом в $2^1/_4$ раза выше, чем у матери старше 20 лет. Отношение шансов оценивается посредством

$$o=\frac{n_{11}\,n_{22}}{n_{12}\,n_{21}}. \tag{5.16}$$

Эту величину часто называют *«перекрестным отношением»*.

Стандартную ошибку оценки отношения шансов можно оценить величиной

$$\text{s. e. }(o)=\frac{o}{\sqrt{n_{\cdot\cdot}}}\sqrt{\frac{1}{p_{11}}+\frac{1}{p_{12}}+\frac{1}{p_{21}}+\frac{1}{p_{22}}}. \tag{5.17}$$

Для данных табл. 5.2

$$\text{s. e. }(o)=\frac{2{,}25}{\sqrt{200}}\sqrt{\frac{1}{0{,}05}+\frac{1}{0{,}2}+\frac{1}{0{,}075}+\frac{1}{0{,}675}}=1{,}00. \tag{5.18}$$

Выражение (5.17) через частоты запишется в виде

$$\text{s. e. }(o)=o\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}}. \tag{5.19}$$

По величине стандартной ошибки можно судить о точности оценки отношения шансов, но не строить доверительные интервалы или проверять гипотезы. В качестве критерия проверки гипотезы о равенстве отношения шансов единице, $\omega=1$, следует использовать классический критерий хи-квадрат; методы построения доверительных интервалов для ω описаны в разд. 5.6.

В [Anscombe, 1956; Gart, 1966; Gart and Zweifel, 1967] изучены выборочные свойства o и его стандартной ошибки. Оценка (5.16) не определена, если n_{12} или n_{21} равны нулю; если же хотя бы одна из четырех частот равна нулю, то не определена s. e. (o) в (5.19). Для устранения этого свойства была предложена модификация оценки отношения шансов

$$o'=\frac{(n_{11}+0{,}5)\,(n_{22}+0{,}5)}{(n_{12}+0{,}5)\,(n_{21}+0{,}5)}. \tag{5.20}$$

Оценка стандартной ошибки имеет теперь вид:

$$\text{ст. е. } (o') = o' \sqrt{\frac{1}{n_{11}+0{,}5} + \frac{1}{n_{12}+0{,}5} + \frac{1}{n_{21}+0{,}5} + \frac{1}{n_{22}+0{,}5}}. \quad (5.21)$$

В последующем изложении мы опишем свойства отношения шансов, важные для мер связи. Преимущества отношения шансов перед другими подходами к мерам связи проиллюстрированы Мостеллером [Mosteller, 1968]. Эдвардс [Edwards, 1963] считает, что эти преимущества настолько велики, что в качестве меры связи в таблицах 2×2 следует использовать только отношение шансов или функции от него.

5.4. Некоторые свойства отношения шансов и его логарифма

Отношение шансов как показатель связи между двумя факторами, один из которых является исходным по отношению ко второму, результирующему (например, смертности или заболеваемости), впервые предложено Корнфилдом [Cornfield, 1951]. Однако отношение шансов использовалось лишь в связи с аппроксимацией другой меры, также предложенной им, — *относительным риском*. Относительный риск определяется просто как отношение условных вероятностей $P(B|A)$ и $P(B|\bar{A})$ появления B в присутствии и в отсутствии A:

$$R = \frac{P(B|A)}{P(B|\bar{A})}. \quad (5.22)$$

R можно оценить величиной

$$z = \frac{p_{11}/p_{1\cdot}}{p_{21}/p_{2\cdot}} = \frac{p_{11}\,p_{2\cdot}}{p_{21}\,p_{1\cdot}}. \quad (5.23)$$

Если появление B маловероятно как при A, так и при $\bar{A}$, то, как показано в задаче 5.3, r приближенно равно o (см. (5.14)). Для табл. 5.2

$$r = \frac{0{,}05 \cdot 0{,}75}{0{,}075 \cdot 0{,}25} = 2{,}0, \quad (5.24)$$

что лишь немного меньше значения отношения шансов $o = 2{,}25$ (см. (5.15)).

Однако отношение шансов, конечно, важно не только как аппроксимация относительного риска. Отношение шансов как мера связи естественным образом возникает в так называемой *логистической модели*. Рассмотрим конкретный пример — связь между курением и раком легких. Смертность от рака легких зависит не только от курения, но и от других факторов, на-

пример от загрязненности воздуха в месте проживания или работы.

Допустим, что мы изучаем связь между курением и раком легких в некоторой группе, однородной по месту проживания и работы. Пусть x — средний показатель загрязненности воздуха для этой группы. Смертность от рака легких среди курящих можно представить в виде

$$P_S = \frac{1}{1 + e^{-(ax + b_S)}}, \tag{5.25}$$

а среди некурящих — в виде

$$P_N = \frac{1}{1 + e^{-(ax + b_N)}}, \tag{5.26}$$

где $e = 2{,}718...$ — основание натурального логарифма. Параметр a связывает смертность со степенью загрязнения воздуха x. Из того, что в (5.25) и (5.26) входит один и тот же параметр a, следует, что нет никакого согласованно работающего воздействия курения и загрязнения воздуха на смертность. Если a больше нуля, то и P_S, и P_N стремятся к единице при увеличении степени загрязненности x.

Согласно модели (5.25)—(5.26) различие уровней смертности курящих и некурящих определяется только различием параметров b_S и b_N. Если загрязненность воздуха отсутствует, т. е. $x = 0$, то смертность среди курящих зависит только от b_S, а среди некурящих — только от b_N.

Теперь рассмотрим шансы курящего из выбранной группы умереть от рака легких. Они равны:

$$\Omega_S = \frac{P_S}{1 - P_S}.$$

Так как

$$1 - P_S = \frac{e^{-(ax + b_S)}}{1 + e^{(-ax + b_S)}},$$

то

$$\Omega_S = \frac{1}{e^{-(ax + b_S)}} = e^{ax + b_S}. \tag{5.27}$$

Аналогично шансы некурящего умереть от рака легких

$$\Omega_N = e^{ax + b_N}. \tag{5.28}$$

Значит, если справедлива логистическая модель, то отношение шансов равно:

$$\omega = \frac{\Omega_S}{\Omega_N} = \frac{e^{ax + b_S}}{e^{ax + b_N}} = e^{(b_S - b_N)} \tag{5.29}$$

и не зависит от x. Натуральный логарифм отношения шансов также не зависит от x и равен просто разности

$$\ln \omega = b_S - b_X . \tag{5.30}$$

Важность этого результата заключается в следующем. Если отношение шансов или его логарифм приближенно равен для многих различных популяций, то правомерным будет вывод, что логистическая модель удовлетворительно описывает изучаемое явление. Основываясь на этом выводе, можно предсказывать отношение шансов в новой популяции и проверять различие между предсказанным и наблюдаемым значениями. Можно прогнозировать влияние изменений управляемого фактора x (например, загрязненности воздуха) на смертность и, конечно, прогнозировать изменение смертности при отказе от курения.

Из представления (5.30) следует, что логарифм выборочного отношения шансов,

$$L = \ln o, \tag{5.31}$$

является важной мерой связи. Стандартную ошибку L изучали Вулф [Woolf, 1955], Холдейн [Haldane, 1956], и Гарт [Gart, 1966]. Как было показано, более хорошей оценкой $\ln \omega$ является

$$L' = \ln o', \tag{5.32}$$

где o' определено в (5.20), а хорошая оценка ее стандартной ошибки есть

$$\text{ст. ош.}(L') = \sqrt{\frac{1}{n_{11}+0{,}5} + \frac{1}{n_{12}+0{,}5} + \frac{1}{n_{21}+0{,}5} + \frac{1}{n_{22}+0{,}5}} . \tag{5.33}$$

Из (5.30) следует, что в логистической модели (5.26) и (5.27) $\ln \omega$ не зависит от x. Даже если в данной задаче применяется не логистическая модель, а модель, основанная на нормальном распределении, $\ln \omega$ почти не зависит от x [Edwards, 1966, Weiss, 1970]. Однако логистическая модель намного удобнее формальной, когда дело касается пропорций и долей. Именно с этих целей логистическая модель неоднократно использовалась [Bartlett, 1935; Winsor, 1948; Dyke and Patterson, 1952, Cox, 1958, 1970; Grizzle, 1961, 1963; Maxwell and Everitt, 1970, Goodberg, 1977].

5.5. Проверка гипотез об отношении шансов

Теоретические результаты. Будем рассматривать четырехклеточную таблицу. Предположим, что наблюдаемые значения маргинальных частот равны $n_{1.}$, $n_{2.}$, $n_{.1}$ и $n_{.2}$, а истинное

значение отношения шансов равно ω. Ожидаемые частоты N_{ij} определяются а) наблюдаемыми маргинальными частотами (см. табл. 5.4) и

Таблица 5.4

Ожидаемые частоты в четырехклеточной таблице

Фактор A	Фактор B		Сумма
	Присутствует	Отсутствует	
Присутствует	N_{11}	N_{12}	$n_{1.}$
Отсутствует	N_{21}	N_{22}	$n_{2.}$
Сумма	$n_{.1}$	$n_{.2}$	$n_{..}$

б) значением ω:

$$\frac{N_{11}\ N_{22}}{N_{12}\ N_{21}} = \omega\,. \tag{5.34}$$

Гипотезу о том, что истинное значение отношения шансов равно ω, можно проверить, сравнивая статистику

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(|n_{ij} - N_{ij}| - 1/2)^2}{N_{ij}} \tag{5.35}$$

со значениями распределения хи-квадрат с одной степенью свободы. Вид статистики (5.35) критерия совпадает с видом (2.7) классической статистики, как идентична и интерпретация N_{ij} как частот, ожидаемых при гипотетическом значении отношения шансов (в (2.7) $\omega=1$, в (5.35) ω произвольно).

Если за нулевую гипотезу принято отношение шансов, равное ω и $\omega\neq1$, то ожидаемые частоты можно найти следующим образом. Определим

$$X = \omega\ (n_{1.} + n_{.1}) + (n_{2.} - n_{.1}) \tag{5.36}$$

и

$$Y = \sqrt{X^2 - 4\ n_{1.}\ n_{.1}\ \omega\ (\omega - 1)}\ , \tag{5.37}$$

тогда

$$N_{11} = \frac{X - Y}{2\ (\omega - 1)}\,; \tag{5.38}$$

$$N_{12} = n_{1.} - N_{11}; \tag{5.39}$$

$$N_{21} = n_{.1} - N_{11}; \tag{5.40}$$

$$N_{22} = n_{.2} - n_{1.} + N_{11}. \tag{5.41}$$

Стивенсом и Корнфилдом [Stevens, 1951; Cornfield, 1956] было доказано, что при фиксированных значениях маргинальных частот и отношения шансов частоты n_{ij} (в каждой из четырех клеток) распределены приближенно нормально со средними N_{ij} и стандартной ошибкой $1/\sqrt{W}$, где

$$W = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{1}{N_{ij}}, \tag{5.42}$$

а N_{ij} определены в (5.38) — (5.41).

Применение. Проверим по данным табл. 5.1 гипотезу о том, что значение отношения шансов $\omega = 5$. Значение X в (5.36) равно:

$$X = 5\ (50 + 25) + (150 - 25) = 500, \tag{5.43}$$

значение Y в (5.37) —

$$Y = \sqrt{500^2 - 4 \cdot 50 \cdot 25 \cdot 5 \cdot 4} = 387{,}30. \tag{5.44}$$

Ожидаемые частоты представлены в табл. 5.5.

Таблица 5.5

Ожидаемые частоты для данных табл. 5.1 при отношении шансов $\omega = 5$

Возраст матери	Вес новорожденного		Сумма
	B	$\overline{B}$	
A	14,1	35,9	50
$\overline{A}$	10,9	139,1	150
Сумма	25	175	200

Для проверки вычислим заново отношение шансов:

$$\frac{14{,}1 \cdot 139{,}1}{35{,}9 \cdot 10{,}9} = 5{,}0. \tag{5.45}$$

Значение статистики (5.35) равно:

$$\chi^2 = \frac{(|10 - 14{,}1| - 0{,}5)^2}{14{,}1} + \frac{(|40 - 35{,}9| - 0{,}5)^2}{35{,}9} +$$
$$+ \frac{(|15 - 10{,}9| - 0{,}5)^2}{10{,}9} + \frac{(|135 - 139{,}1| - 0{,}5)^2}{139{,}1} = 2{,}56. \tag{5.46}$$

Итак, гипотеза «$\omega = 5$» не отвергается.

Другой критерий для проверки этой гипотезы строится на основе результатов предыдущего раздела. Обозначим $\lambda = \ln \omega$. Величина

$$\chi^2 = \frac{(L' - \lambda)^2}{(\text{s. e. } (L'))^2} \tag{5.47}$$

распределена приближенно как хи-квадрат с одной степенью свободы. Для данных нашего примера

$$\lambda = \ln 5 = 1{,}61; \tag{5.48}$$

$$L' = \ln \frac{10{,}5 \cdot 135{,}5}{40{,}5 \cdot 15{,}5} = 0{,}82; \tag{5.49}$$

$$\text{s. e. } (L') = \sqrt{\frac{1}{10{,}5} + \frac{1}{40{,}5} + \frac{1}{15{,}5} + \frac{1}{135{,}5}} = 0{,}438. \tag{5.50}$$

Значение статистики (5.47)

$$\chi^2 = \frac{(0{,}82 - 1{,}61)^2}{0{,}438^2} = 3{,}25 \tag{5.51}$$

больше соответствующего значения (5.46) статистики (5.35), но по-прежнему незначимо.

Значение статистики (5.47), основанной на логарифме отношения шансов, обычно больше значения статистики (5.35), основанной на сравнении N_{ij} и n_{ij}, хотя различие между ними мало́ при больших маргинальных частотах. Если бы статистика (5.35) определялась без поправки на непрерывность, ее значение было бы близко к значению (5.47) и при умеренных объемах выборок (см. задачу 5.5).

Критерий (5.35), описанный в этом разделе, более сложен, но и более точен, чем критерий (5.47). При проверке гипотез о значении ω следует предпочесть критерий (5.35).

5.6. Построение доверительных интервалов для отношения шансов

В качестве $100(1-\alpha)\%$-ного доверительного интервала для ω можно взять множество значений ω, для которых

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(|n_{ij} - N_{ij}| - 1/2)^2}{N_{ij}} \leqslant c^2_{\alpha/2}, \tag{5.52}$$

где N_{ij} — соответствующие ожидаемые частоты (5.38)—(5.41). Верхняя и нижняя границы интервала берутся в точках, где значение (5.52) равно $c^2_{\alpha/2}$.

Статистика (5.52) зависит от ω не явно, а через (5.36) — (5.41), так что процедура вычисления верхней и нижней границ не проста. Однако чрезмерной сложности здесь нет, и вычисления проводятся по следующей схеме.

При фиксированных маргинальных частотах нижней доверительной границе ω_L соответствуют значения ожидаемых частот $N_{11}<n_{11}$, $N_{22}<n_{22}$, $N_{12}>n_{12}$, $N_{21}>n_{21}$. Раскрывая модули разностей в (5.52), получим, что определение нижней границы упрощается до решения уравнения

$$\chi_L^2 = (n_{11} - N_{11} - 1/2)^2 \; W = c_{\alpha/2}^2, \tag{5.53}$$

где W определено в (5.42).

Значение ω_L будет найдено, когда

$$F = (n_{11} - N_{11} - 1/2)^2 \; W - c_{\alpha/2}^2 \tag{5.54}$$

будет равно нулю. Определим

$$T = \frac{1}{2(\omega-1)^2}\left(Y - n_{..} - \frac{\omega-1}{Y}\left[X\ (n_{1.}+n_{.1}) - \right.\right.$$
$$\left.\left. - 2\, n_{1.}\, n_{.1}\ (2\omega - 1)\right]\right), \tag{5.55}$$

$$U = \frac{1}{N_{12}^2} + \frac{1}{N_{21}^2} - \frac{1}{N_{11}^2} - \frac{1}{N_{22}^2}, \tag{5.56}$$

$$V = T\left[(n_{11} - N_{11} - 1/2)^2\, U - 2\, W\, (n_{11} - N_{11} - 1/2)\right]. \tag{5.57}$$

Пусть $\omega_L^{(1)}$ — начальное приближение к ω_L, X и Y — соответствующие значения (5.36) и (5.37), а N_{11}, N_{12}, N_{21}, N_{22} — соответствующие значения (5.38)—(5.41). Если значение F в (5.54) не равно нулю, то следующим, более хорошим приближением к ω_L будет значение

$$\omega_L^{(2)} = \omega_L^{(1)} - \frac{F}{V}. \tag{5.58}$$

Если значение F, соответствующее последнему приближению, не равно нулю (или, скажем модуль F больше 0,05), то нужно продолжить итеративный процесс.

Поиск ω_U, верхней доверительной границы, проводится точно так же, только поправка на непрерывность в (5.53), (5.54) и (5.57) принимает значение $+0{,}5$. Хорошими начальными приближениями к ω_L и ω_U служат границы интервала для логарифма отношения шансов,

$$\omega_L^{(1)} = \exp\left(L' - c_{\alpha/2}\ \text{s. e.}\ (L')\right) \tag{5.59}$$

и

$$\omega_U^{(1)} = \exp\left(L' + c_{\alpha/2}\ \text{s. e.}\ (L')\right). \tag{5.60}$$

Снова обратимся к данным табл. 5.1 и найдем 95%-ный доверительный интервал для ω. Из (5.49) и (5.50)

$$\omega_L^{(1)} = \exp\,(0{,}82 - 1{,}96 \cdot 0{,}438) = 0{,}96 \qquad (5.61)$$

и

$$\omega_U^{(1)} = \exp\,(0{,}82 + 1{,}96 \cdot 0{,}438) = 5{,}37\,. \qquad (5.62)$$

Определим сначала нижнюю границу. Значение X в (5.36), соответствующее $\omega_L^{(1)} = 0{,}96$, равно:

$$X = 0{,}96\,(50 + 25) + (150 - 25) = 197{,}0\,, \qquad (5.63)$$

а значение Y —

$$Y = \sqrt{197^2 - 4 \cdot 50 \cdot 25 \cdot 0{,}96 \cdot (-0{,}04)} = 197{,}49. \qquad (5.64)$$

Тогда значения частот (5.38)—(5.41), соответствующие отношению $\omega = 0{,}96$, равны:

$$N_{11} = \frac{197 - 197{,}49}{2 \cdot (-0{,}04)} = 6{,}13 \qquad (5.65)$$

и $N_{12} = 43{,}87$, $N_{21} = 18{,}87$, $N_{22} = 131{,}13$. Значение W в (5.42) есть

$$W = \frac{1}{6{,}13} + \frac{1}{43{,}87} + \frac{1}{18{,}87} + \frac{1}{131{,}13} = 0{,}2465, \qquad (5.66)$$

а значение F в (5.54) —

$$F = (10 - 6{,}13 - 0{,}5)^2 \cdot 0{,}2465 - 3{,}84 = -1{,}04. \qquad (5.67)$$

Статистика (5.53) отличается на величину 1,04 от требуемого значения $c^2_{\alpha/2} = 3{,}84$, поэтому воспользуемся итеративной процедурой.

На первой итерации получим следующие значения для T, U и V:

$$T = \frac{1}{2 \cdot (-0{,}04)^2}\,(197{,}49 - 200 - \frac{-0{,}04}{197{,}49} \cdot [197 \cdot (50 + 25) - \\ - 2 \cdot 50 \cdot 25 \cdot (1{,}92 - 1)]) = 5{,}22; \qquad (5.68)$$

$$U = \frac{1}{43{,}87^2} + \frac{1}{18{,}87^2} - \frac{1}{6{,}13^2} - \frac{1}{131{,}13^2} = -0{,}0233; \qquad (5.69)$$

$$V = 5{,}22\,[(10 - 6{,}13 - 0{,}5)^2 \cdot (-0{,}0233) - 2 \cdot 0{,}2465 \times \\ \times (10 - 6{,}13 - 0{,}5)] = -10{,}05. \qquad (5.70)$$

Из (5.58) получаем второе приближение

$$\omega_L^{(2)} = 0{,}96 - \frac{-1{,}04}{-10{,}05} = 0{,}86. \qquad (5.71)$$

Это значение оказывается равным с точностью до двух знаков нижней границе 95%-ного доверительного интервала. Соответствующие ожидаемые частоты даны в табл. 5.6.

Таблица 5.6

Ожидаемые частоты для данных табл. 5.1 при отношении шансов $\omega = 0{,}86$

Возраст матери	Вес новорожденного		Сумма
	B	$\bar{B}$	
A	5,66	44,34	50
$\bar{A}$	19,34	130,66	150
Сумма	25	175	200

Значение статистики хи-квадрат в (5.53) равно:

$$\chi_L^2 = (10 - 5{,}66 - 0{,}5)^2 \cdot 0{,}2586 = 3{,}81, \qquad (5.72)$$

что близко к требуемому $c_{\alpha/2}^2 = 3{,}84$.

В задаче 5.6 нужно применить описанную итеративную процедуру при определении верхней доверительной границы. Она равна 5,84. Итак, доверительный интервал для отношения шансов с уровнем доверия 95%, построенный по табл. 5.1, есть

$$0{,}86 \leqslant \omega \leqslant 5{,}84\,. \qquad (5.73)$$

Заметим, что он немного шире интервала (0,96, 5,37), основанного на логарифме отношения шансов. Это явление анализировалось в ряде работ [Gart, 1962; Gart and Thomas, 1972; Fleiss, 1979a]. В этих работах показано, что интервал, построенный по точкам, удовлетворяющим (5.52), шире (но при этом точнее) интервала, построенного по (5.59) и (5.60) и, более того, шире (и точнее) интервалов, построенных многими другими приближенными способами.

Томас и Гарт [Thomas and Gart, 1977] затабулировали нижнюю и верхнюю доверительные границы отношения шансов для некоторых таблиц 2×2. В случае произвольной четырехклеточной таблицы можно использовать описанный метод. Он сложнее альтернативных приближенных методов, но более точен и легко программируется на любых программируемых калькуляторах.

Значения ожидаемых частот, соответствующих верхней и нижней доверительным границам отношения шансов, полезны не только для выводов об отношении шансов. Их можно использовать, строя доверительный интервал для любого параметра (фи-коэффициента, относительного риска и т. д.), оценка которого является функцией частот. В табл. 5.6 даны ожидаемые частоты, соответствующие $\omega_L = 0{,}86$. Нижняя 95%-ная доверительная граница для фи-коэффициента, вычисленная по

этим частотам, равна:

$$\varphi_L = \frac{N_{11}N_{22} - N_{12}N_{21}}{\sqrt{n_{1.}\, n_{2.}\, n_{.1} n_{.2}}} = \frac{5{,}66 \cdot 130{,}66 - 44{,}34 \cdot 19{,}34}{\sqrt{50 \cdot 150 \cdot 25 \cdot 175}} = -0{,}02, \quad (5.74)$$

а для относительного риска —

$$R_L = \frac{N_{11}\, n_{2.}}{N_{21}\, n_{1.}} = \frac{5{,}66 \cdot 150}{19{,}34 \cdot 50} = 0{,}88\,. \quad (5.75)$$

В задаче 5.7 требуется найти верхние доверительные границы для этих параметров.

Методы, предложенные в разд. 5.5 и 5.6, описаны для четырехклеточных таблиц, получаемых методом выбора I. Тем не менее они применимы и для выборок, извлеченных методами II и III. Дело в том, что вероятностная структура частот в клетках не зависит от метода извлечения выборки, поскольку маргинальные частоты фиксированы.

5.7. Привносимый риск

Пусть A обозначает присутствие исследуемого фактора риска, B — присутствие результирующего фактора. Полная вероятность выполнения B есть $P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A})$. К сожалению, единого мнения относительно определения *привносимого риска* (attributable risk) нет. Мы воспользуемся определением Левина [Levin, 1953], смысл которого состоит в выделении доли $P(B)$, привносимой фактором риска.

Присутствие фактора B не обязательно вызвано присутствием A: оно будет для части $P(B|\bar{A})$ людей с $\bar{A}$ (т. е. не имеющих фактора риска). Если фактор риска не играет никакой роли, то в той же пропорции фактор B будет присутствовать и среди людей с фактором $\bar{A}$, так что вклад этой группы в $P(B)$ будет $P(B|\bar{A})P(A)$. Разность между действительным вкладом этой группы $P(B|A)P(A)$ и последней величиной, отнесенная к $P(B)$, и есть привносимый риск по Левину:

$$R_A = \frac{P(B|A)\cdot P(A) - P(B|\bar{A})\cdot P(A)}{P(B)} =$$

$$= \frac{P(A)\,[P(B|A) - P(B|\bar{A})]}{P(B|A)\,P(A) + P(B|\bar{A})\,(1 - P(A))} = \frac{P(A)\,(R-1)}{1 + P(A)\,(R-1)}, \quad (5.76)$$

где R — относительный риск

$$R = \frac{P(B|A)}{P(B|\bar{A})}\,. \quad (5.77)$$

Привносимый риск можно интерпретировать как долю, на которую уменьшается $P(B)$ — вероятность присутствия в по-

пуляции результирующего фактора при устранении фактора риска. Следовательно, его можно с пользой применять в управлении образованием или здравоохранением. Правда, привносимый риск зависит от величины $P(A)$ — доли объектов с фактором риска, поэтому при сравнении популяций, где подверженность фактору риска различна, он неприменим (за некоторыми любопытными исключениями, см. задачу 5.9).

Как указано в [Markush, 1977], если данные получены при перекрестном обследовании (таковы, например, некоторые обследования, проводимые Национальным центром медицинской статистики) или, например, при записи актов гражданского состояния местными отделениями здравоохранения, то привносимый риск можно оценить величиной

$$r_A = \frac{p_{11}p_{22} - p_{12}p_{21}}{p_{.1}p_{2.}} \qquad (5.78)$$

(см. задачу 5.8). Уолтер [Walter, 1976] вывел для стандартной ошибки r_A сложное выражение. Значительно проще связанная с этим выражением оценка стандартной ошибки для $\ln(1-r_A)$, данная Флейсом [Fleiss, 1979b]:

$$\text{s. e. } (\ln\ (1 - r_A)) = \sqrt{\frac{p_{12} + r_A\ (p_{11} + p_{22})}{n..\ p_{21}}} \qquad (5.79)$$

Продемонстрируем использование этой оценки на примере табл. 5.7, в которой представлены данные о весе и выживании детей, родившихся у белого населения в Нью-Йорке в 1974 г.

Таблица 5.7

Зависимость смертности детей, родившихся у белого населения в Нью-Йорке в 1974 г., от их веса при рождении (72730 наблюдений; учитываются только дети, родившиеся живыми) *

Вес новорожденного, г	Умерло	Выжило	Сумма
⩽2500	0,0085	0,0632	0,0717
>2500	0,0058	0,9225	0,9283
Сумма	0,0143	0,9857	1

* Данные за 1 год.

Оценка риска смерти, привносимого низким весом, равна:

$$r_A = \frac{0{,}0085 \cdot 0{,}9225 - 0{,}0632 \cdot 0{,}0058}{0{,}0143 \cdot 0{,}9283} = 0{,}563\,. \qquad (5.80)$$

Оценка стандартной ошибки $\ln(1-r_A)$

$$\text{s. e. } (\ln\ (1-r_A)) = \sqrt{\frac{0{,}0632 + 0{,}563\ (0{,}0085 + 0{,}9225)}{72{,}730 \cdot 0{,}0058}} = 0{,}037\ . \tag{5.81}$$

Вычисляем:

$$\ln(1-r_A) = \ln(1-0{,}563) = -0{,}828\ , \text{ тогда} \tag{5.82}$$

приближенный 95%-ный доверительный интервал для $\ln(1-R_A)$ — это

$$-0{,}828 - 1{,}96 \cdot 0{,}037 \leqslant \ln(1-R_A) \leqslant -0{,}828 + 1{,}96 \cdot 0{,}037, \tag{5.83}$$

т. е.

$$-0{,}901 \leqslant \ln\ (1-R_A) \leqslant -0{,}755\ . \tag{5.84}$$

Отсюда

$$0{,}530 \leqslant R_A \leqslant 0{,}594 \tag{5.85}$$

есть приблизительно 95%-ный доверительный интервал для собственно привносимого риска. Следовательно, при уровне доверия 95% в Нью-Йорке в 1974 г. можно было бы, наверное, предотвратить от 53 до 59% всех смертей новорожденных у белых, если бы полностью было предупреждено недонашивание (т. е. рождение детей с весом 2500 г и менее).

Задачи

5.1. Шансы Ω_A и $\Omega_{\bar{A}}$ определены в (5.5) и (5.8). Докажите, что $\Omega_A = \Omega_{\bar{A}}$, если и только если $P(B|A) = P(B|\bar{A})$.

5.2. Отношение шансов ω определено в (5.13). Докажите, что $\omega > 1$, если и только если $P(B|A) > P(B|\bar{A})$.

5.3. Относительный риск r определен в (5.23), отношение шансов o — в (5.14). Докажите, что r приближенно равен o, если p_{21} мало по отношению к p_{22}, а p_{11} — по отношению к p_{12}. (Указание: $p_2 = p_{22}\ (1+p_{21}/p_{22})$ и $p_1 = p_{12}\ (1+p_{11}/p_{12})$.)

5.4. Давно известно, что психиатры американских общественных психиатрических клиник при постановке первичного диагноза чаще ошибаются в пользу шизофрении, чем в пользу эмоциональных расстройств, в то время как для британских клиник характерно обратное соотношение. Чтобы выяснить, насколько это различие обусловлено различиями в способах постановки диагноза, психиатры Нью-Йорка и Лондона провели совместное исследование. Данные, приводимые ниже, взяты из отчета об этом исследовании [Cooper et al., 1972].

а) Для участия в исследовании было отобрано по 145 пациентов в Нью-Йорке и в Лондоне. В Нью-Йоркской клинике диагноз «шизофрения» был поставлен 82, «эмоциональное расстройство» — 24 пациентам. Число соответствующих диагнозов в Лондоне составило 51 и 67. Составьте четырехклеточную таблицу, игнорируя пациентов с другими диагнозами.

Психиатры, участвовавшие в исследовании, ставили диагноз по стандартному набору критериев после проведения однотипного опроса пациентов. При этом в Нью-Йорке психиатры поставили диагноз «шизофрения» — 43, «эмоциональное расстройство» — 53 пациентам, в Лондоне — соответственно 33 и 85 пациентам. Составьте четырехклеточную таблицу, игнорируя пациентов с другими диагнозами.

Данные, полученные в результате этих опросов, обрабатывались на компьютере с помощью диагностической программы. В Нью-Йорке соотношение компьютерных диагнозов было 67 и 27, в Лондоне — 56 и 37. Составьте четырехклеточную таблицу, игнорируя пациентов с другими диагнозами.

б) Соотношение диагнозов в Нью-Йорке и в Лондоне можно сравнивать по диагнозам, поставленным врачами клиник, психиатрами-исследователями, компьютером. Вычислите для каждого способа диагностирования отношение шансов, что больному в Нью-Йорке поставят скорее диагноз «шизофрения», чем «эмоциональное расстройство», к соответствующим шансам больного в Лондоне.

Насколько отличаются отношения шансов для диагнозов психиатров-исследователей и компьютера? Как они отличаются от отношения шансов для клиник?

в) Для каждого из трех способов диагностики получены данные с высокими значениями частот в клетках. Это означает, что не обязательно использовать улучшенную оценку (5.20). Проверьте в каждом из трех случаев, что оценка отношения шансов (5.14) лишь немного больше оценки (5.20).

5.5. Когда в выражение (5.35) не включаются поправки на непрерывность, его значение обычно близко к значению статистики (5.47). Вычислите

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} (n_{ij} - N_{ij})^2 / N_{ij}$$

по табл. 5.1 и 5.5. Сравните со значением (5.51).

5.6. Найдите с помощью итеративной процедуры разд. 5.6 верхнюю границу 95%-ного доверительного интервала для отношения шансов, соответствующую данным из табл. 5.1. В качестве начального приближения используйте значение $\omega_U^{(1)} = 5{,}37$ из (5.62).

5.7. Найдите ожидаемые частоты, соответствующие верхней границе, найденной в задаче 5.6. Используя полученные значения, вычислите верхние границы 95%-ных интервалов для фи-коэффициента и относительного риска применительно к табл. 5.1.

5.8. Покажите, что выборочная оценка (5.78) привносимого риска популяции, определенного в (5.76), следует из подстановки в (5.76) вместо $P(A)$ и R их оценок p_1. и (5.23).

5.9. Данные о смертности новорожденных в зависимости от веса при рождении для белых приведены в табл. 5.7. Соответствующие данные для небелых родителей по Нью-Йорку за 1974 г. (всего родилось 37 840 живых детей) приведены в следующей таблице.

Вес новорожденного, г	Умерли	Выжили	Сумма
$\leqslant 2500$	0,0140	0,1147	0,1287
> 2500	0,0088	0,8625	0,8713
Сумма	0,0228	0,9772	1

а) Вычислите оценку привносимого риска в этой группе населения Нью-Йорка. Сравните со значением (5.80) для белых.

б) Вычислите оценку стандартной ошибки величины $\ln(1-r_A)$, где r_A получено в а), и приблизительный 95%-ный доверительный интервал для R_A для небелых. Сравните полученный интервал с (5.85).

ЛИТЕРАТУРА

Altham, P. M. E. (1970a). The measurement of association of rows and columns for an $r \times s$ contingency table. *J. R. Stat. Soc., Ser. B*, **32**, 63–73.

Altham, P. M. E. (1970b). The measurement of association in a contingency table: Three extensions of the cross-ratios and metric methods. *J. R. Stat. Soc., Ser. B.*, **32**, 395–407.

Anscombe, F. J. (1956). On estimating binomial response relations. *Biometrika*, **43**, 461–464.

Bartlett, M. S. (1935). Contingency table interactions. *J. R. Stat. Soc. Suppl.*, **2**, 248–252.

Berger, A. (1961). On comparing intensities of association between two binary characteristics in two different populations. *J. Am. Assoc.*, **56**, 889–908.

Carroll, J. B. (1961). The nature of the data, or how to choose a correlation coefficient. *Psychometrika*, **26**, 347–372.

Cooper, J. E., Kendell, R. E., Gurland, B. J., Sharpe, L., Copeland, J. R. M., and Simon, R. (1972). *Psychiatric diagnosis in New York and London.* London: Oxford University Press.

Cornfield, J. (1951). A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast and cervix. *J. Natl. Cancer Inst.*, **11**, 1269–1275.

Cornfield, J. (1956). A statistical problem arising from retrospective studies. Pp. 135–148 in J. Neyman (Ed.). *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, Vol. 4. Berkeley: University of California Press.

Cox, D. R. (1958). The regression analysis of binary sequences. *J. R. Stat. Soc., Ser. B*, **20**, 215–242.

Cox, D. R. (1970). *Analysis of binary data.* London: Methuen.

Craddock, J. M. and Flood, C. R. (1970). The distribution of the χ^2 statistic in small contingency tables. *Appl. Stat.*, **19**, 173–181.

Dixon, W. J. and Massey, F. J. (1969). *Introduction to statistical analysis*, 3rd ed. New York: McGraw-Hill.

Dyke, G. V. and Patterson, H. D. (1952). Analysis of factorial arrangements when the data are proportions. *Biometrics*, **8**, 1–12.

Edwards, A. W. F. (1963). The measure of association in a 2×2 table. *J. R. Stat. Soc., Ser. A*, **126**, 109–114.

Edwards, J. H. (1966). Some taxonomic implications of a curious feature of the bivariate normal surface. *Brit. J. Prev. Soc. Med.*, **20**, 42–43.

Everitt, B. S. (1977). *The analysis of contungency tables.* London: Chapman and Hall.

Fienberg, S. E. (1977). *The analysis of cross-classified categorical data.* Cambridge, Mass.: M.I.T. Press.

Fisher, R. A. (1954). *Statistical methods for research workers*, 12th ed. Edinburgh: Oliver and Boyd.

Русский перевод: Фишер Р.
Статистические методы для исследователей. – М.: Госстатгиз, 1958.

Fleiss, J. L. (1970). On the asserted Invariance of the odds ratio. *Brit. J. Prev. Soc. Med.*, **24**, 45–46.

Fleiss, J. L. (1979a). Confidence intervals for the odds ratio in case-control studies: The state of the art. *J. Chronic Dis.*, **32**, 69–77.

Fleiss, J. L. (1979b). Inference about population attributable risk from cross-sectional studies. *Am. J. Epidemiol.*, **110**, 103–104.

Gart, J. J. (1962). Approximate confidence limits for the relative risk. *J. R. Stat. Soc., Ser. B*, **24**, 454–463.

Gart, J. J. (1966). Alternative analyses of contingency tables. *J. R. Stat. Soc., Ser. B,* 28, 164–179.
Gart, J. J. and Thomas, D. G. (1972). Numerical results on approximate confidence limits for the odds ratio. *J. R. Stat. Soc., Ser. B,* 34, 441–447.
Gart, J. J. and Zweifel, J. R. (1967). On the bias of various estimators of the logit and its variance, with application to quantal bioassay. *Biometrika,* 54, 181–187.
Goodman, L. A. (1964). Simultaneous confidence limits for cross-product ratios in contingency tables. *J. R. Stat. Soc., Ser. B,* 26, 86–102.
Goodman, L. A. and Kruskal, W. H. (1954). Measures of association for cross classifications. *J. Am. Stat. Assoc.,* 49, 732–764.
Goodman, L. A. and Kruskal, W. H. (1959). Measures of association for cross classifications. II: Further discussion and references. *J. Am. Stat. Assoc.,* 54, 123–163.
Goodman, L. A. and Kruskal, W. H. (1963). Measures of association for cross classifications. III: Approximate sampling theory. *J. Am. Stat. Assoc.,* 58, 310–364.
Goodman, L. A. and Kruskal, W. H. (1972). Measures of association for cross classifications. IV. Simplification of asymptotic variances. *J. Am. Stat. Assoc.,* 67, 415–421.
Grizzle, J. E. (1961). A new method of testing hypotheses and estimating parameters for the logistic model. *Biometrics,* 17, 372–385.
Grizzle, J. E. (1963). Tests of linear hypotheses when the data are proportions. *Am. J. Public Health,* 53, 970–976.
Haldane, J. B. S. (1956). The estimation and significance of the logarithm of a ratio of frequencies. *Ann. Hum. Genet.,* 20, 309–311.
Harman, H. H. (1960). *Modern factor analysis.* Chicago: University of Chicago Press.
Русский перевод: Харман Г. Современный факторный анализ. — М.: Статистика, 1972.
Levin, M. L. (1953). The occurrence of lung cancer in man. *Acta Unio Int. Contra Cancrum,* 19, 531–541.
Lord, F. M. and Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley.
Markush, R. E. (1977). Levin's attributable risk statistic for analytic studies and vital statistics. *Am. J. Epidemiol.,* 105, 401–406.
Maxwell, A. E. and Everitt, B. S. (1970). The analysis of categorical data using a transformation. *Brit. J. Math. Stat. Psychol.,* 23, 177–187.
Mosteller, F. (1968). Association and estimation in contingency tables. *J. Am. Stat. Assoc.,* 63, 1–28.
Nunnally, J. (1978). *Psychometric theory,* 2nd ed. New York: McGraw-Hill.
Stevens, W. L. (1951). Mean and variance of an entry in a contingency table. *Biometrika,* 38, 468–470.
Thomas, D. G. and Gart, J. J. (1977). A table of exact confidence limits for differences and ratios of two proportions and their odds ratios. *J. Am. Stat. Assoc.,* 72, 73–76.
Walter, S. D. (1976). The estimation and interpretation of attributable risk in health research. *Biometrics,* 32, 829–849.
Winsor, C. P. (1948). Factorial analysis of a multiple dichotomy. *Hum. Biol.,* 20, 195–204.
Woolf, B. (1955). On estimating the relation between blood group and disease. *Ann Hum. Genet.,* 19, 251–253.
Yule, G. U. (1900). On the association of attributes in statistics. *Philos. Trans. R. Soc. Ser. A,* 194, 257–319.
Yule, G. U. (1912). On the methods of measuring the association between two attributes. *J. R. Stat. Soc.,* 75, 579–642.

Глава 6

Метод выбора II. Проспективные и ретроспективные исследования

При методе II извлечения выборок, согласно определению из разд. 2.1, выборку строят по $n_1 + n_2$ объектам, извлекаемым из двух популяций, где n_1 — заданное число объектов из первой популяции и n_2 — заданное число объектов из второй популяции. Этот метод применяется в сравнительных проспективных исследованиях (когда одна популяция определяется наличием, а вторая — отсутствием исследуемого исходного фактора [MacMahon and Pugh, 1970, Chapter 11]) и в сравнительных ретроспективных исследованиях (когда популяции определяются соответственно наличием и отсутствием исследуемого результирующего фактора [MacMahon and Pugh, 1970, Chapter 12]).

В разд. 6.1 и 6.2 обсуждаются проспективный и ретроспективный анализы данных при сравнительных исследованиях. В разд. 6.3 представлена критика отношения шансов, данная Берксоном, Шепс и Файнстайном. В разд. 6.4 описано построение выводов о привносимом риске при ретроспективных исследованиях. Сравнение ретроспективного и проспективного подходов проведено в разд. 6.5.

6.1. Проспективные исследования

При проведении проспективного сравнительного исследования, называемого также когортным (cohort, forward-going, follow-up), выделяют две выборки, основываясь на наличии или отсутствии некоторого исходного фактора, и оценивают в каждой выборке пропорцию числа объектов, для которых выполняется результирующее условие (например, пропорцию числа заболевших).

Вновь обратимся к примеру из гл. 5, в котором рассматривается связь возраста матери (исходный фактор) с весом новорожденного (результирующий фактор). Чтобы провести исследование по сравнительной проспективной схеме, надо, например, по двум раздельным спискам, один из которых содержит записи о рождениях детей у матерей не старше 20 лет, а другой — у матерей старше 20 лет, построить две независимые друг от друга выборки. Предположим, что мы сформировали выборки объема 100 для каждой из двух возрастных групп и что при рождении зарегистрирован вес каждого ребенка.

Конечно, маловероятно, что результат точно совпадает с результатом для метода выбора I (см. разд. 5.1). Тем не менее для наглядности мы предположим, что наблюдается полное соответствие пропорций. Доля новорожденных с низким весом у матерей не старше 20 лет по табл. 5.2 составляет

$$p(B \mid A) = \frac{p_{11}}{p_{1.}} = \frac{0{,}05}{0{,}25} = 0{,}20 . \qquad (6.1)$$

Значит, в 20% случаев, т. е. у 20 из 100 молодых матерей, вес новорожденного составил 2500 г или меньше, а у остальных — больше 2500 г.

Для матерей старше 20 лет соответствующее значение равно:

$$p(B \mid \overline{A}) = \frac{p_{21}}{p_{2.}} = \frac{0{,}075}{0{,}75} = 0{,}10 . \qquad (6.2)$$

Следовательно, у 10 из 100 матерей старше 20 лет вес ребенка оставил 2500 г или меньше, а у остальных — больше 2500 г. Итак, полученные данные приводятся в виде табл. 6.1.

Таблица 6.1

Зависимость веса новорожденного от возраста матери. Проспективное исследование

Возраст матери	Вес новорожденного		Сумма	Пропорция числа детей с низким весом
	B	$\overline{B}$		
A	20	80	$N_A = 100$	$p(B \mid A) = 0{,}2$
$\overline{A}$	10	90	$N_{\overline{A}} = 100$	$p(B \mid \overline{A}) = 0{,}1$
Сумма	30	170	200	

Вычисление χ^2 для этих данных дает

$$\chi^2 = 3{,}18, \tag{6.3}$$

т. е. зависимость между факторами A и B незначима при уровне 0,05. Следует обратить внимание, что значение (6.3) больше значения χ^2 для табл. 5.1, хотя суммарные объемы выборок (200) равны, и табл. 5.1 и 6.1 согласуются друг с другом по частотам. Этот результат — частный случай общего правила: при равных суммарных объемах выборок проспективное исследование с выборками из A и $\bar{A}$ равного объема приводит к более *мощному* критерию хи-квадрат, чем перекрестное исследование [Lehmann, 1959, p. 146].

В разд. 5.3 отношение шансов ω рассматривается как мера связи между факторами A и B. Поскольку шансы Ω_A и $\Omega_{\bar{A}}$ определяются раздельно в (5.5)—(5.8), легко понять, что оценивать отношение шансов можно как в перекрестном, так и в проспективном исследовании. Оценка отношения шансов равна:

$$o = \frac{p(B \mid A)\, p(\bar{B} \mid \bar{A})}{p(\bar{B} \mid A)\, p(B \mid \bar{A})}. \tag{6.4}$$

Подставив в эту формулу данные из табл. 6.1, получаем:

$$o = \frac{0{,}20 \cdot 0{,}90}{0{,}80 \cdot 0{,}10} = 2{,}25, \tag{6.5}$$

что совпадает с оценкой (5.15) в перекрестном исследовании.

Выражение (5.16) применимо и для данных, найденных в сравнительном проспективном исследовании. Очевидно, значение (5.16) также равно 2,25.

Оценкой стандартной ошибки вычисляемого отношения шансов в проспективном исследовании является

$$\text{s. e. } (o) = o \sqrt{\frac{1}{N_A\, p(B \mid A)\, p(\bar{B} \mid A)} + \frac{1}{N_{\bar{A}}\, p(B \mid \bar{A})\, p(\bar{B} \mid \bar{A})}}. \tag{6.6}$$

С помощью данных табл. 6.1 определяем оценку стандартной ошибки:

$$\text{s. e. } (o) = 2{,}25 \sqrt{\frac{1}{100 \cdot 0{,}2 \cdot 0{,}8} + \frac{1}{100 \cdot 0{,}1 \cdot 0{,}9}} = 0{,}94. \tag{6.7}$$

Эквивалентное выражение для оценки стандартной ошибки можно получить через наблюдаемые частоты (см. задачу 6.1):

$$\text{s. e. } (o) = o \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}, \tag{6.8}$$

что совпадает с (5.19).

Выше мы указали, что критерий хи-квадрат, построенный по данным с равными объемами выборок, полученных в сравнительном проспективном исследовании, мощнее критериев по данным, полученным в перекрестном исследовании. Аналогичное утверждение справедливо и для *точности* оценки отношения шансов. Суммарный объем выборок в табл. 5.1 и 6.1 одинаков. Однако оценка отношения шансов в первом случае (s. e. $(o)=1{,}00$; см. (5.18)) менее точна, чем во втором (s. e. $(o)=0{,}94$; см. (6.7)). Таким образом, если суммарные объемы выборок равны, то проспективное исследование с равными объемами выборок ($N_A=N_{\bar{A}}$) предпочтительнее перекрестного исследования как по мощности критериев, так и по точности оценок в среднеквадратичном.

Воспользовавшись методами из разд. 5.6, можно рассчитать приближенный 95%-ный доверительный интервал для отношения шансов, построенный по данным табл. 6.1, —

$$0{,}93 \leqslant \omega \leqslant 5{,}51; \tag{6.9}$$

он уже (т. е. лучше) соответствующего интервала (5.73) для перекрестного исследования.

Используя данные табл. 6.1, находим значение статистики хи-квадрат без поправок (см. (5.3)):

$$\chi^2_u = \frac{200\cdot(20\cdot 90 - 80\cdot 10)^2}{100\cdot 100\cdot 30\cdot 70} = 3{,}92\,.$$

Этой величине соответствует значение фи-коэффициента (5.2):

$$\varphi = \sqrt{\frac{3{,}92}{200}} = 0{,}14, \tag{6.10}$$

что немного больше значения $\varphi=0{,}13$, полученного в (5.4) по данным табл. 5.1, хотя табл. 5.1 и 6.1 совпадают по пропорциям (6.1) и (6.2).

6.2. Ретроспективные исследования

При проведении ретроспективного сравнительного исследования, называемого также исследованием с опытной и контрольной группами (case-control study), выделяют две выборки, основываясь на наличии или отсутствии результирующего фактора, и оценивают в каждой выборке пропорцию объектов, у которых есть исследуемый исходный фактор.

Исследование зависимости между весом новорожденного и возрастом матери можно провести и по ретроспективной схеме. Для этого следует, например, по двум спискам, один из которых содержит записи о новорожденных с низким, а другой — с нормальным весом, построить две независимые друг от друга выборки. Предположим, что число объектов в каждой вы-

борке равно 100 и что в сведения о новорожденном записывается возраст матери.

Как и в разд. 6.1, допустим, что собранные данные полностью соответствуют результатам, полученным ранее. Исходя из этого условия определим доли $p(A|B)$ и $p(A|\bar{B})$, т. е. пропорции числа матерей не старше 20 лет в группах детей с низким и нормальным весом. Воспользовавшись данными табл. 5.2, получаем

$$p(A|B) = \frac{p_{11}}{p_{.1}} = \frac{0,05}{0,125} = 0,40 . \qquad (6.11)$$

Это означает, что у 40 из 100 детей с низким весом возраст матери не более 20 лет, а у остальных — более 20 лет. Аналогично

$$p(A|\bar{B}) = \frac{p_{12}}{p_{.2}} = \frac{0,20}{0,875} = 0,23 , \qquad (6.12)$$

т. е. у 23 из 100 детей с нормальным весом возраст матери — не более 20 лет, а у остальных — более 20 лет.

Данные, полученные методом выбора II, обычно представляют в виде табл. 2.6, т. е. данные одной выборки располагают под данными другой выборки, а обозначения для результирующего фактора, характеризующего объект, ставят сверху. При ретроспективном исследовании это может привести к недоразумению, так как теперь результирующим фактором формально будет исходный фактор. В [Miettinen, 1970] подчеркивается, что логически результирующий фактор следует после исходного, но при анализе данных в ретроспективном исследовании порядок обратный.

Представим данные, полученные методом выбора II, в виде табл. 6.2. Значение χ^2 для них равно:

$$\chi^2 = 5,93 , \qquad (6.13)$$

что указывает на наличие зависимости при уровне значимости 0,05.

Таблица 6.2

Зависимость веса новорожденного от возраста матери. Ретроспективное исследование

Вес новорожденного	Возраст матери		Сумма	Пропорция числа матерей не старше 20 лет
	A	$\bar{A}$		
B	40	60	$N_B = 100$	$p(A\|B) = 0,40$
$\bar{B}$	23	77	$N_{\bar{B}} = 100$	$p(A\|\bar{B}) = 0,23$
Сумма	63	137	200	

Следует обратить внимание на увеличение χ^2 (2,58 — в перекрестном исследовании, 3,18 — проспективном и 5,93 — в ретроспективном), несмотря на то, что данные выбраны так, чтобы определяющие условные вероятности и суммарные объемы выборок во всех трех случаях совпадали. В том случае, когда суммарные размеры выборок равны, ретроспективное исследование приводит к более мощному критерию хи-квадрат, чем перекрестное исследование. Если же одна из категорий B или $\bar{B}$ результирующего фактора более редкая, чем категории A и $\bar{A}$ исходного фактора, т. е. если

$$|P(B) - 0{,}5| > |P(A) - 0{,}5|, \tag{6.14}$$

то ретроспективное исследование с выборками равного объема из B и $\bar{B}$ эффективнее по мощности критериев проспективного исследования с выборками равного объема из A и $\bar{A}$ [Lehmann, 1959, p. 146].

С помощью данных из табл. 6.2, найдем значение статистики хи-квадрат без поправок

$$\chi^2_u = \frac{200\,(40 \cdot 77 - 60 \cdot 23)^2}{100 \cdot 100 \cdot 63 \cdot 137} = 6{,}70\,.$$

Соответствующее значение фи-коэффициента равно:

$$\varphi = \sqrt{\frac{6{,}70}{200}} = 0{,}18\,, \tag{6.15}$$

что на 40% больше значения $\varphi = 0{,}13$ для данных табл. 5.1 и на 30% больше значения $\varphi = 0{,}14$ для табл. 6.1. Однако если мы хотим, чтобы некоторая мера была чем-либо бóльшим, чем просто неинтерпретируемым индексом, то она должна быть *инвариантна,* т. е. оценки, получаемые при различных схемах исследования, должны быть по меньшей мере близки. Поскольку фи-коэффициент, очевидно, не инвариантен, его не следует использовать как меру связи в проспективных и ретроспективных исследованиях.

Напротив, отношение шансов ω инвариантно ко всем трем типам исследований. Если определить его как

$$\omega = \frac{P(B|A)\,P(\bar{B}|\bar{A})}{P(\bar{B}|A)\,P(B|\bar{A})}\,, \tag{6.16}$$

то может показаться, что оценивание ω осуществимо только в перекрестном, либо сравнительном проспективном исследовании, так как только в этих случаях обеспечены оценки $P(B|A)$ и $P(B|\bar{A})$. Однако выражение

$$\omega = \frac{P(A|B)\,P(\bar{A}|\bar{B})}{P(\bar{A}|B)\,P(A|\bar{B})} \tag{6.17}$$

Т Зак. 2383

эквивалентно (6.16) (см. задачу 6.2). Ясно, что оно позволяет оценивать ω и в ретроспективных исследованиях [Cornfield, 1956]. Оценкой ω при этом является

$$o = \frac{p(A|B)\, p(\bar{A}|\bar{B})}{p(\bar{A}|B)\, p(A|\bar{B})} . \qquad (6.18)$$

Используя данные табл. 6.2, получаем оценку отношения шансов

$$o = \frac{0{,}40 \cdot 0{,}77}{0{,}60 \cdot 0{,}23} = 2{,}23 , \qquad (6.19)$$

что совпадает со значением $o = 2{,}25$, найденным ранее, с точностью до ошибок округления. Как и прежде, отношение шансов можно вычислять через частоты по (5.16).

Стандартную ошибку оценки отношения шансов в сравнительном ретроспективном исследовании можно оценить с помощью величины

$$\text{s. e. } (o) = o \sqrt{\frac{1}{N_B\, p(A|B)\, p(\bar{A}|B)} + \frac{1}{N_{\bar{B}}\, p(A|\bar{B})\, p(\bar{A}|\bar{B})}} . \qquad (6.20)$$

Формулы (5.19) и (6.8) для стандартной ошибки, выраженной через наблюдаемые частоты, верны и для ретроспективного исследования.

Подставив в (6.20) данные из табл. 6.2, получаем:

$$\text{s. e. } (o) = 2{,}23 \sqrt{\frac{1}{100 \cdot 0{,}40 \cdot 0{,}60} + \frac{1}{100 \cdot 0{,}23 \cdot 0{,}77}} =$$

$$= 2{,}23 \sqrt{\frac{1}{40} + \frac{1}{60} + \frac{1}{23} + \frac{1}{77}} = 0{,}70 . \qquad (6.21)$$

Разным значениям χ^2 в различных схемах исследования соответствует и разная точность оценки отношения шансов. Так, оценка стандартной ошибки o равна: в перекрестном исследовании — 1,00 (см. (5.18)), в проспективном — 0,94 (см. (6.7)) и в ретроспективном — 0,70 (см. 6.21)).

Приближенный 95%-ный доверительный интервал для отношения шансов, построенный по данным табл. 6.2 с помощью метода, описанного в разд. 5.6, —

$$1{,}16 \leqslant \omega \leqslant 4{,}33 . \qquad (6.22)$$

Значит, такое же различие характерно и для длины 95%-ного доверительного интервала отношения шансов: она самая большая в перекрестном (см. (5.73)), меньше — в проспективном (см. (6.9)) и наименьшая в ретроспективном исследовании (см. (6.22)).

Итак, если суммарные объемы выборок равны, то по точности оценивания, мощности критериев и длине доверительного интервала сравнительное ретроспективное исследование с равными объемами выборок из B и $\overline{B}$ эффективнее и перекрестного, и проспективного исследований [1].

6.3. Критический анализ отношения шансов

Как мы уже говорили, оценить вероятности $P(B|A)$ и $P(B|\overline{A})$ можно только в перекрестном и проспективном исследовании (A по-прежнему обозначает наличие исходного фактора, B — нежелательного результирующего фактора). Сравнивать результаты ретроспективного исследования с результатами одного из двух других видов исследования можно только в том случае, если связь определяется отношением шансов, т. е. функцией отношения этих вероятностей.

Однако Берксон [Berkson, 1958], а позже Файнстайн [Feinstein, 1973] резко критиковали использование функции *отношения* вероятностей в качестве меры связи, указывая, что при этом теряется информация о величине вероятностей. Так, десятикратное возрастание долей, равных соответственно одной миллионной и одной тысячной, дадут одинаковый результат, хотя увеличение в последнем случае намного серьезнее, чем в первом. Берксон и Файнстайн считают, что на практике, и частности в здравоохранении, правильнее в качестве меры связи использовать просто разность двух долей (вероятностей).

Поясним эту точку зрения с помощью данных табл. 26, взятой из [Smoking and Health, 1964, p. 110]. В табл. 6.3 приводится приблизительное число смертей для курящих и некурящих на 100 000 человек за год.

Если сравнивать только отношения шансов, то, на наш взгляд, курение больше влияет на смертность от рака легких, чем от ишемической болезни сердца. Именно это заключение, сделанное во многих исследованиях, не удовлетворяет Берксона. Он совершенно справедливо заявляет, что использование отношения шансов приводит к полной потере информации о числе смертей, вызванных каждой из причин. Более того,

---

[1] С точки зрения математики в обсуждаемых в разд. 6.1 и 6.2 задачах не имеет значения, какой фактор считается «исходным», а какой — «результирующим». Не следует также широко обобщать вывод этого раздела о преимуществе ретроспективного подхода. Проспективный подход будет эффективнее ретроспективного в терминах разд. 6.2, если в (6.14) будет справедливо противоположное неравенство (см. книгу Лемана, 1959, разд. 4.6). — *Примеч. пер.*

Таблица 6.3

Смертность для курящих и некурящих на 100 000 человек в год от рака легких и ишемической болезни сердца

Заболевание	Курящие	Некурящие	σ	Разность
Рак легких	48,33	4,49	10,8	43,84
Ишемическая болезнь сердца	294,67	169,54	1,7	125,13

он идет дальше и утверждает, что верно оценить зависимость причины смерти от курения позволяет *только* прирост смертности: «. . . с истинно практической точки зрения значение имеет, конечно, только общее увеличение числа смертей» [Berkson, 1958, p. 30].

По мнению Берксона, курение сильнее влияет на смертность от той причины, по которой прирост числа смертей у курящих больше. Поскольку прирост числа смертей от ишемической болезни сердца у курящих по сравнению с некурящими составляет более 120 на 100 000 человек за год, а от рака легких — менее 50, он заключает, что курение сильнее влияет на заболевание ишемической болезнью сердца, чем раком легких.

Шепс [Sheps, 1958, 1961] предложила простую и изящную модификацию индекса Берксона. Пусть p_c — смертность (или в общем случае доля объектов, для которых выполняется неблагоприятное событие) в контрольной выборке; p_s — соответствующая доля в опытной выборке, где ожидается бо́льшая смертность. Значит, предполагается, что $p_s > p_c$.

Шепс утверждает, что увеличение риска p_e в опытной группе по сравнению с контрольной проявляется только через объекты, для которых событие не произошло бы, будь они в контрольной группе. Следовательно, она постулирует модель

$$p_s = p_c + p_e\,(1 - p_c). \tag{6.23}$$

Таким образом, доля p_s в опытной группе — это сумма доли p_c в контрольной группе и прироста p_e среди тех объектов $(1-p_c)$, для которых событие не случилось бы, если бы они остались в контрольной группе. Шепс предлагает использовать p_e как меру дополнительного или избыточного риска. Поскольку, очевидно,

$$p_e = \frac{p_s - p_c}{1 - p_c}, \tag{6.24}$$

p_e можно также назвать *относительным приростом*.

Он отличается от индекса Берксона, $p_s - p_c$, только наличием выражения $1-p_c$, т. е. пропорции числа людей, в действительности подвергшихся дополнительному риску. Если p_c мало, то индексы Берксона и Шепс близки. Например, подставив в (6.24) данные табл. 6.3, находим прирост смертности от рака легких на 100 000 человек: $p_e = 43{,}84/(100\,000 - 4{,}49) = 43{,}84$, что равно числу смертей от рака легких, которые связаны исключительно с курением и которых удалось бы избежать при воздержании населения от курения. Для ишемической болезни сердца соответствующий прирост составляет: $p_e = 125{,}13/(100\,000 - 169{,}54) = 125{,}34$, что равно числу смертей на 100 000 человек, которых удалось бы избежать при воздержании от курения.

Если бы этиология болезни относилась *исключительно* к эпидемиологическим обследованиям населения, то простая разность по Берксону или относительный прирост по Шепсу были бы единственными полезными мерами связи между исследуемым фактором и результирующим событием. При этом ретроспективные исследования надо было бы считать бесполезными, так как они не дают возможности оценить значения обоих индексов. Однако, как указанно в [Cornfield et al., 1959, Greenberg, 1969], этиологическое исследование связано еще и с поиском закономерностей во многих наборах данных, с разработкой моделей причин развития и распределения заболевания по различным слоям населения и с выдвижением гипотез по тем данным, которые можно проверить на других наборах данных.

Исходя из этого можно сказать, что лучшей является та мера связи, которая вытекает из математической модели, остается справедливой в альтернативных моделях, допускает возможность выдвигать и проверять гипотезы о значениях параметров в определенных популяциях, является инвариантной по отношению к схеме исследования при изучении связи. Отношение шансов (или такая его функция, как логарифм) более чем других мер удовлетворяет этим требованиям [Cox, 1970, pp. 20—21]. С помощью ретроспективных исследований можно оценивать отношение шансов, значит, они являются мощным средством широкого научного поиска. Пикок [Peacock, 1971], однако, советует осмотрительнее относиться к предположению о том, что отношение шансов одинаково в популяциях различного вида.

Приведем пример, когда индекс Шепс, как и индекс Берксона, не позволяет выявить закономерность поведения показателей, в отличие от отношения шансов. В табл. 6.4 приведены данные о смертности среди курящих и некурящих различных возрастов, которые взяты из графы в [Smoking and Health,

1964, p. 88]. (В этой графе даны два значения для некурящих в возрасте 75—79 лет. Нами взято значение, которое выглядит более разумным.)

Таблица 6.4

Доли смертей по всем причинам на 100 000 человек в зависимости от возраста и пристрастия к курению

Возрастной интервал	Курящие	Некурящие	o	p_e на 100 000 чел.
45—49	580	270	2,2	310
50—54	1050	440	2,4	610
55—59	1600	850	1,9	750
60—64	2500	1500	1,7	1000
65—69	3700	2000	1,9	1700
70—74	5300	3000	1,8	2400
75—79	9200	4800	2,0	4600

Как это часто бывает, мы придем к различным выводам в зависимости от используемой меры связи. Так, судя по величине p_e, влияние курения на смертность с возрастом постепенно усиливается. По отношению к числу смертей, которых можно избежать при воздержании от курения, это заключение справедливо. Однако увеличение p_e нерегулярно, что исключает сколько-нибудь точную математическую экстраполяцию и даже интерполяцию.

С другой стороны, судя по величине o, курение практически одинаково влияет на смертность во всех возрастных категориях. С эпидемиологической точки зрения это заключение важно, так как позволяет вполне обосованно предполагать, что отношение шансов составит приблизительно 2,0 в следующих возрастных группах: от 45 до 79 лет, до 45 лет, от 79 лет и старше. Если результаты наблюдений покажут отклонение от этой интерполяции или экстраполяции по возрасту, то следует предпринять дальнейшее исследование.

6.4. Оценивание привносимого риска в ретроспективном исследовании

Привносимый риск R_A по [Levin, 1953] определялся в разд. 5.7 как доля объектов с наступившим результирующим событием, для которых наступление события обусловлено

фактором риска. В соответствии с определением (5.76) и утверждениями из разд. 5.6, оценить R_A позволяет только перекрестное исследование, т. е. когда можно одновременно оценить и $P(A)$ — долю популяции, подверженную фактору риска, и R относительный риск. Однако в [Levin, 1953; Walter, 1975, 1976; Taylor, 1977] указано, что при некоторых предположениях R_A можно оценить и в ретроспективном исследовании если $P(B)$ — вероятность результирующего события в популяции — мала, то отношение шансов, o, является оценкой R. Если при этом контрольная группа $\bar{B}$ — случайная выборка из соответствующей части популяции, то оценкой $P(A)$ является $p(A|B)$.

В этих предположениях[1]

$$r_A = \frac{p(A|\bar{B})(o-1)}{1+p(A|\bar{B})(o-1)} = \frac{p(A|B)-p(A|\bar{B})}{1-p(A|\bar{B})} \qquad (6.25)$$

хорошая оценка привносимого риска в популяции [Levin and Bertell, 1978]. Подставив в (6.25) данные табл. 6.2, найдем оценку риска малого веса новорожденного, привносимого фактором молодости матери:

$$r_A = \frac{0{,}4-0{,}23}{1-0{,}23} = 0{,}22\,. \qquad (6.26)$$

При перекрестном исследовании (см. табл. 5.2) $p(A)=0{,}25$ и относительный риск $r=2{,}0$, значит,

$$r_A = \frac{0{,}25\cdot(2{,}0-1)}{1+0{,}25\cdot(2{,}0-1)} = 0{,}20\,, \qquad (6.27)$$

что весьма близко к (6.26) для ретроспективного исследования. В обоих случаях вывод состоит в том, что если женщины в возрасте 20 лет и меньше откажутся от рождения детей, то можно избежать около 1/5 всех рождений детей с низким весом.

Уолтер [Walter, 1975] показал, что при оценивании привносимого риска в ретроспективном исследовании по формуле (6.25) оценкой стандартной ошибки логарифма, $1-r_A$, является

$$s.e.[\ln(1-r_A)] = \sqrt{\frac{p(A|B)}{N_B\,p(\bar{A}|B)} + \frac{p(A|\bar{B})}{N_{\bar{B}}\,p(\bar{A}|\bar{B})}}\,, \qquad (6.28)$$

---

[1] Как нетрудно заметить, правая часть (6.25) совпадает с (6.24).— *Примеч. пер.*

где $N_{\bar{B}}$ — число контрольных объектов; N_B — число опытных объектов. Пользуясь данными табл. 6.2, получаем:

$$\ln\,(1-r_A) = \ln\,(1-0{,}22) = -0{,}25, \tag{6.29}$$

$$\text{s. e.}\,[\ln(1-r_A)] = \sqrt{\frac{0{,}4}{100\cdot 0{,}6} + \frac{0{,}23}{100\cdot 0{,}77}} = 0{,}1\,. \tag{6.30}$$

Приближенный 95%-ный доверительный интервал для $\ln(1-R_A)$ составит

$$-0{,}25 - 1{,}96\cdot 0{,}1 \leqslant \ln\,(1-R_A) \leqslant -0{,}25 + 1{,}96\cdot 0{,}1\,, \tag{6.31}$$

или

$$-0{,}45 \leqslant \ln\,(1-R_A) \leqslant -0{,}05\,. \tag{6.32}$$

В итоге получаем, что доверительный интервал с уровнем доверия около 95% для привносимого риска —

$$0{,}05 \leqslant R_A \leqslant 0{,}36\,. \tag{6.33}$$

6.5. Сравнение ретроспективного и проспективного подходов

Даже если исследователь принимает доводы разд. 6.3 в пользу правомерности применения отношения шансов и, следовательно, ретроспективных исследований, он должен сознавать, что ретроспективные исследования больше подвержены ошибкам, чем проспективные. Например, Хэммонд [Hammond, 1958] считает, что при анализе исторических данных о заболевании могут появиться смещения, поскольку данные собирались только после появления признаков болезни, а часто даже после установления диагноза. Когда пациенту известен диагноз, он может исказить (намеренно или нет) описание факторов, предшествующих болезни.

Другая трудность, указанная Хэммондом, состоит в поиске контрольной группы, адекватной выборке больных, поскольку надо найти группу (группы), сходную с опытной выборкой во всех отношениях. Исключение лишь в том, что объекты контрольной группы не страдают данным заболеванием. Этот и другие недостатки ретроспективного подхода указаны в [Mantel and Haenszel, 1959]. Например, если данные по больным собирают в больницах или клиниках, то ретроспективное исследование может привести к смещению, рассмотренному в разд. 1.3: мы сочтем, что исходный фактор связан с заболе-

нанием, хотя в действительности он связан сильнее с назначенным лечением. Файнстайн [Feinstein, 1973] описывает смещение, очень близкое по характеру к указанному, возникающее за счет различной выживаемости пациентов с фактором риска и без него.

Сказанное не означает, что только ретроспективные исследования подвержены смещениям. Подобные явления наблюдаются и при проспективном анализе [Mainland and Herrera, 1956; Yerushalmy and Palmer, 1959; Mantel and Haenszel, 1959; Greenland, 1977]. Например, в проспективных исследованиях при проведении наблюдения над добровольцами, согласившимися принять участие в опыте, возможно смещение, аналогичное смещению при ретроспективном исследовании по данным о госпитализированных пациентах. Другими проблемами проспективного подхода являются ошибки в диагнозе и недостаточно частые осмотры больных [Schlesselman, 1977].

Что действительно кажется верным, так это то, что ретроспективный подход требует при планировании исследования большей изобретательности, чем проспективный. Например, первое из описанных смещений вызвано тем, что на ответы пациента, возможно, влияет его знание о своей болезни (изучаемой нами). Для устранения этого смещения Левин [Levin, 1951] предложил опрашивать всех больных до постановки окончательного диагноза. Таким способом можно избежать и смещения, вызываемого тем, что объекты из опытной и контрольной групп опрашиваются по-разному. В [Rimmer and Chambers, 1969] утверждается, что более точным методом является опрос не больных, а их родственников.

Долл и Хилл [Doll and Hill, 1952] применяли две контрольные группы, чтобы избежать смещения, возможного при одной контрольной группе. Первую группу составила выборка госпитализированных больных с заболеваниями, отличными от исследуемого, вторую — выборка из населения. Другим способом борьбы со смещениями является связывание (см. гл. 8). По мере того как исследователь определяет, какую информацию объект сообщает верно, а какую — нет, точность ретроспективного подхода может только возрастать. Например, в [Colton, 1955; Klemetti and Saxén, 1967] показано, что обычно правильно сообщают, произошло событие или нет, но не точное время, когда это событие произошло.

Джик и Висси [Jick and Vessey, 1978] описали основные причины ошибок в тех ретроспективных исследованиях, в которых выясняется роль применения прописанных лекарств в развитии заболевания, и указали способы контроля описанных ошибок. «Journal of Chronic Diseases» за январь 1979 г. [Buchan, 1979] посвящен современному состоянию ретроспек-

тивных исследований. Статья [Sackett, 1979], помещенная в этом номере, особенно полезна данным в ней перечнем причин смещений и рекомендациями по их устранению или измерению.

В гл. 12 мы вернемся к обсуждению предупреждения смещений. Учитывая, что смещения поддаются контролю, следует принять убедительную аргументацию Мантела и Ханзела в пользу ретроспективного подхода: «Среди привлекательных свойств, присущих ретроспективному подходу, можно перечислить: возможность получать результаты по данным, доступным в настоящем... Ретроспективное исследование может проводиться ограниченными средствами, которые находятся в распоряжении отдельного ученого... Для очень редких заболеваний ретроспективный подход может быть единственным возможным подходом... Ретроспективный подход в отсутствии серьезных смещений при проведении исследования можно считать, согласно строгой статистической теории, предпочтительным» [Mantel and Haenszel, 1959, p. 720].

Задачи

6.1. Докажите равенство выражений (6.6) и (6.8) стандартной ошибки отношения шансов (*Указание.* Выразите оценки условных вероятностей в (6.6) через наблюдаемые частоты.)

6.2. Докажите равенство (6.16) и (6.17). (*Указание.* Замените все условные вероятности в (6.16) совместными вероятностями с помощью определений из разд. 1.1.)

6.3. Правильное использование фи-коэффициента как меры связи возможно только в методе выбора I (перекрестном). Его значение, вычисленное по данным, полученным методом выбора II (проспективно или ретроспективно), нельзя сравнивать со значением для данных, полученных методом I.

Более того, в общем случае нельзя сравнивать два значения фи-коэффициента, если оба получены в двух проспективных (или ретроспективных) исследованиях с различным числом объектов по категориям отбора, даже если суммарные объемы выборок совпадают.

а) Ретроспективно изучая факторы, связанные с раком полости рта, Уиндер и др. [Wynder, et al., 1958] обследовали 34 больных женщины и 214 женщин, соответствующих им по возрасту и не больных раком полости рта. В группе раковых больных не курили 24% женщин в отличие от 66% в контрольной. Составьте таблицу и вычислите статистику хи-квадрат без поправок и соответствующий фи-коэффициент.

б) Предположим, что Уиндер и др. обследовали теперь 214 больных и 34 здоровых женщин. Предполагая, что пропорции некурящих те же, что и в а), вычислите значения статистики хи-квадрат по новой таблице. Сравните значения фи-коэффициентов.

в) Предположим теперь, что для опытной и контрольной групп обследовано по 124 женщины и что пропорции некурящих среди больных и не больных раком остались прежними. Вычислите соответствующие новым данным значения. Сравните все три фи-коэффициента. Как вы считаете, можно ли сравнивать фи-коэффициенты, полученные в ретроспективных исследованиях, с разными распределениями объектов по контрольной и опытной группам?

ıве при сравнении перекрестного подхода с проспективным
ам подходами применялись три критерия. Еще одним кри-
уммарный объем выборки, требуемый, чтобы стандартная
ания шансов принимала определенное значение. Используем
меров в гл. 5 и 6. Предположим, что истинное значение o=
ужные пропорции известны.

..иженная стандартная ошибка в перекрестном исследовании да-
Какое требуется значение $n_{..}$, чтобы стандартная ошибка была
но равна 0,50?

бозначим суммарный объем выборок в проспективном исследовании
$=N_A+N_{\overline{A}}$. Для простоты допустим, что $N_A=N_{\overline{A}}=N_P/2$. Вычислите
щью (6.6), какое требуется значение N_P, чтобы стандартная ошибка
приближенно равна 0,50. Вычислите процентное уменьшение N_P по
нию с $n_{..}$.

Обозначим суммарный объем выборок в ретроспективном исследова-
ак $N_R=N_B+N_{\overline{B}}$. Допустим, что $N_{\overline{B}}=N_B=N_R/2$. Вычислите по (6.20),
ое требуется N_R, чтобы стандартная ошибка o была приближенно равна
0. Вычислите процентное умеиьшение N_R по сравнению с $n_{..}$; N_R по срав-
нению с N_P.

г) Сравните требуемые объемы выборок. Насколько практичнее (т. е.
ешевле) б) и в) по сравиению с а)?

ЛИТЕРАТУРА

Berkson, J. (1958). Smoking and lung cancer: Some observations on two recent reports. *J. Am. Stat. Assoc.*, **53**, 28–38.

Cornfield, J. (1956). A statistical problem arising from retrospective studies. Pp. 135–148 in J. Neyman (Ed.). *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, Vol. 4. Berkeley: University of California Press.

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., and Wynder, E. L. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *J. Natl. Cancer Inst.*, **22**, 173–203.

Cox, D. R. (1970). *Analysis of binary data*. London: Methuen.

Doll, R. and Hill, A. B. (1952). A study of the etiology of carcinoma of the lung. *Brit. Med. J.*, **2**, 1271–1286.

Feinstein, A. R. (1973). Clinical biostatistics XX. The epidemiologic trohoc, the ablative risk ratio, and retrospective research. *Clin. Pharmacol. Ther.*, **14**, 291–307.

Gray, P. G. (1955). The memory factor in social surveys. *J. Am. Stat. Assoc.*, **50**, 344–363.

Greenberg, B. G. (1969). Problems of statistical inference in health with special reference to the cigarette smoking and lung cancer controversy. *J. Am. Stat. Assoc.*, **64**, 739–758.

Greenland, S. (1977). Response and follow-up bias in cohort studies. *Am. J. Epidemiol.*, **106**, 184–187.

Hammond, E. C. (1958). Smoking and death rates: A riddle in cause and effect. *Am. Sci.*, **46**, 331–354.

Ibrahim, M. A. (Ed.) (1979). The case-control study: Consensus and controversy. *J. Chronic Dis.*, **32**, 1–144.

Jick, H. and Vessey, M. P. (1978). Case-control studies in the evaluation of drug-induced illness. *Am. J. Epidemiol.*, **107**, 1–7.

Klemetti, A. and Saxen, L. (1967). Prospective versus retrospective approach in the search for environmental causes of malformations. *Am. J. Public Health*, **57**, 2071–2075.

Lehmann, E. L. (1959). *Testing statistical hypotheses*. New York: Wiley.

Русский перевод: Леман Э.А.
Проверка статистических гипотез. – 2 изд. – М.: Наука, 1979.

Levin, M. L. (1953). The occurrence of lung cancer in man. *Acta Unio Int. Contra Cancrum*, **19**, 531–541.

Levin, M. L. (1954). Etiology of lung cancer: Present status. *N. Y. State J. Med.*, **54**, 769–777.

Levin, M. L. and Bertell, R. (1978). Re: "Simple estimation of population attributable risk from case-control studies." *Am. J. Epidemiol.*, **108**, 78–79.

MacMahon, B. and Pugh, T. F. (1970). *Epidemiology: Principles and methods*. Boston: Little, Brown.

Mainland, D. and Herrera, L. (1956). The risk of biased selection in forward-going surveys with nonprofessional interviewers. *J. Chronic Dis.*, **4**, 240–244.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.*, **22**, 719–748.

Miettinen, O. S. (1970). Matching and desing efficiency in retrospective studies. *Am. J. Epidemiol.*, **91**, 111–118.

Peacock, P. B. (1971). The noncomparability of relative risks from different studies. *Biometrics*, **27**, 903–907.

Rimmer, J. and Chambers, D. S. (1969). Alcoholism: Methodological considerations in the study of family illness. *Am. J. Orthopsychiatry*, **39**, 760–768.

Sackett, D. L. (1979). Bias in analytic research. *J. Chronic Dis.*, **32**, 51–63.

Schlesselman, J. J. (1977). The effect of errors of diagnosis and frequency of examination on reported rates of disease. *Biometrics*, **33**, 635–642.

Sheps, M. C. (1958). Shall we count the living or the dead? *New Engl. J. Med.*, **259**, 1210–1214.

Sheps, M. C. (1961). Marriage and mortality. *Am. J. Public Health*, **51**, 547–555.

Smoking and Health (1964). Report of the Advisory Committee to the Surgeon General of the Public Health Service. Princeton: Van Nostrand.

Taylor, J. W. (1977). Simple estimation of population attributable risk from case-control studies. *Am. J. Epidemiol.*, **106**, 260.

Walter, S. D. (1975). The distribution of Levin's measure of attributable risk. *Biometrika*, **62**, 371–374.

Walter, S. D. (1976). The estimation and interpretation of attributable risk in health research. *Biometrics*, **32**, 829–849.

Wynder, E. L., Navarrette, A., Arostegui, G. E., and Llambes, J. L. (1958). Study of environmental factors in cancer of the respiratory tract in Cuba. *J. Natl. Cancer Inst.*, **20**, 665–673.

Yerushalmy, J. and Palmer, C. E. (1959). On the methodology of investigations of etiologic factors in chronic diseases. *J. Chronic Dis.*, **10**, 27–40.

Weinstein, M. C. (1974). Allocation of subjects in medical experiments. *New Engl. J. Med.*, **291**, 1278–1285.

Zelen, M. (1969). Play the winner rule and the controlled clinical trial. *J. Am. Stat. Assoc.*, **64**, 131–146.

Zelen, M. (1979). A new design for randomized clinical trials. *New Engl. J. Med.*, **300**, 1242–1245.

Глава 7

Метод выбора III. Планируемые сравнительные испытания

Метод III извлечения выборок применяют в сравнительных клинических испытаниях. При этом сравниваемые методы лечения случайно назначают объектам (пациентам). Общие вопросы сравнительных клинических испытаний обсуждаются в [Hill, 1962, Chapters 1—3], а в [Mainland, 1960] и выпуске «Clinical Pharmacology and Therapeutics» [Roth and Gordon, 1979] предлагаются решения некоторых практических задач, возникающих при проведении клинических испытаний. Этические проблемы рассматриваются в [Fox, 1959; Meier, 1975]. Принципы определения числа пациентов, требуемого в исследовании, и некоторые неприятные последствия выбора слишком малого числа пациентов описаны в [Frieman et al., 1978].

Спорной особенностью при клинических испытаниях является возможность преждевременно прерывать испытания, если в одной из сравниваемых групп наблюдается высокая доля серьезных неблагоприятных реакций, вызывающая тревогу, или если выясняется, что преимущества одного из средств терапевтического лечения над другим настолько велики, что отказывать кому-либо из пациентов в лечении этим средством просто неэтично. В [Report of the Committee for the Assessment of Biometric Aspects of Controlled Trials of Hypoglycemic Agents, 1975] приведен интересный пример прерывания эксперимента по первой причине, а в [Anturane Reinfarction Trial Research Group, 1978] — по второй причине. Мейер [Meier, 1979] предлагает несколько правил, по которым можно принимать решение либо о прекращении испытаний, либо об их продолжении.

В этой главе вы познакомитесь со сравнением лечений только двух видов. Описание анализа данных в простых сравнительных испытаниях дается в разд. 7.1, а в разд. 7.2 рассматривается план с перекрытием. В обоих разделах обсуждается случай, когда градациями результирующего фактора являются

«да» и «нет» (например «выздоровел» и «не выздоровел»). Некоторые альтернативы простой рандомизации, предлагаемые для сравнительных испытаний, обсуждаются в разд. 7.3.

7.1. Простые сравнительные испытания

Предположим, что данные табл. 7.1 получены в результате испытаний, в которых один вид лечения применялся к $n_1 = 80$ объектам, случайно выбранным из группы в 150 больных, другой — к остальным $n_2 = 70$ объектам.

Таблица 7.1

Вымышленные данные сравнительных клинических испытаний

	Число пациентов	Пропорция пациентов, у которых наблюдалось улучшение
Лечение 1	$n_1 = 80$	$p_1 = 0{,}60$
Лечение 2	$n_2 = 70$	$p_2 = 0{,}80$
Всего	$n = 150$	$\bar{p} = 0{,}69$

Проверим статистическую значимость различия двух долей числа объектов, у которых наступило улучшение, с помощью статистики (2.5). Ее значение по данным табл. 7.1 равно:

$$z = \frac{|0{,}80 - 0{,}60| - \frac{1}{2}\left(\frac{1}{80} + \frac{1}{70}\right)}{\sqrt{0{,}69 \cdot 0{,}31\left(\frac{1}{80} + \frac{1}{70}\right)}} = 2{,}47\,, \qquad (7.1)$$

что говорит о значимости различия на уровне 0,05.

Чаще всего в качестве меры отличия эффективности одного вида лечения от другого используют простую разность долей объектов, у которых наступило улучшение

$$d = p_2 - p_1\,. \qquad (7.2)$$

Приблизительная стандартная ошибка d равна:

$$\text{s. e. } (d) = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}\,. \qquad (7.3)$$

Подставив в (7.2) данные табл. 7.1, находим

$$d = 0{,}80 - 0{,}60 = 0{,}20\,, \tag{7.4}$$

из чего следует, что при лечении 100 пациентов первым способом число пациентов, у которых наступит улучшение, примерно на 20 больше, чем при лечении их вторым способом. Оценкой стандартной ошибки d является

$$\text{s. e. } (d) = \sqrt{\frac{0{,}60 \cdot 0{,}40}{80} + \frac{0{,}80 \cdot 0{,}20}{70}} = 0{,}07\,. \tag{7.5}$$

Приближенный 95%-ный доверительный интервал для разности пропорций составляет:

$$0{,}20 - 1{,}96 \cdot 0{,}07 \leqslant P_2 - P_1 \leqslant 0{,}20 + 1{,}96 \cdot 0{,}07, \tag{7.6}$$

или

$$0{,}06 \leqslant P_2 - P_1 \leqslant 0{,}34\,. \tag{7.7}$$

Иногда правдоподобно предположение, что пациент, положительно реагирующий на лечение первого вида, будет положительно реагировать и на лечение второго вида. Такая ситуация может встретиться, когда в лечении первого вида используют нейтральный препарат либо когда первый вид — лечение активным препаратом, а второй — лечение тем же препаратом, но с добавками или в большей дозировке. Следствием этого предположения является то, что увеличение эффективности второго вида лечения по сравнению с первым проявляется на объектах, безразличных к первому (дальнейшее обсуждение и примеры см. в [Sheps, 1958]).

Пусть P_1 — пропорция числа пациентов с улучшением в группе, получавшей первое лечение, P_2 — второе. Обозначим через f ожидаемую долю случаев улучшения от лечения вторым способом среди тех пациентов, состояние которых от лечения первым способом не улучшается. Предполагается, что

$$P_2 = P_1 + f(1 - P_1)\,, \tag{7.8}$$

т. е. доля улучшения при втором способе равна сумме доли улучшения при первом виде лечения и дополнительной доли среди тех, кому лечение первым способом не помогает. Отсюда

$$f = \frac{P_2 - P_1}{1 - P_1}\,, \tag{7.9}$$

так что f можно назвать *относительным приростом* [1].

[1] В разд. 6.3 относительным приростом названа оценка (7.10) индекса (7.9). — *Примеч. пер.*

выборочные пропорции p_1 и p_2 — оценки соответ-
ропорций в популяции, получим оценку относи-
роста:

$$p_e = \frac{p_2 - p_1}{1 - p_1}. \qquad (7.10)$$

ная ошибка приблизительно равна (см. [Sheps,

$$\text{s. e.}(p_e) = \frac{1}{q_1}\sqrt{\frac{p_2 q_2}{n_2} + (1-p_e)^2 \frac{p_1 q_1}{n_1}}. \qquad (7.11)$$

Valter, 1975] показал, что более точные выводы от-
можно получить, считая распределение величины
нормальным со средним $\ln(1-f)$ и стандартной

$$\text{s. e.}[\ln(1-p_e)] = \sqrt{\frac{p_2}{n_2 q_2} + \frac{p_1}{n_1 q_1}}. \qquad (7.12)$$

зовавшись данными табл. 7.1, находим относитель-
ст

$$p_e = \frac{0{,}80 - 0{,}60}{1 - 0{,}60} = 0{,}50, \qquad (7.13)$$

ожидать, что из 100 пациентов, которым пер-
чения не приносит улучшения, 50 помогло бы
ым способом. Значение $\ln(1-p_e) = 0{,}69$, а оценка
тной ошибки

$$\text{s. e.}[\ln(1-p_e)] = \sqrt{\frac{0{,}80}{70 \cdot 0{,}20} + \frac{0{,}60}{80 \cdot 0{,}40}} = 0{,}28. \qquad (7.14)$$

кенный 95%-ный доверительный интервал для $\ln(1-f)$

$$-1{,}96 \cdot 0{,}28 \leqslant \ln(1-f) \leqslant -0{,}69 + 1{,}96 \cdot 0{,}28,$$

$$-1{,}24 \leqslant \ln(1-f) \leqslant -0{,}14, \qquad (7.15)$$

получаем доверительный интервал такого же уровня
ственно относительного прироста:

$$0{,}13 \leqslant f \leqslant 0{,}71. \qquad (7.16)$$

ько иные методы требуются, когда цель исследова-
показать, что два средства терапевтического лече-
квивалентны или их различие с клинической точки зре-
несущественно. Примеры таких исследований и соответст-
их методов анализа даны в [Dunnett and Gent, 1977].

7.2. Планы с перекрытием

В следующей главе представлены методы анализа данных, полученных в планируемых испытаниях с предварительным связыванием (группированием в пары) пациентов по факторам-характеристикам, влияющим на результирующий фактор, и последующим случайным назначением лечения.

Один из вариантов плана со связыванием — *план с перекрытием* (crossover design), когда каждый пациент становится объектом как опытной, так и контрольной выборки, т. е. получает лечение обоих видов.

При этом все пациенты, участвующие в испытаниях, рандомизированно разделяются на две выборки. В первой выборке они будут получать лечение в одном порядке, во второй — в обратном. При анализе таких данных следует остерегаться некоторых неприятных обстоятельств.

Мейер и др. [Meier et al., 1958] показали, что порядок лечений влияет на реакцию пациента. Опишем критерий Гарта [Gart, 1969], применимый в этом случае. Предполагается, что градациями результирующего фактора являются «положительная» и «отрицательная» реакция.

Расположим данные в виде табл. 7.2, где, например, n_{12} обозначает число пациентов с положительной реакцией на лечение A и отрицательной — на лечение B в подвыборке, в которой пациентам назначили лечение в порядке AB.

Таблица 7.2

Представление данных, полученных в исследовании по плану с перекрытием

Порядок AB			Порядок BA		
Реакция на лечение A	Реакция на лечение B		Реакция на лечение A	Реакция на лечение B	
	Положительная	Отрицательная		Положительная	Отрицательная
Положительная	n_{11}	n_{12}	Положительная	m_{11}	m_{12}
Отрицательная	n_{21}	n_{22}	Отрицательная	m_{21}	m_{22}

Если реакция пациента на лечение обоих видов одинакова — либо положительна (число таких пациентов в выборке с порядком AB равно n_{11}), либо отрицательна (их число n_{22}), — то информация о различии двух видов лечения отсутствует. Таких пациентов следует исключить из анализа. Аналогично в выборке с порядком BA нужно проигнорировать $m_{11}+m_{22}$ пациентов с одинаковой реакцией на лечение.

В результате получаем табл. 7.3, где $n=n_{12}+n_{21}$, $m=m_{12}+m_{21}$. Если эффекты лечения A и B одинаковы, то пропорции $p_1=n_{12}/n$ и $p_2=m_{21}/m$ должны быть близки, если же нет, то p_1 и p_2 должны различаться. Гипотезу о равной эффективности двух видов лечения, A и B, можно проверить обычным способом: сравнивая p_1 и p_2 (см. задачу 7.2).

Таблица 7.3

Представление данных из табл. 7.2 для проверки гипотезы о равной эффективности видов лечения

Порядок лечений	Результат		Сумма
	Первое лечение лучше	Второе лечение лучше	
AB	n_{12}	n_{21}	n
BA	m_{21}	m_{12}	m
Сумма	$n_{12}+m_{21}$	$n_{21}+m_{12}$	$n+m$

Одна из возможных опасностей в исследованиях с перекрытием — долговременное действие первого вида лечения, которое может повлиять на реакцию от второго вида лечения. Как показано в [Grizzle, 1965], в присутствии этого *«эффекта наложения»* (carry-over effect) для сравнения двух видов лечения можно использовать только данные, полученные на первом этапе, если эффекты наложения для этих лечений различны. Точнее, можно сравнивать только реакции объектов, получивших лечение A, с реакциями объектов, получивших лечение B. Реакции на второй вид лечения выявляют действие эффекта наложения, но могут оказаться бесполезными, если все, что нас интересует, — это просто эффективность лечений.

Различие эффектов наложения можно свести на нет, выдерживая длительный интервал между двумя курсами лечения. Однако чем больше интервал, тем больше вероятность ухода пациента из испытаний.

Планы с перекрытием надежны, если срок действия лечения мал. В обратном случае их следует избегать.

7.3. Альтернативы простой рандомизации

Обязательность рандомизированного назначения лечений в клинических испытаниях продолжает оставаться предметом спора. В [Gehan and Freireich, 1974; Weinstein, 1974], напри-

мер, изложена критика строгой рандомизации в клинических испытаниях и предложены некоторые альтернативы, тогда как Байэр и др. [Byar et al., 1976], а также Пето и др. [Peto et al., 1976] отстаивают ее. В некоторых случаях отказ от рандомизации следует считать оправданным (например, когда исследователь располагает малым числом пациентов, так что данные для контрольной группы приходится набирать по результатам обследований, проводившихся в недавнем прошлом), но число таких случаев мало.

Продолжаются споры и о планах испытаний, в которых сравнивают контрольную группу пациентов с группой, получавшей экспериментальное лечение. Споры и поиски ведутся по двум вопросам. Первый — необходимость стратифицировать, или связывать, пациентов по априорным факторам (например, по полу, возрасту, тяжести заболевания), второй — как гарантировать равенство распределений по слоям между группами с различным лечением. По одну сторону споров — Пето и др. [Peto et al., 1976], рекомендующие не обременять себя стратификацией полученной выборки пациентов, когда предстоит рандомизированное назначение лечения, особенно при большом количестве пациентов.

Все же нельзя не признать, что среди тех, кто определяет планы клинических испытаний, имеется единодушное мнение: чтобы избежать маловероятного, но очень опасного дисбаланса между группами обработок, желательно какое-либо управление — лучше всего рандомизация. Обычно достаточно раздельной и независимой рандомизации пациентов, попавших в различные слои. Однако следует иметь в виду некоторые недавно предложенные модификации простой рандомизации.

Эти модификации предназначены в первую очередь для исследований, в которых пациенты поступают в клинику последовательно в течение времени, так что число пациентов, оказавшихся в каждом слое к концу исследования, изначально неизвестно. Допустим, например, что новый пациент поступает в клинику в тот момент испытаний, когда в соответствующем ему слое большему числу пациентов назначено лечение A, меньшему — B. Тогда вместо того, чтобы назначать лечение, не учитывая данный дисбаланс, можно применять методы, выравнивающие количество пациентов, получающих лечение A и лечение B. Здесь одна из крайностей — схема «несимметричной монеты» Эфрона [Efron, 1971], согласно которой лечение B в нашем примере будет назначено поступившему пациенту с фиксированной вероятностью больше 0,5 и меньше 1. Другая крайность — схема «минимизации» Тэйвса [Taves, 1974], которая обязывает назначить этому пациенту лечение B. Рандомизация в данной схеме используется только в том случае, если

размеры групп больных, получающих лечения A и B, в момент поступления нового пациента равны, а также при назначении лечения той части пациентов, которая имеется в начале испытаний. Схема, предложенная в [Pocock and Simon, 1975], и ее упрощенный вариант в [Freedman and White, 1976] занимают промежуточное положение между этими двумя крайними методами.

Как заметил Покок [Pocock, 1979], возможность реализовать план не менее важна, чем его теоретическая оптимальность. Реализация описанных схем, выравнивающих размеры групп, в которых проводятся лечения вида A и B сложна. Кроме того, в большинстве случаев вполне удовлетворительна простая рандомизация со стратификацией. Это обстоятельство, по видимому, исключает широкое применение таких усложненных планов.

Сравнительные клинические испытания уникальны, так как их целью является не только проверка научных гипотез, но и медицинская помощь пациентам, участвующим в исследовании. В связи с этими требованиями предложена другая альтернатива простым сравнительным испытаниям — адаптивные испытания. В то время как классический план и его модификации, описанные выше, имеют целью назначить лечения различного вида приблизительно равному числу пациентов, в адаптивном плане требуется назначать в возрастающей доле пациентов то лечение, которое оказывается более эффективным согласно результатам, полученным на текущий момент.

Адаптивные планы находятся в стадии разработки и, естественно, еще не часто применяются на практике, несмотря на их внешнюю привлекательность (см.: [Anscombe, 1963; Colton, 1963; Cornfield et al., 1969; Zelen, 1969; Canner, 1970; Robbins, 1974], где предложены и описаны некоторые способы решения компромисса между двумя требованиями: этическим — назначать более эффективное средство как можно большему числу больных и как можно скорее, и статистическим — стремиться к равенству объемов выборок (обзор дан в [Simon, 1977])).

В клинических испытаниях приходится сталкиваться с нежеланием многих пациентов и лечащих врачей участвовать в «слепом» рандомизированном эксперименте. Особенно часто встречаются отказы, когда заболевание угрожает жизни пациента (например, при раковом или сердечном заболевании). Для преодоления этой ситуации Зелен [Zelen, 1979] предложил смелую идею: сначала рандомизированно разделяют всех пациентов, участвующих в исследовании, на две группы, в которых затем назначают лечение следующим образом.

Всех пациентов первой группы лечат стандартным, традиционным для текущего момента способом. Поскольку им в любом случае назначили бы это лечение, согласие пациентов на участие в эксперименте не требуется.

Всем пациентам второй группы предлагают экспериментальное лечение. Если пациент соглашается, его лечат экспериментальным способом, в противном случае его лечат стандартным способом. По завершении испытаний результаты лечения *всех* пациентов второй группы независимо от вида лечения сравнивают с результатами, полученными в первой группе.

Мощность критериев при использовании плана Зелена снижается, поскольку оценка доли положительных реакций во второй группе — это взвешенное среднее доли среди тех, кого лечили экспериментальным способом, и доли среди пациентов второй группы, лечившихся традиционно. Если экспериментальное лечение лучше стандартного, то это среднее будет ближе к доле положительных реакций в первой группе, чем к доле, которая была бы в группе пациентов с экспериментальным лечением.

В пользу предлагаемого плана срабатывает возможность включать в исследование значительно большее число пациентов: в традиционном плане рандомизированное назначение лечения проводится только среди тех, кто согласился принять участие в эксперименте. Предлагаемый метод позволяет включать в рандомизацию всех подходящих для исследования пациентов. Отказ пациента от участия в эксперименте не означает, что он исключен из испытаний. Увеличение мощности за счет большого числа пациентов может преобладать над ее уменьшением за счет спада различия между двумя группами.

Следует ли принять план Зелена, зависит от доли пациентов, соглашающихся лечиться экспериментальным способом, и от величины смещений, возникающих за счет того, что пациенту и лечащему врачу известен способ лечения. Однако эта информация становится доступной только в процессе работы по этому плану.

Задачи

7.1. Предположим, что два вида лечения, которые мы сравнивали по данным табл. 7.1, исследуются теперь в другой клинике. Данные испытаний таковы:

	Число пациентов	Пропорция улучшений
Лечение 1	100	0,35
Лечение 2	100	0,75
Всего	200	0,55

а) Значимо ли теперь различие между долями улучшения?

б) Вычислите разность (7.2) долей улучшения во второй клинике, а также стандартную ошибку (7.3). Значимо ли различаются разности для двух клиник? (*Указание.* Вычислите значение

$$z = \frac{|d_2 - d_1|}{\sqrt{[\text{s.e.}(d_1)]^2 + [\text{s.e.}(d_2)]^2}},$$

где d_1 — разность (7.4); $\text{s.e.}(d_1)$ — оценка стандартной ошибки (7.5); d_2 и $\text{s.e.}(d_2)$ — соответственно значения статистик (7.2) — (7.3) для данных второй клиники. Сравните z со значениями в табл. А.2 для нормального распределения.)

в) Оцените по данным второй клиники относительный прирост (7.9) с помощью (7.10). Значимо ли различаются относительные приросты для двух клиник? (*Указание.* $L_1 = \ln(1 - p_{e\,(1)})$, $L_2 = \ln(1 - p_{e\,(2)})$. Вычислите

$$z = \frac{|L_1 - L_2|}{\sqrt{[\text{s.e.}(L_1)]^2 + [\text{s.e.}(L_2)]^2}}$$

соотнесите полученное значение со значениями в табл. А.2.)

7.2. Предположим, что мы сравниваем два вида лечения в исследовании по плану с перекрытием. Результаты испытаний таковы:

Порядок *AB*			Порядок *BA*		
Реакция на лечение *A*	Реакция на лечение *B*		Реакция на лечение *A*	Реакция на лечение *B*	
	Положительная	Отрицательная		Положительная	Отрицательная
Положительная	20	15	Положительная	30	10
Отрицательная	5	10	Отрицательная	5	5

а) Рассмотрите выборку с порядком *AB*. Каково значение n (т. е. число пациентов, реакции которых дают информацию о различии *A* и *B*)? Вычислите p_1 (пропорцию среди этих пациентов тех, кто положительно отреагировал на первый вид лечения).

б) Аналогично рассмотрите выборку *BA*. Вычислите m, p_2.

в) Проверьте по p_1 и p_2 гипотезу об одинаковой эффективности лечения *A* и *B*.

ЛИТЕРАТУРА

Anscombe, F. J. (1963). Sequential medical trials. *J. Am. Stat. Assoc.*, **58**, 365–384.

Anturane Reinfarction Trial Research Group (1978). Sulfinpyrazone in the prevention of cardiac death after myocardial infarction: The anturane reinfarction trial. *New Engl. J. Med.*, **298**, 290–295.

Byar, D. P., Simon, R. H., Friedewald, W. T., Schlesselman, J. J., DeMets, D. L., Ellenberg, J. H., Gail, M. H., and Ware, J. H. (1976). Randomized clinical trials: Perspectives on some recent ideas. *New Engl. J. Med.*, **295**, 74–80.

Canner, P. L. (1970). Selecting one of two treatments when the responses are dichotomous. *J. Am. Stat. Assoc.*, **65**, 293–306.

Colton, T. (1963). A model for selecting one two medical treatments. *J. Am. Stat. Assoc.*, **58**, 388–401.

Cornfield, J., Halperin, M., and Greenhouse, S. W. (1969). An adaptive procedure for sequential clinical trials. *J. Am. Stat. Assoc.*, **64**, 759–770.

Dunnett, C. W. and Gent, M. (1977). Significance testing to establish equivalence between treatments, with special reference to data in the form of 2 × 2 tables. *Biometrics*, **33**, 593–602.

Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika*, **58**, 403–417.

Fox, T. F. (1959). The ethics of clinical trials. Pp. 222–229 in D. R. Laurence (Ed.). *Quantitative methods in human pharmacology and therapeutics*. New York: Pergamon Press.

Freedman, L. S. and White, S. J. (1976). On the use of Pocock and Simon's method for balancing treatment numbers over prognostic factors in the controlled clinical trial. *Biometrics*, **32**, 691–694.

Frieman, J. A., Chalmers, T. C., Smith, H., and Kuébler, R. R. (1978). The importance of Beta, the type II error and sample size in the design and interpretation of the randomized control trial: Survey of 71 "negative" trials. *New Engl. J. Med.*, **299**, 690–694.

Gart, J. J. (1969). An exact test for comparing matched proportions in crossover designs. *Biometrika*, **56**, 75–80.

Gehan, E. A. and Freireich, E. J. (1974). Non-randomized controls in cancer clinical trials. *New Engl. J. Med.*, **290**, 198–203.

Grizzle, J. E. (1965). The two-period change-over design and its use in clinical trials. *Biometrics*, **21**, 467–480.

Hill, A. B. (1962). *Statistical methods in clinical and preventive medicine*. New York: Oxford University Press.

Mainland, D. (1960). The clinical trial–some difficulties and suggestions. *J. Chronic Dis.*, **11**, 484–496.

Meier, P. (1975). Statistics and medical experimentation. *Biometrics*, **31**, 511–529.

Meier, P. (1979). Terminating a trial–The ethical problem. *Clin. Pharmacol. Ther.*, **25**, 633–640.

Meier, P., Free, S. M., and Jackson, G. L. (1958). Reconsideration of methodology in studies of pain relief. *Biometrics*, **14**, 330–342.

Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J., and Smith, P. G. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *Brit. J. Cancer*, **34**, 585–612.

Pocock, S. J. and Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*, 31, 103–115.

Pocock, S. J. (1979). Allocation of patients to treatment in clinical trials. *Biometrics*, 35, 183–197.

Report of the Committee for the Assessment of Biometric Aspects of Controlled Trials of Hypoglycemic Agents. (1975). *J. Am. Med. Assoc.*, 231, 583–608.

Robbins, H. (1974). A sequential test for two binomial populations. *Proc. Natl. Acad. Sci.*, 71, 4435–4436.

Roth, H. P. and Gordon, R. S. (Eds.) (1979). Proceedings of the National Conference on Clinical Trials Methodology. *Clin. Pharmacol. Ther.*, 25, 629–766.

Sheps, M. C. (1958). Shall we count the living or the dead? *New Engl. J. Med.*, 259, 1210–1214.

Sheps, M. C. (1959). An examination of some methods of comparing several rates or proportions. *Biometrics*, 15, 87–97.

Simon, R. H. (1977). Adaptive treatment assignment methods and clinical trials. *Biometrics*, 33, 743–749.

Simon R. (1979). Restricted randomization designs in clinical trials. *Biometrics*, 35, 503–512.

Taves, D. R. (1974). Minimization: A new method of assigning patients to treatment and control groups. *Clin. Pharmacol. Ther.*, 15, 443–453.

Walter, S. D. (1975). The distribution of Levin's measure of attributable risk. *Biometrika*, 62, 371–374.

Weinstein, M. C. (1974). Allocation of subjects in medical experiments. *New Engl. J. Med.*, 291, 1278–1285.

Zelen, M. (1969). Play the winner rule and the controlled clinical trial. *J. Am. Stat. Assoc.*, 64, 131–146.

Zelen, M. (1979). A new design for randomized clinical trials. *New Engl. J. Med.*, 300, 1242–1245.

Глава 8

Анализ данных в связанных выборках

В планируемых испытаниях (метод выбора III) часто используют прием, состоящий в подборе и связывании (matching) объектов по идентичности факторов, взаимодействующих с изучаемым результирующим явлением, и в последующем независимом рандомизированном назначении обработок (лечений) в каждом наборе (связке). Для сравнения двух видов лечений объекты группируют в связанные пары, трех видов лечений — в связанные тройки, в общем случае для сравнения m лечений используют m-связки. Целью связывания в планируемых испытаниях является повышение точности оценок сравнения разных видов лечения [Hill, 1962, p. 21].

Связывание часто применяют также в сравнительных проспективных и ретроспективных исследованиях (метод выбора II), однако в них оно применяется скорее не для повышения точности, а для усиления справедливости выводов за счет контроля *мешающих* факторов (см. споры по этому вопросу в [Bross, 1969; Miettinen, 1970 a)]). Например, в исследовании зависимости между раком легких и курением возможными мешающими факторами являются пол и возраст, так как они связаны и с пристрастием к курению, и с риском заболевания раком легких. Поэтому в ретроспективном исследовании влияние данных факторов следует устранить, подбирая в пару каждому объекту с раком легких контрольный объект такого же пола и возраста, но не страдающего раком легких. Поскольку опытные и контрольные объекты будут иметь одинаковый пол и возраст, любое различие между опытной и контрольной выборками будет обусловлено другими факторами. В разд. 10.6 описан еще один способ устранения влияния мешающих факторов.

В методе выбора I связывание применять невозможно.

Разд. 8.1 посвящен анализу парных наблюдений в случае, когда результирующий фактор — дихотомический (типа «да — нет»), разд. 8.2 — анализу парных наблюдений в случае, когда число градаций результирующего фактора больше двух.

3 разд. 8.3 рассматривается анализ данных, получаемых при вязывании каждого опытного объекта с несколькими контрольными, которые образуют одну выборку. Разд. 8.4 посвящен сравнению выборок из нескольких (более двух) популяций, когда объекты из нескольких выборок образуют связанные наборы. Обсуждение преимуществ и недостатков связывания приведено в разд. 8.5.

3.1. Парные наблюдения, дихотомический результирующий фактор

Допустим, что проводится ретроспективное исследование, в котором каждый опытный объект связан с одним контрольным. Требуется сравнить относительную частоту исходного фактора у опытных и контрольных объектов. Соответствующим реальности элементом анализа в данном случае является не отдельный объект, а парное наблюдение, поскольку каждый опытный объект связан с контрольным. Исходя из этого представим данные в виде табл. 8.1.

Таблица 8.1

Данные, полученные по результирующему фактору, для случая парных наблюдений

Опытные объекты	Контрольные объекты		Сумма
	Фактор присутствует	Фактор отсутствует	
Фактор присутствует	a	b	$a+b$
Фактор отсутствует	c	d	$c+d$
Сумма	$a+c$	$b+d$	n

Каждая частота в таблице обозначает число пар. Таким образом, всего изучается n пар. Из них в a парах исходный фактор присутствует у обоих объектов, в b парах присутствует у опытного, но отсутствует у контрольного объекта, в c парах присутствует у контрольного, но отсутствует у опытного объекта и в d парах исходный фактор отсутствует у обоих объектов.

Пропорция контрольных объектов с присутствием фактора равна:

$$p_1 = \frac{a+c}{n},$$

опытных объектов —

$$p_2 = \frac{a+b}{n}.$$

Число пар с исходным фактором у обоих объектов, равное a, очевидно, не влияет на разность этих пропорций,

$$p_2 - p_1 = \frac{b-c}{n}. \qquad (8.1)$$

Как показал МакНемар [McNemar, 1947], и a, и d — число пар, в которых исходный фактор присутствует или отсутствует одновременно у обоих объектов, — не влияют явно на стандартную ошибку разности, если истинные пропорции равны, причем

$$\text{s. e. } (p_2 - p_1) = \frac{\sqrt{b+c}}{n}. \qquad (8.2)$$

Квадрат отношения (8.1) к (8.2) с поправкой на непрерывность можно использовать для проверки статистической значимости различия между p_1 и p_2. Вводя поправку, предложенную Эдвардсом [Edwards, 1948], найдем статистику [1]

$$\chi^2 = \left\{ \frac{|p_2 - p_1| - 1/n}{\text{s. e. } (p_2 - p_1)} \right\}^2 = \frac{(|b-c|-1)^2}{b+c}. \qquad (8.3)$$

Получаемые значения χ^2 можно соотносить с таблицей распределения хи-квадрат с одной степенью свободы (см. [McNemar, 1947; Mosteller, 1952; Stuart, 1957]). Если значение χ^2 велико, можно сделать вывод, что опытная и контрольная группы различаются пропорциями объектов, обладающих исходным фактором. Заметим, что в статистику (8.3) вносят вклад только те пары, в которых опытный и контрольный объекты различаются по отношению к присутствию исходного фактора. Мощность критерия (8.3) изучалась в [Miettinen, 1968; Bennet and Underwood, 1970][2].

---

[1] Это выражение, вычисленное в (8.3) с помощью (8.1) и (8.2), можно рассматривать как частный случай статистики Мантела — Хаизела (10.50), если считать каждую пару отдельной группой ($n_{i1} = n_{i2} = 1$, $i = 1, \ldots, g$, где $g = n$ — число пар). — *Примеч. пер.*

[2] Как и в случае расслоения (см. гл. 10), здесь выводы относятся к некоторому «среднему» различию между опытными и контрольными объектами. В такой ситуации выводы зависят от состава мешающих параметров в связках. При этом если при каких-то значениях мешающих параметров связь имеет одно направление (эффект положителен), а при других — противоположное (эффект отрицателен), то различие нивелируется. Как раз это явление имеется в виду в разд. 8.5, где говорится об опасности не обнаружить «взаимодействие». Мощность критериев при альтернативах с разнородностью зависимости в связках теряется. Данное примечание относится ко всем критериям, описанным в гл. 8. — *Примеч. пер.*

Проиллюстрируем критерий, основанный на (8.3) — критерий МакНемара — с помощью данных табл. 8.2.

Таблица 8.2

Данные для демонстрации критерия МакНемара

Опытные объекты	Контрольные объекты		
	Фактор присутствует	Фактор отсутствует	Сумма
Фактор присутствует	15	20	35
Фактор отсутствует	5	60	65
Сумма	20	80	100

Пропорция контрольных объектов с фактором равна: $p_1 = 20/100 = 0{,}2$, пропорция опытных объектов — $p_2 = 35/100 = 0{,}35$; стандартная ошибка разности (8.2) составит: $\sqrt{20+5}/100 = 0{,}5$. Теперь найдем значение статистики (8.3):

$$\chi^2 = \left(\frac{|0{,}35 - 0{,}2| - 0{,}01}{0{,}05}\right)^2 = 7{,}84 .$$

Это значение можно также вычислять, используя b и c — числа пар с различием контрольных и опытных объектов по отношению к исходному фактору:

$$\chi^2 = \frac{(|20 - 5| - 1)^2}{20 + 5} = 7{,}84 .$$

Поскольку χ^2 больше 6,63, критического значения при уровне значимости 0,01, следует признать, что опытные и контрольные объекты различаются по отношению к присутствию исходного фактора.

Как было указано в гл. 5 и 6, важной мерой степени связи между исходным фактором и заболеванием является *отношение шансов* (например, отношение шансов заболеть при наличии фактора к шансам при его отсутствии). Мантел и Ханзел [Mantel and Haenszel, 1959], Корнфилд и Ханзел [Cornfield and Haenszel, 1960] разработали и исследовали метод оценивания отношения шансов для случая связанных пар. Пользуясь данными в виде табл. 8.1, можно выразить оценку просто как

$$o = \frac{b}{c}, \qquad (8.4)$$

а ее стандартную ошибку можно оценить величиной

$$\text{s. e. } (o) = o\sqrt{\frac{1}{b} + \frac{1}{c}} \tag{8.5}$$

(см. [Ejigou and McHugh, 1977]). Для частот табл. 8.2 оценка отношения шансов равна: $o = 20/5 = 4{,}0$, а оценка ее стандартной ошибки —

$$\text{s. e. } (o) = 4\sqrt{\frac{1}{20} + \frac{1}{5}} = 2{,}0\,.$$

Приближенный доверительный интервал для истинного отношения шансов ω можно получить таким образом. Определим

$$P = \frac{\omega}{\omega + 1}\,. \tag{8.6}$$

Оценкой P является

$$p = \frac{b}{b + c}\,. \tag{8.7}$$

Для поиска $100(1-\alpha)\%$-ного доверительного интервала для параметра P,

$$P_L \leqslant P \leqslant P_U\,, \tag{8.8}$$

можно использовать метод, описанный в разд. 1.4, а для поиска доверительного интервала для ω достаточно подставить (8.6) в (8.8) и обратить неравенства:

$$\frac{P_L}{1 - P_L} \leqslant \omega \leqslant \frac{P_U}{1 - P_U}\,. \tag{8.9}$$

Воспользовавшись данными табл. 8.2, получаем, что $b + c = 20 + 5 = 25$ и $p = 20/25 = 0{,}8$. По формулам (1.26) и (1.27) приближенные 95%-ные доверительные интервалы для P и истинного отношения шансов ω составляют:

$$0{,}587 \leqslant P \leqslant 0.924, \tag{8.10}$$

$$1{,}42 \leqslant \omega \leqslant 12{,}16. \tag{8.11}$$

Более простой подход, основанный на (1.29), дает

$$0{,}623 \leqslant P \leqslant 0{,}977 \tag{8.12}$$

в качестве приближенного 95%-ного доверительного интервала для P и

$$1{,}65 \leqslant \omega \leqslant 42{,}48 \tag{8.13}$$

в качестве приближенного 95%-ного доверительного интервала для ω. Нижние границы в (8.11) и (8.13) хорошо согласуются, зато верхние сильно различаются. Поскольку в данном случае

$p = 0{,}80$ находится вне интервала (0,3; 0,7), предложенного в разд. 1.4 в качестве множества близости двух подходов, следует предпочесть результат (8.11).

До сих пор мы описывали анализ четырехклеточных таблиц со связанными парами для сравнительного ретроспективного исследования. Такой же анализ, включая проверку по критерию МакНемара и оценивание отношения шансов, можно проводить и при сравнительном проспективном исследовании со связанными парами. Обнаружив с помощью критерия МакНемара значимое различие в планируемых испытаниях со связанными парами, следует провести еще точечное и интервальное оценивания разности двух результирующих пропорций или относительного прироста.

Подходящим способом представления данных в таких испытаниях (мы предположили, что новый вид лечения сравнивается с традиционным) является табл. 8.3. Как и в табл. 8.1, каждая частота обозначает число пар.

Таблица 8.3

Данные клинических испытаний со связыванием объектов в пары

Новый метод лечения	Традиционный метод лечения		Сумма
	Выздоровели	Не выздоровели	
Выздоровели	a	b	$a+b$
Не выздоровели	c	d	$c+d$
Сумма	$a+c$	$b+d$	n

Пропорция выздоровевших при традиционном лечении равна:

$$p_1 = \frac{a+c}{n},$$

а пропорция выздоровевших при лечении новым методом —

$$p_2 = \frac{a+b}{n}.$$

Далее,

$$p_2 - p_1 = \frac{b-c}{n},$$

и оценка стандартной ошибки этой разности при различных истинных пропорциях равна:

$$\text{s. e. } (p_2 - p_1) = \frac{\sqrt{n\,(b+c) - (b-c)^2}}{n\sqrt{n}} = \frac{\sqrt{(a+d)\,(b+c) + 4\,bc}}{n\sqrt{n}}. \tag{8.14}$$

Приближенный $100(1-\alpha)\%$-ный доверительный интервал для разности двух истинных пропорций числа выздоравливающих дается неравенствами:

$$(p_2 - p_1) - c_{\alpha/2}\ \text{s. e. } (p_2 - p_1) - \frac{1}{n} \leqslant P_2 - P_1 \leqslant$$
$$\leqslant (p_2 - p_1) + c_{\alpha/2}\ \text{s. e. } (p_2 - p_1) + \frac{1}{n}. \tag{8.15}$$

Заметим, что все четыре наблюдаемые частоты входят в выражение для стандартной ошибки (8.14), в отличие от выражения для стандартной ошибки (8.2), применимой только при проверке гипотезы о равенстве истинных пропорций.

Принимая предположение, что преимущество лечения новым методом проявляется через пациентов, у которых традиционное лечение не улучшает состояние, относительную эффективность нового лечения можно оценить относительным приростом

$$p_e = \frac{p_2 - p_1}{1 - p_1} = \frac{b - c}{b + d}. \tag{8.16}$$

Стандартная ошибка относительного прироста будет оцениваться величиной

$$\text{s. e. } (p_e) = \frac{1}{(b+d)^2}\sqrt{(b + c + d)(bc + bd + cd) - bcd}. \tag{8.17}$$

Заметим, что число пар a, в которых оба объекта выздоровели, не входит ни в оценку относительного прироста, ни в оценку его стандартной ошибки. Приближенный $100(1-\alpha)\%$-ный доверительный интервал для истинного значения параметра определяется границами

$$p_e \pm c_{\alpha/2}\ \text{s. e. } (p_e).$$

В табл. 8.4 представлены иллюстративные данные. Пропорция выздоровевших среди пациентов, которых лечили традиционным методом, равна $p_1 = 50/75 = 0{,}67$, а пропорция числа выздоровевших среди прошедших новое лечение — $p_2 = 65/75 =$

Таблица 8.4

Вымышленные данные планируемых испытаний со связыванием объектов в пары

Новый метод лечения	Традиционный метод лечения		Сумма
	Выздоровели	Не выздоровели	
Выздоровели	40	25	65
Не выздоровели	10	0	10
Сумма	50	25	75

0,87. Теперь находим значение статистики МакНемара (8.3) для проверки значимости различия двух пропорций

$$\chi^2 = \frac{(|25-10|-1)^2}{25+10} = 5{,}60 .$$

Различие значимо при уровне 0,05.

Разность двух пропорций равна:

$$p_2 - p_1 = \frac{25-10}{75} = 0{,}20 ,$$

а оценка (8.14) ее стандартной ошибки —

$$\text{s. e. } (p_2 - p_1) = \frac{1}{75\sqrt{75}} \sqrt{(40+0)\cdot(25+10) + 4\cdot 25\cdot 10} = 0{,}08.$$

Подставив в (8.15) полученные значения, определим приближенный 95%-ный доверительный интервал для $P_2 - P_1$:

$$0{,}20 - 1{,}96\cdot 0{,}08 - 1/75 \leqslant P_2 - P_1 \leqslant 0{,}20 + 1{,}96\cdot 0{,}08 + 1{,}75,$$

или

$$0{,}03 \leqslant P_2 - P_1 \leqslant 0{,}37 .$$

Значение относительного прироста (8.16) равно:

$$p_e = (25-10)/25 = 0{,}60.$$

Значит, можно ожидать, что из каждых 100 пациентов, которые не выздоровели при традиционном лечении, 60 выздоровеют при лечении новым методом. Оценка (8.17) стандартной ошибки относительного прироста равна:

$$\text{s. e. } (p_e) = \frac{1}{(25+0)^2} \sqrt{(25+10+0)\cdot(25\cdot 10+25\cdot 0+10\cdot 0) - 25\cdot 10\cdot 0} = 0{,}15 .$$

Приближенный 95%-ный доверительный интервал для истинного значения параметра $-0{,}60 \pm 1{,}96 \cdot 0{,}15$, т. е. (0,30; 0,90).

8.2. Парные наблюдения. Число категорий результирующего фактора больше двух

Часто реакция объекта на лечение или степень наличия у него фактора поддается более точной классификации, чем простой дихотомии (наличие — отсутствие), рассматривавшейся в предыдущем разделе. Например, реакция на лечение может иметь следующие градации: улучшение, без изменений, ухудшение, а градациями пристрастия к курению могут быть: не курит, выкуривает от одной до десяти сигарет за день, выкуривает 11—20 сигарет за день, выкуривает за день более 20 сигарет. Когда подлежащие сравнению выборки не взаимосвязаны, можно применять методы, описанные в гл. 9. Здесь же мы рассмотрим случай со связанными в пары объектами, которые классифицируются по k непересекающимся категориям ($k>2$).

Подходящим способом представления данных является табл. 8.5. Каждый элемент таблицы обозначает число пар. Например, $n_{..}$ — общее число пар; $n_{1.}$ — число пар, в которых опытный объект относится к категории 1; $n_{.2}$ — число пар, в которых контрольный объект относится к категории 2; n_{12} — число пар, в которых опытный объект относится к категории 1, а контрольный — к категории 2. Различие между опытным и контрольным объектами описывается k разностями: $d_1 = (n_{1.} - n_{.1})$, $d_2 = (n_{2.} - n_{.2})$, ..., $d_k = (n_{k.} - n_{.k})$. Очевидно, что разности не зависят от частот $n_{11}, n_{22}, \ldots, n_{kk}$, т. е. от количества пар, в которых реакции обоих объектов относятся к одной и той же категории.

Бхапкаром [Bhapkar, 1966], Гриззлом и др. [Grizzle et al., 1969], а также Айэлендом и др. [Ireland et al., 1969] были предложены весьма сложные статистики для проверки значимости k разностей $d_1, d_2, \ldots, d_k$. Более простую, но все же требующую обращения матрицы статистику описали в своих работах Стюарт [Stuart, 1955] и Максвелл [Maxwell, 1970]. Флейсом и Эвериттом [Fleiss and Everitt, 1971] для статистики Стюарта—Максвелла при $k=3$ было получено простое выражение, а именно:
если

$$\bar{n}_{ij} = \frac{n_{ij} + n_{ji}}{2}, \qquad (8.18)$$

Таблица 8.5

Данные исследования со связыванием объектов в пары в случае k взаимоисключающих категорий результирующего фактора

Категория результирующего фактора для опытного объекта	Категория результирующего фактора для контрольного объекта				Сумма
	1	2	...	k	
1	n_{11}	n_{12}	...	n_{1k}	$n_{1.}$
2	n_{21}	n_{22}	...	n_{2k}	$n_{2.}$
.					
k	n_{k1}	n_{k2}	...	n_{kk}	$n_{k.}$
Сумма	$n_{.1}$	$n_{.2}$	...	$n_{.k}$	$n_{..}$

то будет получена статистика

$$\chi^2 = \frac{\bar{n}_{23}\, d_1^2 + \bar{n}_{13}\, d_2^2 + \bar{n}_{12}\, d_3^2}{2\,(\bar{n}_{12}\,\bar{n}_{13} + \bar{n}_{12}\,\bar{n}_{23} + \bar{n}_{13}\,\bar{n}_{23}}, \qquad (8.19)$$

значения которой можно сравнивать со значениями распределения хи-квадрат с двумя степенями свободы. Если χ^2 — значимо велико, можно утверждать, что распределения опытных и контрольных объектов по категориям различаются.

Допустим, два врача независимо друг от друга диагностируют пациентов выборки, состоящей из 100 психических больных (табл. 8.6). Подставив данные этой таблицы в (8.19), получим значение статистики Стюарта—Максвелла, которое значимо при двух степенях свободы на уровне 0,01.

$$\chi^2 = \frac{\frac{5+5}{2}(40-60)^2 + \frac{0+10}{2}(40-30)^2 + \frac{5+15}{2}(20-10)^2}{2\left(\frac{5+15}{2}\cdot\frac{0+10}{2} + \frac{5+15}{2}\cdot\frac{5+5}{2} + \frac{0+10}{2}\cdot\frac{5+5}{2}\right)} =$$

$$= 14{,}00\,.$$

Следовательно, можно сделать вывод о различии распределений для диагностов.

Таблица 8,6

Данные для иллюстрации критерия Стюарта — Максвелла

Диагност *B*	Диагност *A*			Сумма
	Шизофрения	Эмоциональное расстройство	Другие психические заболевания	
Шизофрения	35	5	0	40
Эмоциональное расстройство	15	20	5	40
Другие психические заболевания	10	5	5	20
Сумма	60	30	10	100

Если обнаружено значимое различие двух распределений, как в этом примере, то на следующем шаге анализа следует выделить категории (в случае более трех категорий — комбинации категорий) со значимым различием (см. общее обсуждение в [Fleiss and Everitt, 1971]. Для этого нужно лишь сжать исходную таблицу в таблицу 2×2 и применить критерий МакНемара (8.3). При этом в критерии значимости обязательно должно учитываться то обстоятельство, что применение к одним и тем же данным нескольких критериев повышает вероятность ошибочного вывода о значимости различия. В случае, который мы рассматриваем, подходящим методом учета такого обстоятельства (см. [Miller, 1966, Section, 6.2]) является сравнение статистики МакНемара и критического значения распределения хи-квадрат с $(k-1)$ степенями свободы.

Мы продемонстрируем метод поиска значимо различающихся категорий на примере данных табл. 8.6. Чтобы определить, различаются ли пропорции числа пациентов, которым психиатры поставили диагноз шизофрения, сформируем таблицу 2×2 (табл. 8.7). Психиатр *A* определил шизофрению у 60% пациентов, психиатр *B* — только у 40%. Значение статистики МакНемара —

$$\chi^2 = \frac{(|5-25|-1)^2}{5+25} = 12{,}03\,.$$

Критическое значение распределения хи-квадрат с двумя степенями свободы при уровне значимости 0,05 равно 5,99 (см. табл. А.1). Поскольку полученное значение статистики Мак-

Таблица 8.7

Таблица 2×2 для сравнения долей пациентов с диагнозом шизофрения, поставленным диагностами *A* и *B*

Диагност *B*	Диагност *A*		Сумма
	Шизофрения	Не шизофрения	
Шизофрения	35	5	40
Не шизофрения	25	35	60
Сумма	60	40	100

Немара больше 5,99, можно утверждать, что *A* чаще склонен ставить диагноз шизофрения, чем *B*.

В задаче 8.1 требуется сравнить пропорции числа пациентов с диагнозом «эмоциональные расстройства», поставленным психиатрами *A* и *B*, а также пропорции числа диагностированных больных с другими патологиями.

Если k категорий результирующего фактора упорядочены (см. примеры в начале раздела), то при анализе это упорядочение должно учитываться. Представим, что в результате испытаний, в которых лечение двух видов назначалось объектам в каждой паре случайно, получили данные табл. 8.8.

Таблица 8.8

Данные для демонстрации анализа упорядоченного результирующего фактора

Новый метод лечения	Традиционный метод лечения			Сумма
	Улучшение	Без изменений	Ухудшение	
Улучшение	40	20	10	70
Без изменений	6	6	8	20
Ухудшение	4	4	2	10
Сумма	50	30	20	100

Значение статистики Стюарта—Максвелла значимо (см. задачу 8.2а), поэтому приступим к более подробному анализу.

Можно провести анализ по каждой категории описанным выше способом. Однако такой путь неэффективен, поскольку игнорирует упорядочение категорий. Если вопрос состоит в том, чаще ли реакции на первый вид лечения относятся к категориям на одном конце шкалы и соответственно реже — к категориям на другом конце шкалы по сравнению с реакциями на второй вид лечения, то лучше воспользоваться следующим методом.

Рассмотрим разность $d_1 - d_3$. Если новое лечение эффективнее традиционного, т. е. оно чаще приводит к положительному исходу (и, значит, величина d_1 — положительная) и реже — к отрицательному (d_3 — отрицательная), то значение $d_1 - d_3$ будет положительным. В противном случае — $d_1 - d_3$ будет отрицательным. В обоих случаях гипотезу об отсутствии различия между лечениями (по отношению к первой и последней категориям реакций на лечение) можно проверить, сравнивая значение

$$\chi^2 = \frac{(d_1 - d_3)^2}{2\,(\bar{n}_{12} + 4\,\bar{n}_{13} + \bar{n}_{23})} \qquad (8.20)$$

со значениями распределения хи-квадрат с одной степенью свободы (если сравнение было запланировано заранее) или с двумя степенями свободы (если сравнение было продиктовано данными). Этот критерий и критерий для более общего случая, когда число категорий больше трех, выведен в [Fleiss and Everitt, 1971]. В задаче 8.2 б требуется применить его для данных табл. 8.8.

8.3. Связывание с несколькими контрольными объектами

Две связанные выборки часто можно формировать, связывая каждый опытный объект (пациента, которого лечили новым методом) с несколькими контрольными объектами (пациентами, которых лечили традиционно). Связывание с несколькими контрольными объектами особенно выгодно, когда число объектов, доступных для формирования контрольной выборки, велико по сравнению с возможным числом опытных объектов и когда для получения необходимой информации требуются небольшие усилия.

Предположим, что все объекты характеризуются отсутствием или наличием фактора. Общий метод анализа, пригодный даже в случае различного числа контрольных объектов, связанных с каждым опытным объектом, был разработан Мантелом и Ханзелом [Mantel and Haenszel, 1959]. Кокс [Cox. 1966] предложил другой, более сложный метод анализа для общего

случая. Мы будем считать, что каждый опытный объект связан с одинаковым числом (скажем, m—1) контрольных и рассмотрим только метод Мантела и Ханзела.

Допустим, всего имеется N m-связок, состоящих из одного опытного и m—1 контрольных объектов. Обозначим через x_i — число контрольных объектов с наличием фактора в i-й m-связке (т. е. x_i может принимать значения 0,1,..., m—1), а через n_i суммарное число объектов с фактором в i-й m-связке, включая опытный и контрольный объекты. Значит, если опытный объект i-й связки имеет фактор, то $n_i = x_i + 1$, в противном случае — $n_i = x_i$.

Определим

$$A = \sum_{i=1}^{N} x_i \tag{8.21}$$

число контрольных объектов с наличием фактора,

$$B = \sum_{i=1}^{N} n_i \tag{8.22}$$

суммарное число объектов выборки с присутствующим фактором. Ясно, что число опытных объектов с наличием фактора равно $B-A$. Доля контрольных объектов с фактором определяется величиной

$$p_1 = \frac{A}{N(m-1)}, \tag{8.23}$$

а доля опытных объектов с присутствующим фактором — величиной

$$p_2 = \frac{B-A}{N} \tag{8.24}$$

Чтобы проверить значимость различия между p_1 и p_2, следует подсчитать значение статистики [1]

$$\chi^2 = \left(\frac{p_2 - p_1}{\text{s. e. } (p_2 - p_1)} \right)^2 = \frac{[(m-1)B - mA]^2}{mB - \sum_{i=1}^{N} n_i^2} \tag{8.25}$$

[1] Выражение (8.25) может быть получено из выражения (10.50) для статистики Мантела — Ханзела, если считать каждую связку отдельной группой, т. е. положить в (10.50) $n_{i1} = m-1$, $n_{i2} = 1$, $p_{i1} = x_i/(m-1)$, $p_{i2} = n_i - x_i$, $p_i = n_i/m$ и $g = N$, и пренебречь поправкой на непрерывность. Можно без труда вывести из (10.50) выражение для более общей статистики, когда i-й опытный объект связан с различным числом m_i контрольных объектов.

Если сохранить в (8.25) поправку на непрерывность, то при $m=2$ (связанные пары) мы получим статистику МакНемара (8.3). На примере статистик (8.3) и (8.25) хорошо видна тесная связь методов анализа связанных и расслоенных данных (см. гл. 10). — *Примеч. пер.*

и обратиться к таблице распределения хи-квадрат с одной степенью свободы (см. [Miettinen, 1969; Pike and Morrow, 1970]). Миеттинен [Miettinen, 1969] изучил мощность критерия, основанного на (8.25), и вывел правило выбора подходящего (в терминах некой условной стоимости) значения $m-1$ для числа контрольных объектов на каждый опытный объект.

Продемонстрируем описанный метод с помощью данных табл. 8.9.

Таблица 8.9

Данные по связанным тройкам

Тройка	Наличие фактора у опытного объекта*	Число контрольных объектов с фактором ($=x_i$)	Общее число объектов с фактором ($=n_i$)	n_i^2
1	1	2	3	9
2	1	1	2	4
3	1	1	2	4
4	1	1	2	4
5	1	1	2	4
6	1	0	1	1
7	1	0	1	1
8	1	0	1	1
9	0	1	1	1
10	0	0	0	0
Сумма	$B-A=8$	$A=7$	$B=15$	29

* 1 — да, 0 — иет.

Пусть исследованы $N=10$ связанных троек ($m=3$). На один опытный объект приходится $m-1=2$ контрольных объекта. Пропорция (8.23) контрольных объектов с фактором —

$$p_1 = \frac{7}{10 \cdot 2} = 0{,}35,$$

пропорция (8.24) опытных объектов с фактором —

$$p_2 = \frac{8}{10} = 0{,}80 .$$

статистика (8.25) проверки значимости различия двух пропорций равна:

$$\chi^2 = \frac{(2 \cdot 15 - 3 \cdot 7)^2}{3 \cdot 15 - 20} = 5{,}06 .$$

Поскольку это значение больше критического значения 3,84 распределения хи-квадрат при уровне значимости 0,05, следует признать, что пропорция опытных объектов с фактором больше пропорции контрольных объектов с фактором.

Оценка Мантела—Ханзела предположительно общего отношения шансов по N m-связкам [Mantel and Haenszel, 1959, p. 736] находится по формуле [1]:

$$o = \frac{(m-1)\,(B-A) - \sum_{i=1}^{N} x_i\,(n_i - x_i)}{A - \sum_{i=1}^{N} x_i\,(n_i - x_i)} . \qquad (8.26)$$

Величина $\Sigma\, x_i (n_i - x_i)$ — просто сумма чисел контрольных объектов с фактором по m-связкам, в которых опытный объект имеет фактор.

В табл. 8.9 только в первых восьми тройках опытный объект имеет фактор. Суммарное число контрольных объектов с фактором в этих тройках составит $\Sigma x_i (n_i - x_i) = 6$, так что оценка (8.26) отношения шансов равна:

$$o = \frac{2 \cdot 8 - 6}{7 - 6} = 10 .$$

Миеттинен в своей работе [Miettinen, 1970 b] описывает более сложный метод оценивания отношения шансов в случае

---

[1] Если строго следовать определению отношения шансов, то в обозначениях данного раздела (p_1 относится к контрольным, p_2 — к опытным объектам) в качестве его оценки o надо брать величину, обратную (8.26). Выражение (8.26) мы получим из (10.47) аналогично выводу выражения (8.25) из (10.50) (поменяв при этом местами индексы 1 и 2, т. е. полагая $n_{i1} = 1$, $n_{i2} = m - 1$, $p_{i1} = n_i - x_i$, $p_{i2} = x_i/(m-1)$).

В случае связанных пар ($m = 2$) величиной, обратной (8.26), является (8.4).— *Примеч. пер.*

m-связок и находит приближенное выражение для стандартной ошибки предложенной оценки.

Мощность критерия, основанного на (8.25), и точность оценки (8.26) повышаются с увеличением числа $m-1$ контрольных объектов, связанных с опытным объектом. Однако если число контрольных объектов, связанных с опытным объектом, больше трех или четырех, то повышение мощности и точности обычно незначительно [Miettinen, 1969, Ury. 1975]. Поэтому поиск пяти или более контрольных объектов для каждого опытного объекта является, как правило, бесполезным занятием.

8.4. Сравнение m связанных выборок

В предыдущем разделе мы рассмотрели случай, когда все $m-1$ контрольных объектов в связке образуют однородную группу. В настоящем разделе обсудим сравнение m выборок, определяющих одновременно несколько m-связок. Мы снова ограничимся случаем дихотомического фактора (общий случай обсуждается в [Koch and Reinfurt, 1971]).

Задача рассматриваемого вида возникает, например, при сравнительном проспективном исследовании по четверкам объектов, связанных по полу и возрасту, в которых первый объект не курит, второй объект выкуривает от одной до десяти сигарет в день, третий — от 11 до 20 сигарет и четвертый — более 20 сигарет в день. Для сравнения пропорций числа заболевших в четырех получаемых выборках можно будет использовать методы этого раздела. В [Doll and Hill, 1952] приведен пример ретроспективного исследования с тремя связанными выборками (больные раком легких, больные с другим заболеванием, контрольная выборка из населения).

Методы данного раздела пригодны также для анализа результатов планируемых испытаний, в которых $m>2$ видов лечения сравниваются путем формирования m-связок из m сходных пациентов. В каждой связке пациентам случайно назначают разное лечение. Эти методы также применимы, когда выборка объектов изучается в m различных условиях. Пример такого рода — сравнение пропорций объектов, положительно реагирующих на m диагностических тестов, когда каждый пациент выборки диагностируется с помощью нескольких методов.

В табл. 8.10 представлены данные, полученные при исследовании m связанных выборов, по N наблюдений в каждой.

Таблица 8.10

Данные, полученные из m связанных выборок

m-связка	Выборка 1	Выборка 2	...	Выборка m	Сумма
1	X_{11}	X_{12}	...	X_{1m}	S_1
2	X_{21}	X_{22}	...	X_{2m}	S_2
...					
N	X_{N1}	X_{N2}	...	X_{Nm}	S_N
Сумма	T_1	T_2	...	T_m	T
Пропорция	p_1	p_2	...	p_m	$\bar{p}$

Величины X в таблице принимают значение 0 (если реакция отрицательная) или 1 (если реакция положительная). Значит, S_1 — число положительных реакций в первой m-связке, T_1 — число положительных реакций в первой выборке, T — суммарное число положительных реакций во всех выборках и т. д. Дополнительно определим

$$p_j = \frac{T_j}{N} \tag{8.27}$$

пропорцию объектов j-й выборки с положительной реакцией;

$$P_n = \frac{S_n}{m} \tag{8.28}$$

пропорцию положительных реакций в n-й m-связке;

$$\bar{p} = \frac{1}{m}\sum_{j=1}^{m} p_j = \frac{1}{N}\sum_{n=1}^{N} P_n = \frac{T}{Nm} \tag{8.29}$$

общую пропорцию положительных реакций.

Основной интерес представляет вопрос значимости различий среди пропорции $p_1, \ldots, p_m$. Выпишем статистику, предложенную Кохрэном [Cochran, 1950]:

$$Q = \frac{N^2(m-1)}{m} \cdot \frac{\sum_{j=1}^{m} (p_j - \bar{p})^2}{N\bar{p}(1-\bar{p}) - \sum_{n=1}^{N} (P_n - \bar{p})^2} =$$

$$= (m-1) \cdot \frac{m \sum_{j=1}^{m} T_j^2 - T^2}{mT - \sum_{n=1}^{N} S_n^2} . \qquad (8.30)$$

Сравнивая значение (8.30) с критическим значением распределения хи-квадрат с $m-1$ степенями свободы, можно проверять гипотезу об отсутствии различий в m пропорциях.

Рассмотрим данные табл. 8.11, взятые из [Fleiss, 1965 a]. Пропорции числа пациентов, имеющих, по мнению экспертов, психические расстройства на религиозной почве, колеблются от 0 до 0,375. Значение Q в (8.30), используемое, чтобы выяснить, объясняется ли различие случайными колебаниями или действительным расхождением в оценках экспертов, равно:

$$Q = 7 \cdot \frac{8(1^2 + 0^2 + \ldots + 0^2 + 3^2) - 15^2}{8 \cdot 15 - (0^2 + 1^2 + \ldots + 0^2 + 6^2)} = 14{,}71 .$$

Обоснованность вычисления Q для случая, когда каждый объект многократно классифицируется, установлена Флейсом [Fleiss, 1965 b].

По табл. А.1 критическое значение для $m-1=7$ степеней свободы при уровне значимости 0,05 составит 14,07. Раз мы получили значение $Q=14{,}71$, превышающее критическое значение, следует сделать вывод, что эксперты по-разному судят о наличии расстройств на религиозной почве.

Обнаружив значимое различие, можно попытаться выделить различающиеся выборки или группы выборок (в нашем примере экспертов или групп экспертов). Прием, который при этом часто оказывается полезным, состоит в разбиении Q на отдельные компоненты, в каждой из которых определяется степень разнородности. Общий метод разбиения статистики хи-квадрат описан в [Everitt, 1977, Chapter, 3]. Здесь мы продемонстрируем его для статистики (8.30).

Таблица 8.11

Оценка восьми экспертов о наличии или отсутствии психических расстройств * на религиозной почве у восьми пациентов

Пациент	Эксперт								Сумма ($=S_n$)
	1	2	3	4	5	6	7	8	
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	1	0	0	0	1
3	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0
5	0	0	1	0	0	1	0	1	3
6	0	0	1	1	1	1	0	1	5
7	0	0	0	0	0	0	0	0	0
8	1	0	1	1	1	1	0	1	6
Сумма (T_j)	1	0	3	2	3	3	0	3	$T=15$
Пропорция (p_j)	0,125	0	0,375	0,250	0,375	0,375	0	0,375	$\bar{p}=0,234$

* 1 или 0 соответственно.

Допустим, m выборок представляют две группы. Первой группе соответствует m_1 выборок, второй — m_2 выборок. В примере, приведенном в начале раздела, первую группу представляет выборка некурящих ($m_1=1$), вторую — остальные три выборки ($m_2=3$). Определим суммарное число положительных реакций в первой группе выборок:

$$U_1 = \sum_{j=1}^{m_1} T_j \tag{8.31}$$

и по второй группе выборок:

$$U_2 = \sum_{j=m_1+1}^{m} T_j \,. \tag{8.32}$$

Теперь найдем статистику для проверки значимости различия между пропорциями положительных реакций в первой и во второй группах:

$$Q_{\text{diff}} = \frac{(m-1)}{m_1 m_2} \cdot \frac{(m_2 U_1 - m_1 U_2)^2}{m T - \sum_{n=1}^{N} S_n^2}. \qquad (8.33)$$

Она имеет одну степень свободы.

Снова рассмотрим данные табл. 8.11. Из восьми экспертов первые пять — из Нью-Йорка, остальные — из Кентукки. Они естественным образом составляют две группы ($m_1=5$ и $m_2=3$). Разумно проверить, есть ли различие между двумя группами экспертов по принимаемым ими решениям. В первой группе суммарное число положительных решений равно:

$$U_1 = 1+0+3+2+3 = 9,$$

во второй —

$$U_2 = 3+0+3 = 6.$$

Вычислим значение (8.33):

$$Q_{\text{diff}} = \frac{7}{5 \cdot 3} \cdot \frac{(3 \cdot 9 - 5 \cdot 6)^2}{120 - 71} = 0{,}09,$$

что соответствует незначительному различию между группами экспертов из Нью-Йорка и Кентукки.

На следующем шаге анализа сравним между собой m_1 выборок первой группы с помощью статистики

$$Q_1 = \frac{m(m-1)}{m_1} \cdot \frac{m_1 \sum_{j=1}^{m_1} T_j^2 - U_1^2}{m T - \sum_{n=1}^{N} S_n^2} \qquad (8.34)$$

и m_2 выборок второй группы с помощью

$$Q_2 = \frac{m(m-1)}{m_2} \cdot \frac{m_2 \sum_{j=m_1+1}^{m} T_j^2 - U_2^2}{m T - \sum_{n=1}^{N} S_n^2}. \qquad (8.35)$$

статистики Q_1 и Q_2 имеют соответственно m_1-1 и m_2-1 степеней свободы. Легко проверить, что

$$Q = Q_{\text{diff}} + Q_1 + Q_2 .$$

Отсюда ещё, что сумма степеней свободы статистик (8.33)—(8.35) равна числу степеней свободы суммарной статистики Q из (8.30), т. е. $1+m-1+m_2-1=m-1$.

По данным табл. 8.11 определим различие в оценках экспертов из Нью-Йорка:

$$Q_1 = \frac{8 \cdot 7}{5} \cdot \frac{5\,(1^2+0^2+3^2+2^2+3^2)-9^2}{120-71} = 7{,}77 .$$

Эта величина при $5-1=4$ степенях свободы слишком мала, чтобы признать различие значимым. Различие в оценках экспертов из Кентукки равно:

$$Q_2 = \frac{8 \cdot 7}{3} \cdot \frac{3\,(3^2+0^2+3^2)-6^2}{120-71} = 6{,}86 ,$$

что значимо на уровне 0,05 при $3-1=2$ степенях свободы. Сумма

$$Q_{\text{diff}} + Q_1 + Q_2 = 0{,}09 + 7{,}77 + 6{,}86 = 14{,}72$$

равна, с точностью до ошибок округления суммарному значению $Q=14{,}71$.

Несколько другой подход к разбиению Q предложен Кохреном [Cochran, 1950, p. 265]. Он приводит к чуть меньшим значениям Q_1 и Q_2 (см. также [Tate and Brown, 1979]). По данным табл. 8.11 оба метода разбиения приведут к одному и тому же заключению: между оценками восьми экспертов из Нью-Йорка есть различие, которое обусловлено, по сути, различием оценок экспертов из Кентукки.

Предположим теперь, что m выборок представляют m уровней количественной шкалы переменной (среднее число выкуриваемых за день сигарет). Пусть x_j обозначает значение этой переменной в j-й выборке ($j=1, \ldots, m$). Определим статистику

$$b = \frac{\sum_{j=1}^{m} l_j\,(x_j - \bar{x})}{N \sum_{j=1}^{m} (x_j - \bar{x})^2} , \tag{8.36}$$

где

$$\bar{x} = \frac{1}{m} \sum_{j=1}^{m} x_j \tag{8.37}$$

— среднее значение x в этих выборках. Статистика b — коэффициент регрессии для прямой, описывающей поведение данных. Он соответствует среднему изменению доли наступления изучаемого события на единичное изменение x.

Статистическую значимость b (т. е. отличие от нуля) можно оценить, соотнося значение статистики

$$\chi^2_{\text{slope}} = \frac{m\,(m-1)\,N^2 \sum_{j=1}^{m} (x_j - \bar{x})^2}{m\,T - \sum_{n=1}^{N} S_n^2} \cdot b^2 \qquad (8.38)$$

с таблицей распределения хи-квадрат с одной степенью свободы. Если b значимо положителен (или отрицателен), можно утверждать, что пропорция увеличивается (уменьшается) с увеличением x.

Для иллюстрации примем довольно нелепое предположение, что номер эксперта в табл. 8.11 обозначает стаж его работы в качестве эксперта. Тогда интересно выяснить, систематически ли изменяется вероятность решения эксперта о наличии у больных психических расстройств с ростом его стажа. По данным табл. 8.11

$$\bar{x} = 4{,}5, \text{ а } \sum (x_j - \bar{x})^2 = 42.$$

Найдем по (8.38) коэффициент регрессии для прямой, связывающей p с x:

$$b = \frac{7{,}5}{8 \cdot 42} = 0{,}02.$$

Это соответствует среднему увеличению вероятности решения о наличии расстройств на год стажа, равному 0,02. Соответствующее значение (8.38) составит

$$\chi^2_{\text{slope}} = \frac{8 \cdot 7 \cdot 64 \cdot 42}{8 \cdot 15 - 71} (0{,}02)^2 = 1{,}23,$$

т. е. оно незначимо при сколь-нибудь разумном уровне значимости. Следовательно, мы не имеем права утверждать, что пропорция числа психических больных систематически изменяется с увеличением стажа эксперта.

Описанные в этом разделе статистики критериев не меняют значение при удалении m-связок, в которых все m реакций положительны или все отрицательны. Бергер и Гоулд [Berger and Gold, 1973], а также Бхапкар и Соумс [Bhapkar and Somes, 1977] показали, что в больших выборках при гипотезе

о равенстве истинных вероятностей Q имеет приближенно распределение хи-квадрат с $m-1$ степенями свободы, только при условии, что все парные вероятности $P(X_{ni}=1$ и $X_{nj}=1;$ $i \neq j)$ равны. Сиджер и Габриелссон [Seeger and Gabrielsson, 1968], а затем Тэйт и Браун [Tate and Brown, 1970] изучили точность аппроксимации распределения Q с помощью распределения хи-квадрат в случае, когда это условие не выполняется и когда размеры выборок малы. Видимо, эта аппроксимация приемлема, если произведение числа выборок (m) и числа m-связок, остающихся после удаления связок с одинаковой реакцией всех объектов в связке, превышает 24. В табл. 8.11 четыре пациента (под номерами 1, 3, 4, 7) одинаково проклассифицированы всеми экспертами. Произведение $m=8$ и числа оставшихся пациентов (4) равно 32. Это означает, что аппроксимация подходит.

Беннетт [Bennett, 1967, 1968] предложил метод сравнения m связанных выборок, сильно отличающийся от подхода Кохрэна [Cochran, 1959]. Информацию о статистиках критериев (их выражения сложнее Q, за исключением случая $m=3$; см. [Mantel and Fleiss, 1975]) и их мощности читатель найдет в этих двух работах Беннета.

8.5. Преимущества и недостатки связывания

В сравнительных проспективных и ретроспективных исследованиях связывание объектов обычно применяется для обеспечения сходства между сравниваемыми выборками по мешающим факторам, сопряженным с изучаемыми факторами (см., например, [Billewicz, 1965; Miettinen, 1970 a]). Следовательно, возможное увеличение эффективности исследований за счет использования связанных выборок (т. е. повышение мощности критериев значимости и точности оценки степени связи) имеет второстепенную важность. Тем не менее этому вопросу было уделено некоторое внимание. Кохрэн [Cochran, 1950] и Вустер [Worcester, 1964] показали, что связывание не обязательно повышает эффективность. Она будет повышаться, если есть сильная зависимость между факторами, по которым проводится связывание и факторами, которые составляют предмет изучения. Если зависимость между факторами связывания и исследуемыми факторами слаба или вовсе отсутствует, то, согласно результату Юклиса [Youkeles, 1963], эффективность может даже понижаться. Если же число связок превышает 30, то связывание по мешающим факторам, видимо, не влияет на эффективность.

Напротив, в планируемых клинических испытаниях со случайным назначением лечений различного вида главной целью

связывания является повышение эффективности. Чейс [Chase, 1968] доказал, что связывание по меньшей мере настолько же эффективно, как его отсутствие. С другой стороны, Биллевич [Billewicz, 1964, 1965] показал, что увеличение эффективности часто весьма ограничено и может не оправдывать усилия, затрачиваемые на поиск подходящих друг другу в связку объектов. Он продемонстрировал, как увеличивается время, требующееся для завершения исследования, с увеличением числа факторов связывания или с уменьшением относительных частот в каких-либо категориях факторов связывания. Если число факторов связывания слишком велико или если велико число категорий некоторых факторов связывания (даже при небольшом числе последних), то в конце сбора данных исследователь может обнаружить, что осталось много несвязанных объектов.

В сравнительных проспективных и ретроспективных исследованиях связывание или какие-либо другие методы контроля мешающих факторов часто необходимы, чтобы обеспечить их сопоставимость. Кроме связывания, контроль может осуществляться с помощью расслоения или регрессионными методами [Cochran, 1968; Rubin, 1973; McKinlay, 1975a]. Если выбранным средством контроля является расслоение, применимы методы анализа, описанные в гл. 10.

МакКинлей [McKinlay, 1977] провела критический анализ связывания, цель которого контроль нежелательных мешающих факторов, и указала в качестве основного достоинства его доступность в понимании и относительную простоту анализа получаемых данных. Главными недостатками названы, во-первых, возможные дополнительные затраты, обусловленные как поиском подходящих объектов, так и необходимостью исключать из анализа несвязанные объекты, и во-вторых, вероятность не обнаружить имеющееся *взаимодействие* — явление, при котором величина различия или связи изменяется от подгруппы к подгруппе. МакКинлей [McKinlay, 1975b, 1977] указывает также, что, хотя связывание может обеспечить большую точность, чем планы, в которых не учитываются мешающие факторы, тем не менее оно не обязательно дает большую точность, нежели другие планы с контролем мешающих факторов.

Тем не менее вполне может случиться, что исследователь, проводя сравнительное проспективное или ретроспективное исследование, предпочтет связывание другим методам контроля мешающих факторов. Например, при исследовании с помощью госпитализированных больных лучшим методом учета возможного наличия эпидемии, видимо, является связывание по дате госпитализации. Проводить связывание, однако, следует по

небольшому числу факторов (как правило, не более чем по четырем, а лучше всего по двум), каждый из которых определяется небольшим числом категорий (для возраста, например, обычно достаточна градация на интервалы по 10 лет). Если исследователь все же желает одновременно вести контроль по большому числу мешающих факторов, можно воспользоваться многомерными методами, предложенными в [Althouser and Rubin, 1970; Miettinen, 1976].

Задачи

8.1. Пользуясь данными табл. 8.6, определите, значимо ли различаются решения диагностов по пропорциям числа пациентов, отнесенных ими к больным с эмоциональным расстройством, с шизофренией, с заболеванием, отличным от двух первых? (*Указание.* Сравните значение статистики МакНемара с критическим значением хи-квадрат с двумя степенями свободы.)

8.2. Рассмотрите данные табл. 8.8.

а) Вычислите значение статистики Стюарта — Максвелла (8.19). Значимо ли различаются распределения результирующего фактора для пациентов, которых лечили новым методом, и для пациентов с традиционным методом лечения.

б) Вычислите значения d_1, d_3, $d_1 - d_3$. Что означает знак этой разности для различия двух методов лечения? Вычислите значение статистики критерия (8.20). Значимо ли новый метод лучше традиционного?

8.3. Без поправки на непрерывность статистика МакНемара выглядит так

$$\chi^2_u = \frac{(b-c)^2}{b+c}.$$

Покажите, что при $m=2$ выражение (8.25) эквивалентно χ^2_u. (*Указание.* В обозначениях табл. 8.1 выражение (8.21) равно $a+c$, (8.22) равно $2a+b+c$, выражение $\Sigma n^2/t$ равно $4a+b+c$.)

8.4. Докажите, что при $m=2$ значение Q (8.30) равно χ^2_u. (*Указание.* Докажите, что $T_1=a+c$; $T_2=a+b$; $T=2a+b+c$; $\Sigma S_n^2 = 4a+b+c$ при $m=2$.)

ЛИТЕРАТУРА

Althouser, R. P. and Rubin, D. B. (1970). The computerized construction of a matched sample. *Am. J. Sociol.*, **76**, 325–346.

Bennett, B. M. (1967). Tests of hypotheses concerning matched samples. *J. R. Stat. Soc., Ser. B*, **29**, 468–474.

Bennett, B. M. (1968). Note on X^2 tests for matched samples. *J. R. Stat. Soc., Ser. B*, **30**, 368–370.

Bennett, B. M. and Underwood, R. E. (1970). On McNemar's test for the 2 × 2 table and its power function. *Biometrics*, **26**, 339–343.

Berger, A. and Gold, R. Z. (1973). Note on Cochran's Q-test for the comparison of correlated proportions. *J. Am. Stat. Assoc.*, **68**, 989–993.

Bhapkar, V. P. (1966). A note on the equivalence of two test criteria for hypotheses in categorical data. *J. Am. Stat. Assoc.*, **61**, 228–235.

Bhapkar, V. P. and Somes, G. W. (1977). Distribution of Q when testing equality of matched proportions. *J. Am. Stat. Assoc.*, **72**, 658–661.

Billewicz, W. Z. (1964). Matched samples in medical investigations. *Brit. J. Prev. Soc. Med.*, **18**, 167–173.

Billewicz, W. Z. (1965). The efficiency of matched samples: An empirical investigation. *Biometrics*, **21**, 623–644.

Bross, I. D. J. (1969). How case-for-case matching can improve design efficiency. *Am. J. Epidemiol.*, **89**, 359–363.

Chase, G. R. (1968). On the efficiency of matched pairs in Bernoulli trials. *Biometrika*, **55**, 365–369.

Cochran, W. G. (1950). The comparison of percentages in matched samples. *Biometrika*, **37**, 256–266.

Cochran, W. G. (1968). The effectiveness of subclassification in removing bias in observational studies. *Biometrics*, **24**, 295–313.

Cornfield, J. and Haenszel, W. (1960). Some aspects of retrospective studies. *J. Chronic Dis.*, **11**, 523–534.

Cox, D. R. (1966). A simple example of a comparison involving quantal data. *Biometrika*, **53**, 215–220.

Doll, R. and Hill, A. B. (1952). A study of the etiology of carcinoma of the lung. *Brit. Med. J.*, **2**, 1271–1286.

Edwards, A. L. (1948). Note on the "correction for continuity" in testing the significance of the difference between correlated proportions. *Psychometrika*, **13**, 185–187.

Ejigou, A. and McHugh, R. (1977). Estimation of relative risk from matched pairs in epidemiologic research. *Biometrics*, **33**, 552–556.

Everitt, B. S. (1977). *The analysis of contingency tables.* London: Chapman and Hall.

Fleiss, J. L. (1965a). Estimating the accuracy of dichotomous judgments. *Psychometrika*, **30**, 469–479.

Fleiss, J. L. (1965b). A note on Cochran's Q test. *Biometrics*, **21**, 1008–1010.

Fleiss, J. L. and Everitt, B. S. (1971). Comparing the marginal totals of square contingency tables. *Brit. J. Math. Stat. Psychol.*, **24**, 117–123.

Grizzle, J. E., Starmer, C. F., and Koch, G. G. (1969). Analysis of categorical data linear models. *Biometrics*, **25**, 489–504.

Hill, A. B. (1962). *Statistical methods in clinical and preventive medicine.* New York: Oxford University Press.

Ireland, C. T., Ku, H. H., and Kullback, S. (1969). Symmetry and marginal homogeneity of an $r \times r$ contingency table. *J. Am. Stat. Assoc.*, **64**, 1323–1341.

Koch, G. G. and Reinfurt, D. W. (1971). The analysis of categorical data from mixed models. *Biometrics*, **27**, 157–173.

McKinlay, S. M. (1975a). The design and analysis of the observational study: A review. *J. Am. Stat. Assoc.*, **70**, 503–520.

McKinlay, S. M. (1975b). A note on the chi-square test for pair-matched samples. *Biometrics*, **31**, 731–735.

McKinlay, S. M. (1977). Pair-matching: A reappraisal of a popular technique. *Biometrics*, **33**, 725–735.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, **12**, 153–157.

Mantel, N. and Fleiss, J. L., (1975). The equivalence of the generalized McNemar tests for marginal homogeneity in 2^3 and 3^2 tables. *Biometrics*, **31**, 727–729.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.*, **22**, 719–748.

Maxwell, A. E. (1970). Comparing the classification of subjects by two independent judges. *Brit. J. Psychiatry*, **116**, 651–655.

Miettinen, O. S. (1968). The matched pairs design in the case of all-or-none responses. *Biometrics*, **24**, 339–352.

Miettinen, O. S. (1969). Individual matching with multiple controls in the case of all-or-none responses. *Biometrics*, **25**, 339–355.

Miettinen, O. S. (1970a). Matching and design efficiency in retrospective studies. *Am. J. Epidemiol.*, **91**, 111–118.

Miettinen, O. S. (1970b). Estimation of relative risk from individually matched series. *Biometrics*, **26**, 75–86.

Miettinen, O. S. (1976). Stratification by a multivariate confounder score. *Am. J. Epidemiol.*, **104**, 609–620.

Miller, R. G. (1966). *Simultaneous statistical inference*. New York: McGraw-Hill.

Mosteller, F. (1952). Some statistical problems in measuring the subjective response to drugs. *Biometrics*, **8**, 220–226.

Pike, M. C. and Morrow, R. H. (1970). Statistical analysis of patient-control studies in epidemiology: Factor under investigation an all-or-none variable. *Brit. J. Prev. Soc. Med.*, **24**, 42–44.

Rubin, D. B. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, **29**, 185–203.

Sceger, P. and Gabrielsson, A. (1968). Applicability of the Cochran Q test and the F test for statistical analysis of dichotomous data for dependent samples. *Psychol. Bull.*, **69**, 269:277.

Stuart, A. (1955). A test for homogeneity of the marginal distribution in a two-way classification. *Biometrika*, **42**, 412–416.

Stuart, A. (1957). The comparison of frequencies in matched samples. *Brit. J. Stat. Psychol.*, 10, 29–32.

Tate, M. W. and Brown, S. M. (1970). Note on the Cochran Q test. *J. Am. Stat. Assoc.*, 65, 155–160.

Ury, H. K. (1975). Efficiency of case-control studies with multiple controls per case: Continuous or dichotomous data. *Biometrics*, **31**, 643–649.

Worcester, J. (1964). Matched samples in epidemiologic studies. *Biometrics*, **20**, 840–848.

Youkeles, L. H. (1963). Loss of power through ineffective pairing of observations in small two-treatment all-or-none experiments. *Biometrics*, **19**, 175–180.

Глава 9

Сравнение пропорций в нескольких независимых выборках

В предыдущих главах, за редким исключением, мы ограничивались сравнением двух пропорций. В этой главе мы обсудим сравнение нескольких пропорций. В разд. 9.1 мы исследуем таблицу сопряженности $m \times 2$, $m > 2$, когда m групп не упорядочены. Разд. 9.2 и 9.3 посвящены случаю, когда существует естественное упорядочение m групп. В разд. 9.2 мы рассмотрим гипотезу, согласно которой пропорции изменяются монотонно (т. е. монотонно возрастают или уменьшаются) для m упорядоченных групп, причем пропорция линейно зависит от номера группы. В разд. 9.3 — общую гипотезу монотонности пропорций для упорядоченных групп. В разд. 9.4 мы обсудим сравнение двух или более групп по шкале качественно упорядоченных категорий.

Процедуры, описанные в этой главе, подходят для любого из трех методов извлечения выборок, рассмотренных ранее (разд. 2.1). В методе выбора III m выборок представляют группы, к которым применялись m различных обработок, причем объекты распределены по группам случайно. В методе II исследователь выбирает либо определенное число объектов из каждой среди m групп, либо определенное число объектов, обладающих и не обладающих некоторым свойством. В методе выбора I эти числа становятся известными только по завершении сбора данных. Как и в случае сравнения двух выборок ($m=2$) (см. разд. 6.1 и 6.2), при $m>2$ метод II с равными объемами выборок превосходит по мощности и точности метод I.

9.1. Сравнение m пропорций

Предположим, что исследуется m выборок. Каждый объект характеризуется наличием или отсутствием некоторого свойства.

Таблица 9.1

Пропорции в m независимых выборках

Выборка	Объем выборки	Число объектов с признаком	Число объектов без признака	Пропорция объектов с признаком
1	$n_{1.}$	n_{11}	n_{12}	p_1
2	$n_{2.}$	n_{21}	n_{22}	p_2
.				
.				
m	$n_{m.}$	n_{m1}	n_{m2}	p_m
Всего	$n_{..}$	$n_{.1}$	$n_{.2}$	$\bar{p}$

овокупность данных можно представить в виде табл. 9.1, в которой

$$p_i = \frac{n_{i1}}{n_{i.}}, \tag{9.1}$$

$$\bar{p} = \frac{n_{.1}}{n_{..}} = \frac{\Sigma\, n_{i.}\, p_i}{\Sigma\, n_{i.}}. \tag{9.2}$$

Для проверки значимости различия m пропорций можно равнить величину

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{2} \frac{(n_{ij} - n_{i.}\ n_{.j}/n_{..})^2}{n_{i.}\ n_{.j}/n_{..}} \tag{9.3}$$

критическими значениями таблицы распределения хи-квад-.1 (см. табл. А.1) с $m-1$ степенью свободы. Эквивалентное более наглядное выражение статистики этого критерия записывается в виде:

$$\chi^2 = \frac{1}{\bar{p}\,\bar{q}} \sum_{i=1}^{m} n_{i.}\ (p_i - \bar{p})^2, \tag{9.4}$$

де $\bar{q} = 1 - \bar{p}$.

Рассмотрим, например, данные (табл. 9.2) четырех обследований, обсуждаемых Дорном [Dorn, 1954]. В каждом обсле-

довании было зарегистрировано число курящих среди больных раком легких.

Таблица 9.

Пристрастие к курению среди больных раком легких по данным четырех исследований

Исследование	Число пациентов	Число курящих	Пропорция курящих
1	$86(=n_{1.})$	83	$0,965(=p_1)$
2	$93(=n_{2.})$	90	$0,968(=p_2)$
3	$136(=n_{3.})$	129	$0,949(=p_3)$
4	$82(=n_{4.})$	70	$0,854(=p_4)$
Всего	$397(=n..)$	372	$0,937\ (=\bar{p})$

Для этих данных значение χ^2 (9.4) есть

$$\chi^2 = \frac{1}{0,937 \cdot 0,063}[86 \cdot (0,965 - 0,937)^2 + 93 \cdot (0,968 - 0,937)^2 + 136 \cdot (0,949 - 0,937)^2 + 82 \cdot (0,854 - 0,937)^2] = 12,56, \quad (9.5$$

которое значимо на уровне 0,01 при трех степенях свободы.

Если выяснилось, что пропорции различаются значимо, можно попытаться выделить выборку или группу выборок, которые обусловливают различие. Методы выделения источников значимого различия для таблицы сопряженности общего вида приведены в следующих работах [Irwin, 1949; Lancaster, 1950; Kimball, 1954; Kastenbaum, 1960; Castellan, 1965; Knoke, 1976]. Здесь мы опишем простейший метод для таблицы $m \times 2$.

Предположим, что m выборок разделены на две группы. В первой группе — m_1 выборок, во второй — m_2, $m_1 + m_2 = m$. Определим

$$n_{(1)} = \sum_{i=1}^{m_1} n_{i.} \quad (9.6$$

— суммарное число объектов в первой группе и

$$n_{(2)} = \sum_{i=m_1+1}^{m} n_{i.} \quad (9.7$$

— во второй группе.

Пусть пропорция в первой группе —

$$\bar{p}_1 = \frac{\sum_{i=1}^{m_1} n_{i.}\, p_i}{n_{(1)}}, \tag{9.8}$$

во второй группе —

$$\bar{p}_2 = \frac{\sum_{i=m_1+1}^{m} n_{i.}\, p_i}{n_{(2)}}. \tag{9.9}$$

Тогда выражение

$$\chi^2_{\text{diff}} = \frac{1}{\bar{p}\,\bar{q}} \cdot \frac{n_{(1)}\, n_{(2)}}{n_{..}} (\bar{p}_1 - \bar{p}_2)^2 \tag{9.10}$$

с одной степенью свободы можно использовать для проверки значимости различия между $\bar{p}_1$ и $\bar{p}_2$. Заметим, что χ^2_{diff} распределен по хи-квадрат без поправок на непрерывность. В отличие от (9.10) эти поправки следовало бы использовать для четырехклеточной таблицы, полученной при объединении всех данных из первых m_1 выборок в первое множество и всех данных из остальных m_2 выборок во второе множество.

Статистику

$$\chi^2_1 = \frac{1}{\bar{p}\,\bar{q}} \sum_{i=1}^{m_1} n_{i.}\, (p_i - \bar{p}_1)^2 \tag{9.11}$$

с $m_1 - 1$ степенями свободы можно использовать для проверки значимости различия m_1 пропорций первой группы. Аналогично статистику

$$\chi^2_2 = \frac{1}{\bar{p}\,\bar{q}} \sum_{i=m_1+1}^{m} n_{i.}\, (p_i - \bar{p}_2)^2 \tag{9.12}$$

с $m_2 - 1$ степенями свободы можно использовать для второй группы. Легко проверить, что статистики (9.10)—(9.12) дают в сумме значение χ^2 в (9.4).

Если $\bar{p}_1$ и $\bar{p}_2$ заметно различаются, выражение $\bar{p}\bar{q}$ следует заменить выражениями $\bar{p}_1\bar{q}_1 = \bar{p}_1(1-\bar{p}_1)$ в (9.11) и $\bar{p}_2\bar{q}_2 = \bar{p}_2(1-\bar{p}_2)$ (9.12). Эти поправки несильно влияют на величину χ^2, но сумма (9.11) и (9.12) с этими поправками и (9.10) в общем случае уже не равна значению (9.4).

Если же разбиение выборок на группы диктуется самими данными, а не планом эксперимента, необходимы более серьезные модификации методов. Например, первые три из четырех выборок в табл. 9.2, судя по сходству их пропорций, образуют одну однородную группу, в то время как четвертую выборку можно выделить как вторую группу. Чтобы избежать ошибочных выводов, которые возможны при сравнениях, диктуемых данными, нужно сопоставить величины (9.10)—(9.12) с критическим значением хи-квадрат с $m-1$ степенью свободы, но не с критическими значениями хи-квадрат с 1, m_1-1 и m_2-1 степенями свободы соответственно [Miller, 1966, Section 6.2].

Например, для данных табл. 9.2 первое множество, образованное тремя группами, $m_1=3$, состоит из

$$n_{(1)} = 86 + 93 + 136 = 315$$

больных раком легких, среди которых доля курильщиков составляет:

$$\bar{p}_1 = \frac{83 + 90 + 129}{315} = 0{,}959 .$$

Второе множество, представляющее только одно обследование, $m_2=1$, состоит из $n_2=82$ пациентов, среди которых доля курильщиков — $\bar{p}_2=0{,}854$.

Значимость различия между $\bar{p}_1$ и $\bar{p}_2$ определяется величиной χ^2_{diff} (9.10):

$$\chi^2_{\text{diff}} = \frac{1}{0{,}937 \cdot 0{,}063} \cdot \frac{315 \cdot 82}{397} (0{,}959 - 0{,}854)^2 = 12{,}15 . \qquad (9.13)$$

Значимость различий среди p_1, p_2, p_3 (все эти пропорции из группы 1) определяется величиной χ^2_1 (9.11):

$$\chi^2_1 = [86 \cdot (0{,}965 - 0{,}959)^2 + 93 \cdot (0{,}968 - 0{,}959)^2 + 136 \times$$
$$\times (0{,}949 - 0{,}959)^2]/(0{,}937 \cdot 0{,}063) = 0{,}41 . \qquad (9.14)$$

Статистика χ^2_2 здесь неприменима, поскольку группа 2 состоит из одной выборки.

Прежде всего заметим, что

$$\chi^2_{\text{diff}} + \chi^2_1 = 12{,}15 + 0{,}41 = 12{,}56 .$$

Эта величина равна значению суммарной статистики хи-квадрат (9.5). Далее, замена произведения $0{,}937 \cdot 0{,}063$ в (9.14) на значение $\bar{p}_1\bar{q}_1 = 0{,}959 \cdot 0{,}041$ лишь слегка увеличивает значение χ^2_1 (до 0,62). Наконец, вспомним, что разделение обусловлено самими данными, а не планировалось априорно. Поэтому обе величины χ^2_{diff} и χ^2_1 надо сравнивать с критическим значением хи-квадрат с $m-1=3$ степенями свободы. Поскольку

критическое значение для уровня значимости 0,05 есть 7,81, будет получен вывод, что пропорция курильщиков среди больных в обследовании 4 отличается от пропорций в обследованиях 1—3 (так как $\chi^2_{\text{diff}} = 12{,}15 > 7{,}81$), а пропорции в обследованиях 1—3 не различаются (так как $\chi^2_1 = 0{,}41 < 7{,}81$).

9.2. Монотонное поведение пропорций в случае количественно упорядоченных выборок

Методы предыдущего раздела справедливы в довольно общем случае, но неэффективны при естественном упорядочении m выборок, если оно существует. В этом разделе мы предполагаем, что выборки упорядочены на количественной шкале, в частности, будем считать, что номер выборки прямо связан со значением показателя x. Используем данные [National Center for Health Statistics, 1970, Table 1 and 6] в качестве примера (табл. 9.3).

Таблица 9.3

Увеличение числа женщин, жалующихся на бессонницу, с возрастом

Возрастной интервал	Число женщин ($=n_{i.}$)	Пропорция жалующихся на бессонницу ($=p_i$)	Середина интервала ($=x_i$)
18—24	534	0,280	21,5
25—34	746	0,335	30,0
35—44	784	0,337	40,0
45—54	705	0,428	50,0
55—64	443	0,538	60,0
65—74	299	0,590	70,0
Всего	3511 ($=n_{..}$)	0,393 ($=\bar{p}$)	42,15 ($=\bar{x}$)

В зависимости от выдвигаемой гипотезы о поведении пропорций необходимы различные методы анализа (см. [Yates, 1948]). Здесь мы рассмотрим зависимость простейшего вида — линейную. Пусть P_i обозначает пропорцию в популяции, из которой была извлечена i-я выборка. Мы проверяем гипотезу, что

$$P_i = \alpha + \beta x_i, \qquad (9.15)$$

где β — коэффициент регрессии, указывающий величину изменения пропорции с изменением x на единицу, а α — свободный член, равный ожидаемой пропорции, когда $x=0$.

Эти два параметра можно оценить так. Определим среднее значение x в полученном наборе данных:

$$\bar{x} = \sum_{i=1}^{m} \frac{n_{i.}\, x_i}{n_{..}} \,. \tag{9.16}$$

Оценками для коэффициента β и свободного члена будут

$$b = \frac{\sum_{i=1}^{m} n_{i.}\,(p_i - \bar{p})\,(x_i - \bar{x})}{\sum_{i=1}^{m} n_{i.}\,(x_i - \bar{x})^2} \tag{9.17}$$

и

$$a = \bar{p} - b\,\bar{x}. \tag{9.18}$$

Вычисление b можно несколько упростить, заметив, что числитель в (9.17) есть

$$\sum_{i=1}^{m} n_{i.}\, p_i\, x_i - n_{..}\,\bar{p}\,\bar{x}\,, \tag{9.19}$$

а знаменатель —

$$\sum_{i=1}^{m} n_{i.}\, x_i^2 - n_{..}\,\bar{x}^2\,. \tag{9.20}$$

Для описания зависимости получаем простое уравнение

$$\hat{p}_i = \bar{p} + b\,(x_i - \bar{x})\,. \tag{9.21}$$

Для данных табл. 9.3 $\bar{p}=0{,}393$, $\bar{x}=42{,}15$ и

$$b = 0{,}0064\,. \tag{9.22}$$

$$\hat{p}_i = 0{,}393 + 0{,}0064\;(x_i - 42{,}15)\,, \tag{9.23}$$

что подразумевает увеличение доли взрослых женщин, жалующихся на бессонницу, на 0,64 % с увеличением возраста на год.

Полезно вычислить оценку пропорции, соответствующей каждому x_i, чтобы сравнить ее с выборочной, p_i. Если значения p_i и $\hat{p}_i$ близки для всех или большинства категорий, то

ожно заключить, что (9.15) хорошо описывает данные, т. е. P_i изменяется приблизительно линейно с ростом x_i. Если различие между p_i и $\hat{p}_i$ велико, следует считать связь между P_i и x_i сложнее линейной. Вычислив разности $p_i - \hat{p}_i$, можно также выделять категории с наибольшим отклонением от линейности.

В табл. 9.4 сравниваются пропорции из табл. 9.3 и их оценки (9.23). Согласие выглядит вполне удовлетворительным.

Таблица 9.4

Наблюдаемые и линейно предсказываемые доли жалующихся на бессонницу в зависимости от возраста

x_i	$n_{i.}$	p_i	$\hat{p}_i$
21,5	534	0,280	0,261
30,0	746	0,335	0,315
40,0	784	0,337	0,379
50,0	705	0,428	0,443
60,0	443	0,538	0,507
70,0	299	0,590	0,571

Для проверки линейности связи между P_i и x_i можно использовать статистику хи-квадрат [Cochran, 1954; Armitage, 955]:

$$\chi^2_{\text{lin}} = \sum_{i=1}^{m} n_{i.} \, (p_i - \hat{p}_i)^2 / \bar{p}\bar{q} \qquad (9.24)$$

$m-2$ степенями свободы. Гипотеза линейности будет отвергнута, если значение χ^2_{lin} будет большим. Мощность этого критерия изучалась Чемпеном и Нэмом [Chapman and Nam, 1968].

Вычисление χ^2_{lin} упрощается, если сначала вычислить

$$\chi^2_{\text{slope}} = b^2 \sum_{i=1}^{m} n_{i.} \, (x_i - \bar{x})^2 / \bar{p}\bar{q}, \qquad (9.25)$$

так как можно показать, что

$$\chi^2_{\text{lin}} = \chi^2 - \chi^2_{\text{slope}}, \qquad (9.26)$$

где χ^2 дано выражением (9.4). Статистику χ^2_{slope} с одной степенью свободы можно использовать для проверки гипотезы о коэффициенте β. Если значение χ^2_{slope} велико, можно сделать вывод, что коэффициент значимо отличается от нуля, и, следовательно, P_i увеличивается с ростом x_i, если b положителен, или уменьшается, если b отрицателен.

Для данных табл. 9.3 значение суммарной статистики хи-квадрат (9.4) для проверки гипотезы о равенстве доли страдающих бессонницей во всех группах есть

$$\chi^2 = 140{,}72. \tag{9.27}$$

Это значение (при пяти степенях свободы) указывает на высокую значимость различия пропорций в рассматриваемых возрастных группах. Однако с ее помощью нельзя выявить увеличение доли страдающих бессонницей с возрастом.

Для данных табл. 9.4 статистика хи-квадрат (9.24), используемая для проверки линейности,—

$$\chi^2_{\text{lin}} = \frac{534\cdot(0{,}280-0{,}261)^2 + \ldots + 299\cdot(0{,}590-0{,}571)^2}{0{,}393\cdot 0{,}607} = 10{,}76 \tag{9.28}$$

с 4 степенями свободы, значима на уровне 0,05. Таким образом, связь доли женщин, жалующихся на бессонницу, с возрастом не является точно линейной, но отклонения от линейности (т. е. разности между наблюдаемой долей и ее линейным прогнозом) достаточно малы, чтобы считать гипотезу о линейности разумной.

Статистика хи-квадрат (9.25) имеет значение

$$\chi^2_{\text{slope}} = \frac{0{,}0064^2\cdot 757964{,}8}{0{,}393\cdot 0{,}607} = 130{,}15, \tag{9.29}$$

которое (при одной степени свободы) указывает, что коэффициент регрессии в уравнении (9.23), $b = 0{,}0064$, значимо отличается от нуля. Разность суммарной статистики хи-квадрат, 140,72, и хи-квадрат для проверки гипотезы о коэффициенте регрессии, 130,15, должна согласно (9.26) равняться хи-квадрат для проверки линейности, т. е. 10,76. Это верно, если пренебречь ошибками округления.

Выводы, следующие из этого более подробного анализа, состоят в том, что доля женщин, жалующихся на бессонницу, монотонно увеличивается с возрастом, и что характер этой зависимости практически линеен. Если бы линейность проверялась по статистике хи-квадрат, скажем, на уровне 0,01 или 0,005, а не на уровне всего лишь 0,05, последний вывод был бы неверен.

Немного отличные варианты оценок и статистик критериев, описанных выше, даны в [Mantel, 1963; Chapman and Nam, 1968; Wood, 1978], где также рассмотрены сравнение и объединение оценок регрессионных прямых по нескольким независимым выборкам. Процедуры, представленные здесь, справедливы, когда p_i достаточно удалены от 0 или 1, как в нашем примере.

9.3. Общая гипотеза о монотонном поведении пропорций для упорядоченных выборок

Мы предполагали в разд. 9.2, что m выборок могут быть упорядочены на количественной шкале. В этом разделе рассматривается более общий случай. Мы предполагаем лишь, что шкала категорий (выборок) порядковая. Допустим, мы имеем дело с данными табл. 9.5. Значение χ^2 (9.4) для этих данных есть

$$\chi^2 = 28{,}74 \qquad (9.30)$$

с 3 степенями свободы с высоким уровнем значимости (менее 0,001, см. табл. А.1).

Таблица 9.5

Вымышленные данные о долях пациентов, выздоровевших в течение одного месяца, в зависимости от начальной тяжести заболевания

Начальная тяжесть	Всего	Число выздоровевших за один месяц	Пропорция выздоровевших за один месяц
Слабая	30 ($=n_{1.}$)	25	0,83 ($=p_1$)
Умеренная	25 ($=n_{2.}$)	22	0,88 ($=p_2$)
Сильная	20 ($=n_{3.}$)	12	0,60 ($=p_3$)
Очень сильная	25 ($=n_{4.}$)	6	0,24 ($=p_4$)
Всего	100 ($=n_{..}$)	65	0,65($=\bar{p}$)

Вывод о том, что доли выздоровевших значимо различаются, справедлив, однако явно недостаточен в том смысле, что не описывает близкое к монотонному уменьшение доли выздоровевших с усилением тяжести заболевания. Гипотеза о монотонном поведении доли выздоровевших в зависимости от тяжести

заболевания представляется разумной, и, значит, необходим другой метод анализа. Метод, описанный в предыдущем разделе, не приемлем, поскольку невозможно естественным образом присвоить числовые значения четырем степеням тяжести.

Хассан [Chassan, 1960, 1962] предложил простой критерий для проверки гипотезы о том, что m пропорций расположены в определенном порядке, однако Бартоломью [Bartholomew, 1963] показал, что критерий Хассана неудовлетворителен. Этот критерий применим только в том случае, если выборочные пропорции точно упорядочены в соответствии с гипотезой и не применим, даже если отклонение от гипотетического порядка незначительно (например, для p_1 и p_2 в табл. 9.5). Мы опишем более мощную процедуру [Bartholomew, 1959a, 1959b].

Предположим, что порядок пропорций по гипотезе таков: $p_1 > p_2 > \ldots > p_m$, но наблюдаются отклонения от этого порядка. Например, для пропорций в табл. 9.5 ожидалось, что $p_1 > p_2 > p_3 > p_4$, однако мы получили упорядочение $p_2 > p_1 > p_3 > p_4$.

Если наблюдаются отклонения от предполагаемого порядка, вычисляют взвешенные средние тех соседних пропорций, которые нарушают порядок. Усреднения проводят, пока замена наблюдаемых пропорций усредненными не приведет к порядку, соответствующему гипотезе. Пересчитанные пропорции обозначим p'. Для пропорций табл. 9.5 надо вычислить среднее p_1 и p_2:

$$\bar{p}_{1,2} = \frac{30 \cdot 0{,}83 + 25 \cdot 088}{30 + 25} = 0{,}85. \qquad (9.31$$

Замена p_1 и p_2 на $\bar{p}_{1,2}$ приводит к табл. 9.6.

Таблица 9.6

Преобразование пропорций табл. 9.5 к гипотетическому порядку

Начальная степень тяжести	Всего	Преобразованные пропорции
Слабая	30 ($=n_{1.}$)	0,85 ($=p'_1$)
Умеренная	25 ($=n_{2.}$)	0,85 ($=p'_2$)
Сильная	20 ($=n_{3.}$)	0,60 ($=p_3=p'_3$)
Очень сильная	25 ($=n_{4.}$)	0,24 ($=p_4=p'_4$)
Всего	100 ($=n_{..}$)	0,65 ($=\bar{p}$)

Пересчитанные пропорции уже не нарушают порядок (в противном случае надо было бы продолжать усреднения). После пересчета вычисляется значение статистики

$$\bar{\chi}^2 = \frac{1}{\bar{p}\,\bar{q}} \sum_{i=1}^{m} n_{i.}\,(p_i' - \bar{p})^2. \qquad (9.32)$$

Для пропорций табл. 9.6

$$\bar{\chi}^2 = \frac{1}{0{,}65 \cdot 0{,}35}[30 \cdot (0{,}85-0{,}65)^2 + 25 \cdot (0{,}85-0{,}65)^2 +$$
$$+ 20 \cdot (0{,}60-0{,}65)^2 + 25 \cdot (0{,}24-0{,}65)^2] = 28{,}27. \qquad (9.33)$$

Однако теперь критическое значение надо определять уже не по таблицам распределения хи-квадрат, а по табл. А.5—А.7. Когда сравниваются три пропорции $m=3$, вычисляем

$$c = \sqrt{\frac{n_{1.}\,n_{3.}}{(n_{1.}+n_{2.})\,(n_{2.}+n_{3.})}} \qquad (9.34)$$

и ищем в табл. А.5 критическое значение, соответствующее выбранному уровню значимости, проводя, в случае необходимости интерполирование. Если $m=4$, вычисляем

$$c_1 = \sqrt{\frac{n_{1.}\,n_{3.}}{(n_{1.}+n_{2.})\,(n_{2.}+n_{3.})}} \qquad (9.35)$$

и

$$c_2 = \sqrt{\frac{n_{2.}\,n_{4.}}{(n_{2.}+n_{3.})\,(n_{3.}+n_{4.})}}, \qquad (9.36)$$

а затем находим в табл. А.6 критическое значение, осуществляя, если необходимо, интерполирование и по c_1, и по c_2. Если размеры выборок равны и $m \leqslant 12$, можно использовать табл. А.8.

Для данных табл. 9.6, где $m=4$,

$$c_1 = \sqrt{\frac{30 \cdot 20}{(30+25)\,(25+20)}} = 0{,}49,$$

$$c_2 = \sqrt{\frac{25 \cdot 25}{(25+20)\,(20+25)}} = 0{,}56.$$

Визуальная интерполяция по табл. А.6 (c_1 приближенно равно 0,5, а значение c_2 находится где-то посередине между 0,5 и 0,6) показывает, что значение $\bar{\chi}^2$ должно превысить 9,0, чтобы достичь уровня значимости 0,005. Полученное в (9.33) значение $\bar{\chi}^2 = 28{,}27$ много больше этого критического значения.

Теперь стоит сравнить значение, найденное по табл. А.6, с соответствующим значением в табл. А.1 для стандартного кри-

терия хи-квадрат с $m-1=3$ степенями свободы. Если гипотеза об упорядочении не выдвинута, при уровне значимости 0,005 χ^2 должно превышать 12,8 (а не как ранее 9,0). Таким образом, если гипотетическое упорядочение в популяции действительно имеет место, то критерий Бартоломью оказывается мощнее стандартного критерия хи-квадрат. Однако, если гипотетическое упорядочение не верно, процесс усреднения, проводимый перед вычислением $\bar{\chi}^2$ по (9.32), может уменьшить значение $\bar{\chi}^2$ до незначимого. Дальнейший анализ и обобщения критерия Бартоломью даны в [Barlow et al, 1972].

9.4. Ридит-анализ

Предположим, что данные представляют собой две или более выборок. Объекты в каждой выборке распределены по нескольким упорядоченным категориям. Пусть k обозначает число категорий. Будем, например, исследовать тяжесть травм, получаемых водителями в автомобильных авариях. Градациями тяжести могут быть взяты отсутствие травмы, слабая травма и т. д. до смертельной травмы. Такая шкала явно субъективна и, вероятно, не слишком надежна. Тем не менее эта шкала выглядит предпочтительнее простой дихотомии (легкая травма или отсутствие травмы и тяжелая или смертельная травма), поскольку она все же до некоторой степени достоверна и описывает событие полнее, чем если бы оно описывалось в более грубой системе «да — нет».

Возникает задача — как корректно представлять данные и сравнивать выборки. Когда сравниваются две выборки, данные можно представить в виде табл. 9.7. Пропорции (p_{11}, ..., p_{k1} указывают частотное распределение в выборке 1, а пропорции (p_{12}, ..., p_{k2}) — в выборке 2. Частотное распределение в объединенной выборке есть ($\bar{p}_1$, ..., $\bar{p}_k$), где

$$\bar{p}_i = \frac{n_1 p_{i1} + n_2 p_{i2}}{n}, \tag{9.37}$$

($i=1$, ..., k) и $n=n_1+n_2$. Значение статистики хи-квадрат с $k-1$ степенью свободы можно найти по формуле

$$\chi^2 = \frac{n_1 n_2}{n} \sum_{i=1}^{k} \frac{(p_{i1} - p_{i2})^2}{\bar{p}_i} \tag{9.38}$$

(см. задачу 9.4), но тогда будет утрачена принципиальная информация о естественном упорядочении k категорий.

Часто используют такой прием. Категориям присваивают номера от нуля (для наименее тяжелой травмы) до некоторого

Таблица 9.7

Распределение отиосительных частот в двух выборках

Категория	Выборка 1 (объем выборки $=n_1$)	Выборка 2 (объем выборки $=n_2$)	Объединенная выборка (объем выборки $=n$)
1	p_{11}	p_{12}	$\bar{p}_1$
2	p_{21}	p_{22}	$\bar{p}_2$
k	p_{k1}	p_{k2}	$\bar{p}_k$
Сумма	1	1	1

значения (для наиболее тяжелой травмы), а затем вычисляют средние и стандартные отклонения и применяют критерии, основанные на t-статистиках, или дисперсионный анализ. Этот способ, порождающий на первый взгляд систему числовых показателей, имеет много недостатков. Во-первых, он создает иллюзию более высокой точности, чем в действительности. Во-вторых, получаемые результаты зависят от выбираемой системы числовых характеристик, а ее разумный выбор далеко не прост.

Снова обратимся к исследованию травм при автомобильных авариях и будем считать, что градаций тяжести травмы семь, начиная с отсутствия травмы и слабой травмы и заканчивая очень тяжелой и смертельной. Естественно пронумеровать семь градаций последовательно от 0 до 6. Такую нумерацию нельзя считать объективной, так как она подразумевает, что разница между слабой травмой и отсутствием травмы эквивалентна разнице между очень тяжелой и смертельной травмами. Последние отличаются значительно сильнее, но это большее отличие можно отразить, лишь присвоив последней категории значение, превышающее 6. Однако решать, каким именно должно быть это значение, можно довольно произвольно. Если исходить из предположения, что справедлива логистическая модель (см. разд. 9.4), можно применять процедуру по [Shell, 1964].

Оставим попытки описать категории численными значениями и вместо этого будем работать только с естественным упоря-

дочением. Методом, в котором эффективно учитывается это естественное упорядочение, является ридит-анализ (ridit analysis). По сути, единственное предположение в ридит-анализе состоит в том, что дискретным категориям соответствуют интервалы шкалы для существующего, но ненаблюдаемого непрерывного распределения. Предположения о нормальности или другом виде распределения не используются.

Ридит-анализ введен Броссом [Bross, 1958] и применялся к исследованию автомобильных аварий [Bross, 1960], рака [Wynder et al., 1960] и шизофрении [Spitzer et al., 1965]. Результаты теоретического изучения ридит-анализа отражены в работе [Kantor et al., 1968]. Критика ридит-анализа изложена в [Mantel, 1979].

Ридит-анализ начинают с выбора популяции, которая используется как стандартная или как контрольная группа. Термин «ридит» образован от первых букв (rid) сочетания «relative to an identified distribution» (по сравнению с определенным распределением). Для контрольной группы надо оценить ридит-значения каждого интервала, т. е. пропорцию объектов, которые на предполагаемой непрерывной шкале имеют значение, равное или меньшее среднего значения интервала. Порядок арифметических вычислений продемонстрирован в табл. 9.8, где использованы данные из [Bross, 1958, p. 20].

1. Столбец 1 описывает распределение в контрольной группе по различным категориям. В табл. 9.8 представлено распределение травм по семи категориям в выборке объема 179.

2. Элементы столбца 2 — уменьшенные вдвое значения соответствующих элементов столбца 1.

3. Элементы столбца 3 — суммы значений столбца 1 в предыдущих категориях («накопленная» сумма со сдвигом вниз).

4. Элементы столбца 4 — суммы соответствующих элементов в столбцах 2 и 3.

5. Наконец, в столбце 5 — элементы столбца 4, поделенные на объем выборки (в данном случае — 179).

Результатом этих вычислений являются ридит-значения, соответствующие различным категориям. Таким образом, ридит-значение категории — не более, чем пропорция объектов контрольной группы в предыдущей категории плюс половина пропорции в данной категории. Если предполагаемая модель ненаблюдаемого непрерывного распределения — равномерное распределение в каждом интервале, то ридит-значение категории — пропорция объектов контрольной группы со значением, меньшим либо равным среднему соответствующего интервала.

Пусть известно распределение объектов другой выборки по тем же категориям. Тогда можно вычислить ридит-среднее (см. ниже) для этой выборки, которое можно интерпретировать как

вероятность. Ридит-среднее группы — это вероятность того, что случайно извлеченный из нее объект имеет значение с большим номером категории, чем объект, случайно извлеченный из контрольной группы.

В контексте нашего примера ридит-среднее 0,5 соответствует тому, что для сравниваемой группы не характерна ни меньшая, ни большая тяжесть травм, чем в контрольной группе. Заметим, что в контрольной группе ридит-среднее всегда равно 0,5. Это естественно, поскольку если два объекта случайно выбраны из одной популяции, то в половине случаев один объект будет иметь больший номер категории, а во второй половине случаев — другой объект.

Если ридит-среднее сравниваемой группы больше 0,5, то объект, случайно извлеченный из нее, будет чаще иметь бо́льший номер категории, чем объект, случайно извлеченный из контрольной группы.

Таблица 9.8

Как вычислять ридит-значения
(на примере данных о тяжести травм у водителей)

Степень тяжести	1	2	3	4	Ридит-значение
Отсутствует	17	8,5	0	8,5	0,047
Слабая	54	27,0	17	44,0	0,246
Умеренная	60	30,0	71	101,0	0,564
Сильная	19	9,5	131	140,5	0,785
Тяжелая	9	4,5	150	154,5	0,863
Очень тяжелая	6	3,0	159	162,0	0,905
Смертельная	14	7,0	165	172,0	0,961

В нашем примере мы бы говорили, что в сравниваемой группе травмы тяжелее, чем в контрольной. Если, наконец, ридит-среднее сравниваемой группы меньше 0,5, то можно заключить, что объекты из нее чаще имеют меньшие номера категорий, чем объекты из контрольной группы.

Как пример, рассмотрим гипотетические данные (табл. 9.9) о распределении тяжести травм, полученных при авариях водителями, находившимися в состоянии легкого опьянения.

Ридит-среднее группы — это просто сумма произведений наблюдаемой частоты и соответствующего ридит-значения, деленная на суммарную частоту (т. е. число объектов в группе). Ри-

Таблица 9.9

Тяжесть травм, полученных водителями в состоянии легкого опьянения в автомобильных авариях

Степень тяжести	Число водителей	Ридит-значение	Произведение
Отсутствует	5	0,047	0,235
Слабая	10	0,246	2,460
Умеренная	16	0,564	9,024
Сильная	5	0,785	3,925
Тяжелая	3	0,863	2,589
Очень тяжелая	6	0,905	5,430
Смертельная	5	0,961	4,805
Сумма	50		28,468

дит-среднее для водителей в легком опьянении равно:

$$\bar{r} = \frac{28{,}468}{50} = 0{,}57. \qquad (9.39)$$

Значит, шансы, что водитель в состоянии легкого опьянения получит более серьезную травму, чем водитель из контрольной группы, если они попадут в аварию, равны 4 к 3 (0,57/0,43).

Селвин [Selvin, 1977] показал, насколько тесно ридит-анализ связан с так называемыми ранговыми критериями, используемыми в непараметрической статистике, и как благодаря этому сходству можно найти стандартную ошибку ридит-среднего. Пусть N_j — число объектов контрольной выборки i-й категории; $N = \Sigma N_j$ — суммарное число объектов контрольной выборки; n_j и n — соответствующие величины сравниваемой выборки. Если объем контрольной выборки не намного больше объема сравниваемой выборки, то стандартная ошибка среднего в сравниваемой выборке есть

$$\text{s.e.}(\bar{r}) = \frac{1}{2\sqrt{3n}} \sqrt{1 + \frac{n+1}{N} + \frac{1}{N(N+n-1)} - \frac{\Sigma (N_i + n_j)^3}{N(N+n)(N+n-1)}}. \qquad (9.40)$$

Два частотных распределения из обсуждаемого примера приведены в табл. 9.10.

Таблица 9.10

Частотные распределения в контрольной и сравниваемой группах

Степень тяжести	Контрольная группа (N_j)	Сравниваемая группа (n_j)	Сумма (N_j+n_j)
Отсутствует	17	5	22
Слабая	54	10	64
Умеренная	60	16	76
Сильная	19	5	24
Тяжелая	9	3	12
Очень тяжелая	6	6	12
Смертельная	14	5	19
Сумма	179 ($=N$)	50 ($=n$)	229 ($=N+n$)

Стандартная ошибка ридит-среднего для водителей, попавших в аварию в состоянии легкого алкогольного опьянения, согласно (9.40) равна:

$$\text{s.e.}(\bar{r}) = \frac{1}{2\sqrt{150}}\sqrt{1+\frac{51}{179}+\frac{1}{179\cdot 228}-\frac{735{,}907}{179\cdot 229\cdot 228}} = 0{,}045. \tag{9.41}$$

Значимость различия ридит-средних (т. е. полученного значения и 0,5) можно установить, вычислив

$$z = \frac{\bar{r}-0{,}5}{\text{s. e. }(\bar{r})} \tag{9.42}$$

и обратившись к табл. А.2 нормального распределения. В нашем примере

$$z = \frac{0{,}57-0{,}50}{0{,}045} = 1{,}56. \tag{9.43}$$

Поскольку z не достигает критического значения, мы должны заключить, что травмы, полученные водителями в состоянии легкого опьянения, не тяжелее травм у водителей в контрольной группе.

Когда объем контрольной группы N намного больше объема сравниваемой группы, выражение для стандартной ошибки упрощается до

$$\text{s.e.}(\bar{r}) = \frac{1}{2\sqrt{3n}}. \qquad (9.44)$$

Для данных нашего примера эта аппроксимация дает оценку стандартной ошибки 0,041, что незначительно меньше значения 0,045, полученного по точной формуле (9.40).

Предположим теперь, что мы должны проанализировать выборку из 50 водителей, попавших в аварию в состоянии сильного алкогольного опьянения. Предположим также, что ее ридит-среднее равно 0,73. Интересно сравнить травмы водителей в легком и сильном опьянении. Теперь нам не нужно определять новую контрольную группу. Все, что требуется, это вычесть одно ридит-среднее из другого и прибавить 0,5. Тогда мы получим значение 0,66 (0,73—0,57+0.50), т. е. шансы, что травма водителя в состоянии сильного опьянения будет тяжелее травмы водителя в состоянии легкого опьянения, равны примерно 2 к 1.

Если одно ридит-среднее вычислено по выборке объема N_1, а второе — по выборке объема N_2, то стандартная ошибка равна приблизительно

$$\text{s.e.}(\bar{r}_2 - \bar{r}_1) = \frac{\sqrt{N_1 + N_2}}{2\sqrt{3\,N_1 N_2}}. \qquad (9.45)$$

Эта формула является хорошим приближением ошибки ридит-среднего для случая с одной сравниваемой группой, если величины N и n сравнимы по величине, а N_1 и N_2 заменяют на N и n (см. выражение (9.40)). Такая оценка стандартной ошибки завышена, но обычно это завышение незначительно (см. задачу 9.5, задания в и г).

При $N_1 = N_2 = 50$ стандартная ошибка приближенно равна:

$$\text{s.e.}(\bar{r}_2 - \bar{r}_1) = \frac{\sqrt{100}}{2\sqrt{3 \cdot 50 \cdot 50}} = 0{,}06. \qquad (9.46)$$

Значимость различия $\bar{r}_1$ и $\bar{r}_2$ можно проверить, сопоставляя

$$z = \frac{\bar{r}_2 - \bar{r}_1}{\text{s.e.}(\bar{r}_2 - \bar{r}_1)} \qquad (9.47)$$

со значениями в табл. А.2. В нашем примере

$$z = \frac{0{,}73 - 0{,}57}{0{,}06} = 2{,}67, \qquad (9.48)$$

указывает на различие ридит-средних при уровне значимости 0,01. Мы пришли к выводу, что водители, попавшие в аварии в состоянии сильного опьянения, чаще получают тяжелые травмы, чем водители в состоянии легкого опьянения.

Хочет предупредить читателя о возможности аномального результата в случае двух сравниваемых групп, когда оцениваемая вероятность будет меньше нуля или больше единицы. Рассмотрим гипотетические данные табл. 9.11, где одно частотное распределение есть зеркальное отражение другого.

Таблица 9.11

Гипотетические данные о тяжести травм в двух сравниваемых группах

Тяжесть	Группа А	Группа В
Нет	46	1
	34	2
Легкая	9	3
	5	5
	3	9
Тяжелая	2	34
Сильная	1	46
Итого	100	100

Используя ридит-значения для табл. 9.8, легко проверить, что два ридит-средних равны соответственно: $\overline{r}_A = 0{,}25$, $\overline{r}_B = 0{,}89$. Ход, изложенный выше, дает для вероятности, что объект, случайно выбранный из группы В, получит более тяжелую травму, чем случайно выбранный объект из группы А, невозможное значение $(0{,}89—0{,}25)+0{,}50=1{,}14$.

Когда — как в этом гипотетическом случае — частотные распределения в двух сравниваемых группах сильно различаются, следует выбрать одну из сравниваемых групп в качестве новой контрольной группы и пересчитать заново ридит-среднее, т. е. искомую вероятность. Задача 9.5 посвящена такому подходу к обработке табл. 9.11.

Задачи

9.1. Докажите равенство выражений (9.3) и (9.4).

9.2. Оценка коэффициента β дана в (9.17). Докажите равенство числителя и знаменателя (9.17) соответственно выражениям (9.19)—(9.20).

9.3. Три выборки пациентов психоневрологической больницы в Нью-Йорке изучались в совместном проекте [Cooper et al., 1972]. Число пациентов, которым был поставлен диагноз «эмоциональное расстройство», указано в таблице.

Выборка	Возрастной интервал	Число пациентов	Из них с эмоциональными расстройствами	Пропорция
1	20–34	105	2	—
2	20–59	192	13	—
3	35–59	145	24	—
Всего		442	39	

а) Вычислите пропорции пациентов с данным диагнозом и проверьте значимость их различия.

б) Проверьте значимость различия пропорции в объединении первых двух выборок и пропорции в третьей выборке. Проверьте различие пропорции в первых двух выборках.

в) Пациенты в выборке 1 в целом моложе пациентов в выборке 2, которые в свою очередь моложе пациентов в выборке 3. Поскольку вероятность эмоционального расстройства с возрастом увеличивается, можно выдвинуть гипотезу, что p_1 должно быть меньше p_2, а p_2 — меньше p_3. Расположены ли пропорции в этом порядке? Вычислите значения χ^2 в (9.32) и c в (9.34). Проверьте гипотезу по табл. 1.6[1].

---

[1] В задаче 9.3 имеется некоторое несоответствие между выдвигаемой гипотезой и планом проведенного исследования. Дело в том, что выборка сформирована по пациентам клиники, так что фактически проверяется предположение: «доля пациентов с эмоциональными расстройствами среди (госпитализируемых) пациентов с другими психическими патологиями увеличивается с возрастом». Правда, если принять за аксиому, что с возрастом растет доля психических заболеваний любого вида, то отвержение нулевой гипотезы «$p_1=p_2=p_3$» в пользу альтернативы «$p_1<p_2<p_3$» означает подтверждение предположения, выдвинутого в задаче. С такой точки зрения исследование по принятой схеме приводит просто к понижению мощности критериев. Кроме того, из-за сильного перекрывания возрастных интервалов выводы могут сильно зависеть от распределения пациентов внутри возрастных интервалов. Наконец, с течением времени условия жизни меняются, что вносит смещения при сравнении различных возрастных категорий, не учитываемые в этой задаче. Последнее обстоятельство, указанное научным редактором, игнорируется и в других задачах и примерах, где возраст является одним из факторов (см. гл. 5, 6, разд. 9.2).— *Примеч. пер.*

9.4. Пусть частоты, соответствующие данным табл. 9.7, есть $(n_{11}, \ldots, n_{k1})$ и $(n_{12}, \ldots, n_{k2})$, так что $p_{i1}=n_{i1}/n_1$ и p_{i2}/n_2, $i=1, \ldots, k$. Обозначим $n_{i.}=n_{i1}+n_{i2}$, $p_i=n_{i.}/n_.$. Классическое выражение для статистики хи-квадрат есть

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(n_{i1} - \dfrac{n_{i.} n_1}{n_.}\right)^2}{\dfrac{n_{i.} n_1}{n_.}} + \sum_{i=1}^{k} \frac{\left(n_{i2} - \dfrac{n_{i.} n_2}{n_.}\right)^2}{\dfrac{n_{i.} n_2}{n_.}}.$$

Покажите равенство этого выражения и (9.38).

9.5. По данным табл. 9.11:

а) Найдите ридит-среднее группы B, используя группу A в качестве контрольной.

б) Найдите ридит-среднее группы A, используя группу B в качестве контрольной. Какая связь между результатами а) и б)?

в) Найдите с помощью (9.40) стандартную ошибку ридит-среднего при сравнении групп A и B.

г) Оцените эту же стандартную ошибку с помощью (9.45). Как отличаются оценки в) и г)?

ЛИТЕРАТУРА

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, **11**, 375–385.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972). *Statistical inference under order restrictions.* New York: Wiley.

Bartholomew, D. J. (1959a). A test of homogeneity for ordered alternatives. *Biometrika*, **46**, 36–48.

Bartholomew, D. J. (1959b). A test of homogeneity for ordered alternatives II. *Biometrika*, **46**, 328–335.

Bartholomew, D. J. (1963). On Chassan's test for order. *Biometrics*, **19**, 188–191.

Bross, I. D. J. (1958). How to use ridit analysis. *Biometrics*, **14**, 18–38.

Bross, I. D. J. (1960). How to cut the highway toll in half in the next ten years. *Public Health Rep.*, **75**, 573–581.

Castellan, N. J. (1965). On the partitioning of contingency tables. *Psychol. Bull.*, **64**, 330–338.

Chapman, D. G. and Nam, J. (1968). Asymptotic power of chi-square tests for linear trends in proportions. *Biometrics*, **24**, 315–327.

Chassan, J. B. (1960). On a test for order. *Biometrics*, **16**, 119–121.

Chassan, J. B. (1962). An extension of a test for order. *Biometrics*, **18**, 245–247.

Cochran, W. G. (1954). Some methods of strengthening the common χ^2 tests. *Biometrics*, **10**, 417–451.

Cooper, J. E., Kendell, R. E., Gurland, B. J., Sharpe, L., Copeland, J. R. M., and Simon, R. (1972). *Psychiatric diagnosis in New York and London.* London: Oxford University Press.

Dorn, H. F. (1954). The relationship of cancer of the lung and the use of tobacco. *Am. Stat.*, 8, 7–13.

Irwin, J. O. (1949). A note on the subdivision of chi-square into components. *Biometrika*, **36**, 130–134.

Kantor, S., Winkelstein, W., and Ibrahim, M. A. (1968). A note on the interpretation of the ridit as a quantile rank. *Am. J. Epidemiol.*, 87, 609–615.

Kastenbaum, M. A. (1960). A note on the additive partitioning of chi-square in contingency tables. *Biometrics*, **16**, 416–422.

Kimball, A. W. (1954). Short-cut formulas for the exact partition of χ^2 in contingency tables. *Biometrics*, **10**, 452–458.

Knoke, J. D. (1976). Multiple comparisons with dichotomous data. *J. Am. Stat. Assoc.*, **71**, 849–853.

Lancaster, H. O. (1950). The exact partitioning of chi-square and its application to the problem of pooling small expectations. *Biometrika*, **37**, 267–270.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *J. Am. Stat. Assoc.*, 58, 690–700.

Mantel, N. (1979). Ridit analysis and related ranking procedures–Use at your own risk. *Am. J. Epidemiol.*, **109**, 25–29.

Miller, R. G. (1966). *Simultaneous statistical inference.* New York: McGraw-Hill.

National Center for Health Statistics (1970). Selected symptoms of psychological distress in the United States. *Data from National Health Survey*, Series 11, No. 37.

Selvin, S. (1977). A further note on the interpretation of ridit analysis. *Am. J. Epidemiol.*, **105**, 16–20.

Snell, E. J. (1964). A scaling procedure for ordered categorical data. *Biometrics*, **20**, 592–607.

Spitzer, R. L., Fleiss, J. L., Kernohan, W., Lee, J., and Baldwin, L. T. (1965). The Mental Status Schedule: Comparing Kentucky and New York schizophrenics. *Arch. Gen. Psychiatry*, **12**, 448–455.

Wood, C. L. (1978). Comparison of linear trends in binomial proportions. *Biometrics*, **34**, 496–504.

Wynder, E. L., Bross, I. D. J., and Hirayama, T. (1960). A study of the epidemiology of cancer of the breast. *Cancer*, **13**, 559–601.

Yates, F. (1948). The analysis of contingency tables with groupings based on quantitative characteristics. *Biometrika*, **35**, 176–181.

Глава 10

Совместный анализ нескольких четырехклеточных таблиц

Часто бывает так, что данные о связи между фактором риска A и заболеванием B представлены несколькими четырехклеточными таблицами. Например, если возможность связи между A и B весьма высока, а сама связь достаточно важна, то, скорее всего, она будет независимо изучаться многими исследователями. После того как была обнаружена связь в популяции одного вида, можно ожидать, что исследователи займутся ее поиском и изучением в популяциях других видов. Наконец, при исследовании отдельной популяции может понадобиться расслоение сравниваемых выборок по факторам, которые связаны с изучаемым выходным фактором. В результате для каждого слоя будет составлена своя четырехклеточная таблица.

Предположим, что в каждой из g групп изучалась связь между A и B и соответственно получено g четырехклеточных таблиц. Возникают следующие вопросы:

1. Можно ли утверждать, что степень связи, та или иная ее характеристика, одинаковы во всех группах?

2. Допустим, мы приняли решение, что степень связи одинакова. Значима связь, общая («типичная») для совокупности групп?

3. Допустим, мы приняли, что типичная связь значима. Какая оценка типичного значения меры связи является наилучшей? Какова ее стандартная ошибка? Как строить доверительный интервал для истинного значения меры?

В разд. 10.1 дана простая статистическая схема, по которой будут решаться эти вопросы. Раздел 10.2 описывает метод, рассчитанный на логарифм отношения шансов, разд. 10.3 описывает метод Корнфилда и Гарта, разд. 10.4 — метод Мантела — Ханзела. В разд. 10.5 эти методы сравниваются для различных планов исследования, в разд. 10.6 показано, как можно использовать эти методы в качестве альтернатив связыванию (гл. 8) для контроля мешающих факторов. В разд. 10.7 обсуждаются некоторые популярные, но в общем случае неверные методы совместного анализа совокупности четырехклеточных таблиц.

Методы, рассмотренные в этой главе, являются частными случаями методов анализа более сложных частотных таблиц. Общие методы логлинейного анализа и логистической регрессии прекрасно описаны в работах [Cox, 1970; Bishop et. al., 1976; Everitt, 1977; Fienberg, 1977] [1].

10.1. Построение и интерпретация некоторых критериев хи-квадрат

Чтобы решать поставленные вопросы, необходимо некоторое знакомство с теорией критериев хи-квадрат. Пусть y_i обозначает значение выбранной меры связи с i-й из g групп. Этой мерой может служить разность двух пропорций, логарифм отношения шансов и т. д.

Пусть s.e. (y_i) обозначает стандартную ошибку y_i. Величину, обратную ее квадрату, обозначим

$$w_i = \frac{1}{[\text{s.e. } (y_i)]^2}. \qquad (10.1)$$

w_i будет служить весом, соответствующим y_i. Если стандартная ошибка s.e. (y_i) велика, и, значит, точность определения y_i низка, то значение w_i мало. Если стандартная ошибка мала и, значит, точность оценки y_i высока, то значение w_i велико. Это разумно, поскольку более точной оценке следует приписывать больший вес.

Договоримся, что y_i — такая мера, что ее значение, равное нулю, указывает на отсутствие связи. Тогда при истинности гипотезы об отсутствии связи в i-й группе величина

---

[1] С задачами, рассмотренными в данной главе (и отчасти в других главах книги), близко связаны проблемы статистической теории люсианов, развиваемой в нашей стране. См., например, Орлов А. И. Устойчивость в социально-экономических моделях.— М.: Наука, 1979; Орлов А. И. Случайные множества с независимыми элементами (люсианы) и их применение//Алгоритмическое и программное обеспечение прикладного статистического анализа.— М.: Наука, 1980.— С. 287—308; Орлов А. И. Парные сравнения в асимптотике Колмогорова//Экспертные оценки в задачах управления.— М.: ВНИИСИ, 1982.— С. 58—66; Рыданова Г. В. Непараметрический анализ зависимости дихотомических векторов//Вестник МГУ. Сер. Выч. матем. и киберн. 1983.— № 4.— С. 48—52; Рыданова Г. В. Предельное распределение статистики критерия независимости дихотомических векторов в асимптотике растущего числа параметров//Вестник МГУ. Сер. Выч. матем. и киберн., 1987.— № 2.— С. 54—58; Орлов А. И., Рыданова Г. В. О некоторых результатах статистики объектов нечисловой природы//Программно-алгоритмическое обеспечение анализа данных в медико-биологических исследованиях.— Пущино, 1988.— С. 61—71; Рыданова Г. В. Проверка однородности и согласованности случайных бинарных векторов (люсианов) в асимптотике растущего числа параметров//Случайный анализ.— М.: МГУ, 1987.— С. 68—81.— *Примеч. пер.*

$$\chi_i = \frac{y_i}{\text{s. e. } (y_i)} = y_i \sqrt{w_i} \tag{10.2}$$

имеет приближенно стандартное нормальное распределение, а

$$\chi_i^2 = w_i y_i^2 \tag{10.3}$$

— приближенно распределение хи-квадрат с 1 степенью свободы. Пусть гипотеза об отсутствии связи в i-й группе неверна. Тогда следует ожидать, что χ_i^2 будет принимать большие значения, т. е. весьма вероятно, что критерий хи-квадрат отвергнет эту гипотезу.

Нас интересует, однако, не какая-то отдельная группа, а вся совокупность групп. Анализ совокупности удобно начинать с вычисления

$$\chi^2_{\text{total}} = \sum_{i=1}^{g} \chi_i^2 = \sum_{i=1}^{g} w_i y_i^2. \tag{10.4}$$

При отсутствии связи во всех g группах χ^2_{total} имеет распределение хи-квадрат с g степенями свободы, так как сумма g независимых величин с распределением хи-квадрат с 1 степенью свободы распределена по хи-квадрат с g степенями свободы, а g групп независимы по предположению.

Когда χ^2_{total} значимо велика, можно утверждать, что в одной или более групп связь есть, хотя нам неизвестно, постоянна ли она или меняется от группы к группе. χ^2_{total} не дает на этот счет информации. Однако, как сейчас будет показано, с ее помощью можно упростить другие вычисления.

Разобьем χ^2_{total} на две компоненты:

$$\chi^2_{\text{total}} = \chi^2_{\text{homog}} + \chi^2_{\text{assoc}}. \tag{10.5}$$

Величина χ^2_{homog} описывает степень однородности, или равенства, g мер связи, χ^2_{assoc} описывает значимость средней степени связи. Разделение (10.5) проще всего провести, вычисляя сначала χ^2_{assoc}, а потом простым вычитанием получая χ^2_{homog}.

Член χ^2_{assoc} вычисляется таким образом. Общая мера связи для совокупности групп определяется как взвешенное среднее отдельных мер с весами (10.1):

$$\bar{y} = \sum_{i=1}^{g} w_i y_i \Big/ \sum_{i=1}^{g} w_i. \tag{10.6}$$

При гипотезе о равенстве общей меры связи нулю среднее $\bar{y}$ есть нуль, а стандартная ошибка

$$\text{s.e.}\ (\bar{y}) = \frac{1}{\sqrt{\sum_{i=1}^{g} w_i}}\ . \tag{10.7}$$

Отсюда получаем при условии справедливости этой гипотезы, что

$$\chi_{\text{assoc}} = \frac{\bar{y}}{\text{s.e.}\ (\bar{y})} = \frac{\sum_{i=1}^{g} w_i\, y_i}{\sqrt{\sum_{i=1}^{g} w_i}} \tag{10.8}$$

распределена приближенно по стандартному нормальному закону и что

$$\chi^2_{\text{assoc}} = \bar{y}^2 \sum_{i=1}^{g} w_i = \frac{\left(\sum_{i=1}^{g} w_i\, y_i\right)^2}{\sum_{i=1}^{g} w_i} \tag{10.9}$$

распределена приближенно по хи-квадрат с 1 степенью свободы. χ^2_{homog} легко получить вычитанием;

$$\chi^2_{\text{homog}} = \chi^2_{\text{total}} - \chi^2_{\text{assoc}} = \sum_{i=1}^{g} w_i\, y_i^2 - \bar{y}^2 \sum_{i=1}^{g} w_i\ . \tag{10.10}$$

Эквивалентное выражение для χ^2_{homog} :

$$\chi^2_{\text{homog}} = \sum_{i=1}^{g} w_i\ (y_i - \bar{y})^2, \tag{10.11}$$

полезно в двух отношениях. Во-первых, с его помощью можно провести арифметическую проверку. Во-вторых, оно показывает, что χ^2_{homog} действительно измеряет степень разнородности y_i по группам. При гипотезе об однородности связи χ^2_{homog} распределена приближенно по хи-квадрат с $g-1$ степенями свободы.

Итак, теперь мы можем решать вопросы, поставленные в начале главы.

1. Однородность связи можно проверять с помощью χ^2_{homog} по таблице распределения хи-квадрат с $g-1$ степенями свободы. Если отклонение статистики значимо, следует разбить χ^2_{homog} на соответствующие компоненты, чтобы выделить те

группы, связь в которых отличается от связи в остальных группах (см. задачу 10.1).

2. Если χ^2_{homog} не значима, можно проверить значимость общей меры связи с помощью χ^2_{assoc} и таблицы распределения хи-квадрат с 1 степенью свободы.

3. Наилучшая оценка общей меры — $\bar{y}$ (см. (10.6)). Ее стандартная ошибка дана в (10.7). Приближенный 100 $(1-\alpha)$ %-ный доверительный интервал для общей меры есть

$$\bar{y} \pm c_{\alpha/2} \text{ s.e. } (\bar{y}), \tag{10.12}$$

где $c_{\alpha/2}$ — значение, отделяющее на верхнем хвосте функции стандартного нормального распределения долю $\alpha/2$.

Обычно надеются обнаружить, что значение χ^2_{homog} мало (т. е. можно сделать вывод об однородности связи), а значение χ^2_{assoc} велико (т. е. можно сделать вывод о существовании связи для совокупности групп).

Вопрос о том, стоит ли вообще проводить проверку гипотезы об однородности связи, стал предметом споров. Например, Бишоп и др. [Bishop et al., 1975, p. 147] утверждают, что перед построением выводов о предположительно одинаковой в группах (типичной) мере связи нужно обязательно проверять однородность связи. С другой стороны, Мантел и др. [Mantel et al., 1977] предостерегают, говоря, что при интерпретации результатов таких проверок надо иметь в виду, что сильное влияние на вывод о наличии или отсутствии однородности оказывает выбор конкретной меры связи.

На практике перед началом более сложного анализа разумно, во-первых, проверить по данным, что g мер указывают по крайней мере на связь одного знака (одного направления), если не на точное совпадение значений. Во-вторых, стоит проверить по χ^2_{homog} при консервативном уровне значимости (например, 0,01), что меры связи различаются не очень сильно и что последующие выводы о предположительно-типичной мере связи применимы к отдельным группам.

Теперь остается только применить эту технику к конкретным мерам связи. Далее в этой главе мы будем пользоваться следующими обозначениями: n_{i1} — число объектов в первой выборке i-й группы; p_{i1} — пропорция объектов, обладающих изучаемым фактором; n_{i2} и p_{i2} — соответствующие значения во второй выборке i-й группы. Суммарное число объектов в i-й группе — $n_{i.} = n_{i1} + n_{i2}$, общая пропорция объектов с присутствием фактора

$$\bar{p}_i = \frac{n_{i1}\, p_{i1} + n_{i2}\, p_{i2}}{n_i} . \tag{10.13}$$

Дополняющая пропорция $\bar{q}_i = 1 - \bar{p}_i$.

10.2. Совместный анализ логарифма отношения шансов

Отношение шансов само по себе

$$o_i = \frac{p_{i1}\,(1-p_{i2})}{p_{i2}\,(1-p_{i1})}, \qquad (10.14)$$

таково, что равенство нулю не следует из отсутствия связи, тогда как этим свойством обладает логарифм отношения шансов. Поэтому в качестве меры связи мы рассмотрим

$$y_i = L_i = \ln\,(o_i). \qquad (10.15)$$

Квадрат стандартной ошибки L_i равен приближенно:

$$[\text{s.e.}\,(L_i)]^2 = \frac{1}{w_i} = \frac{1}{n_{i1}\,p_{i1}\,(1-p_{i1})} + \frac{1}{n_{i2}\,p_{i2}\,(1-p_{i2})}, \qquad (10.16)$$

что равно сумме обратных исходных частот (см. (6.6) и 6.8)). Вес w_i — величина, обратная сумме обратных частот.

Логарифм отношения шансов был проанализирован по схеме разд. 10.1 в [Gart, 1962; Sheehe, 1966]. Мы проиллюстрируем результаты этих работ с помощью данных табл. 10.1, в которой представлены пропорции пациентов в Нью-Йорке и в Лондоне, которым психиатры, живущие при клиниках, поставили диагноз шизофрения [Cooper et al., 1972].

Таблица 10.1

Данные о числе диагнозов шизофрении, поставленных психиатрами клиник в Нью-Йорке и в Лондоне в ходе трех исследований

Исследование	Нью-Йорк		Лондон	
	n_{i1}	p_{i1}	n_{i2}	p_{i2}
$i=1$ (возраст 20—34 лет)	105	0,771	105	0,324
$i=2$ (возраст 20—59 лет)	192	0,615	174	0,397
$i=3$ (возраст 35—59 лет)	145	0,566	145	0,359

В табл. 10.2 даны результаты вычислений для анализа логарифма отношения шансов. Чтобы уменьшить смещение[Naylor, 1967, Gart 1970, 1971], вычисляются $L_i' = \ln(o_i')$ и $w_i' = 1/[\text{s.e.}\,(L_i')]$, в которых к каждой частоте аналогично (5.20) и (5.33) добавляется постоянная 0,5. Все значения отдельных статистик $w_i\,(L_i')^2$ приближенно равны значениям стандартных статистик

хи-квадрат с 1 степенью свободы, вычисленных с поправками, и все высоко значимы.

Таблица 10.2

Анализ данных табл. 10.1 с помощью логарифмов отношения шансов

Исследования	o_i'	L_i'	w_i'	$w_i' L_i'$	$w_i' (L_i')^2$
1	6,894	1,931	10,410	20,102	38,816
2	2,415	0,881	21,868	19,266	16,973
3	2,314	0,839	17,357	14,563	12,218
Сумма			49,635	53,931	68,007

Суммарная статистика хи-квадрат равна:

$$\chi^2_{\text{total}} = \sum_{i=1}^{3} w_i' \ (L_i')^2 = 68{,}01, \qquad (10.17)$$

а статистика хи-квадрат для проверки однородности отношения шансов равна:

$$\chi^2_{\text{homog}} = \sum_{i=1}^{3} w_i' (L_i')^2 - \frac{\left(\sum_{i=1}^{3} w_i' L_i' \right)^2}{\sum_{i=1}^{3} w_i'} =$$

$$= 68{,}01 - \frac{(53{,}931)^2}{49{,}635} = 9{,}41, \qquad (10.18)$$

что при двух степенях свободы указывает на различие трех значений отношения шансов, значимое на уровне 0,01. Подробному анализу разнородности отношения шансов в табл. 10.2 посвящена задача 10.1.

Хотя три значения логарифма отношения шансов не равны, они по крайней мере имеют одинаковый знак. Дальнейший анализ данных в предположении одинакового (типичного) отношения шансов может быть оправдан, если выводы относятся не к диагнозам в отдельных возрастных группах, а ко всей совокупности психиатрических пациентов Нью-Йорка и Лондона в целом. В задаче 10.2 требуется провести анализ по описываемой ниже схеме для данных только из групп 2 и 3, где пациенты старше по возрасту, и отношения шансов сходны.

Оценкой логарифма отношения шансов в предположении его равенства в группах является

$$\bar{L}' = \frac{\sum_{i=1}^{3} w_i' L_i'}{\sum_{i=1}^{3} w_i'} = \frac{53,931}{49,635} = 1,087 \qquad (10.19)$$

со стандартной ошибкой, оцениваемой величиной

$$\text{s.e.}(\bar{L}') = \frac{1}{\sqrt{\sum_{i=1}^{3} w_i'}} = \frac{1}{\sqrt{49,635}} = 0,142. \qquad (10.20)$$

Значение статистики для проверки значимости среднего логарифма отношения шансов равно:

$$\chi^2_{\text{assoc}} = \left(\frac{\bar{L}'}{\text{s.e.}(\bar{L}')}\right)^2 = \left(\frac{1,087}{0,142}\right)^2 = 58,60, \qquad (10.21)$$

что при одной степени свободы дает, очевидно, высокую значимость. Следовательно, можно сделать вывод, что психиатры-клинисты больному с психическими расстройствами, госпитализированному в Нью-Йорке, поставят диагноз шизофрения с большей вероятностью, чем больному, госпитализированному в Лондоне.

Приближенный 95%-ный доверительный интервал для λ, логарифма предполагаемого типичного отношения шансов, есть

$$\bar{L}' - 1,96 \cdot \text{s.e.}(\bar{L}') \leqslant \lambda \leqslant \bar{L}' + 1,96\ \text{s.e.}(\bar{L}'),$$

$$1,087 - 1,96 \cdot 0,142 \leqslant \lambda \leqslant 1,087 + 1,96 \cdot 0,142$$

и окончательно

$$0,809 \leqslant \lambda \leqslant 1,365. \qquad (10.22)$$

Как правило, желательно представить окончательные результаты в терминах отношения шансов, а не его логарифма. Среднее отношение шансов оценивается величиной

$$\bar{o}' = e^{\bar{L}'} = \exp(\bar{L}'). \qquad (10.23)$$

Приближенный 95%-ный интервал для ω, предположительно типичного отношения шансов, есть

$$\exp(\bar{L}' - 1,96 \cdot \text{s.e.}(\bar{L}')) \omega \leqslant \exp(\bar{L}' + 1,96\ \text{s.e.}(\bar{L}')). \qquad (10.24)$$

Для данных нашего примера имеем:

$$\bar{o}' = \exp(1,087) = 2,97, \qquad (10.25)$$

а приближенный 95%-ный доверительный интервал:

$$\exp(0{,}809) \leqslant \omega \leqslant \exp(1{,}365),$$

или

$$2{,}25 < \omega < 3{,}92. \qquad (10.26$$

10.3. Метод Корнфилда и Гарта

В этом разделе будет применяться теория, описанная в разд. 5.5. Методы разд. 10.1 здесь не используются. Представим данные для i-й группы в виде табл. 10.3.

Таблица 10.3

Обозначения для данных i-й группы

Выборка	Результирующий признак		Сумма
	Присутствует	Отсутствует	
1	X_i	$n_{i1}-X_i$	n_{i1}
2	m_i-X_i	$n_{i2}-m_i+X_i$	n_{i2}
Сумма	m_i	$n_{i.}-m_i$	$n_{i.}$

Как доказано в [Cornfield, 1956], если все четыре маргинальные частоты фиксированы, а ω — истинное значение отношения шансов, то X_i распределена приближенно нормально со средним x_i и стандартной ошибкой

$$\text{s.e.}(X_i) = \frac{1}{\sqrt{W_i(x_i)}}, \qquad (10.27)$$

где

$$W_i(x_i) = \frac{1}{x_i} + \frac{1}{n_{i1}-x_i} + \frac{1}{m_i-x_i} + \frac{1}{n_{i2}-m_i+x_i}, \qquad (10.28)$$

а x_i — единственный корень квадратного уравнения

$$\frac{x_i(n_{i2}-m_i+x_i)}{(n_{i1}-x_i)(m_i-x_i)} = \omega, \qquad (10.29)$$

лежащий в интервале:

$$\text{большее из } (0;\ m_i-n_{i2}) \leqslant x_i \leqslant \text{меньшее из } (n_{i1},\ m_i). \qquad (10.30)$$

В явном виде квадратное уравнение (10.29) есть

$$x_i^2(\omega-1) - x_i[\omega(n_{i1}+m_i)+n_{i2}-m_i] + \omega n_{i1}m_i = 0. \qquad (10.31)$$

Гарт [Gart, 1970] обобщил этот результат на случай нескольких четырехклеточных таблиц. Проверка данных на разнородность отношения шансов в группах начинается с оценивания отношения шансов в предположении его однородности. Соответствующую оценку $\hat{\omega}$ нельзя получить в явном виде. Она вычисляется по системе из $(g+1)$ уравнений:

$$\frac{\hat{x}_i\,(n_{i2}-m_i+\hat{x}_i)}{(n_{i1}-\hat{x}_i)\,(m_i-\hat{x}_i)}=\hat{\omega}, \tag{10.32}$$

где $\hat{x}_i$ лежит в интервале (10.30), $i=1, \ldots, g$, и

$$\sum_{i=1}^{g} X_i=\sum_{i=1}^{g}\hat{x}_i. \tag{10.33}$$

Найти оценку $\hat{\omega}$ можно как методом проб и ошибок, так и любым обычным итеративным методом решения сложных уравнений. Начальным приближением к $\hat{\omega}$ может служить $\bar{o}'$ из (10.23) или оценка Мантела—Ханзела (см. разд. 10.4).

Оценка $\hat{\omega}$ по данным табл. 10.1 равна:

$$\hat{\omega}=3{,}04, \tag{10.34}$$

что немного больше оценки (10.25), основанной на логарифме отношения шансов. В табл. 10.4 приведены соответствующие значения $\hat{x}$ и $W_i(\hat{x}_i)$. Заметьте, что $\Sigma X_i=\Sigma\hat{x}_i=281$.

Таблица 10.4

Значения для данных табл. 10.1, соответствующие $\hat{\omega}=3{,}04$

Исследование	n_{i1}	n_{i2}	m_i	X_i	$\hat{x}_i$	$W_i(\hat{x}_i)$
1	105	105	115	81	71,601	0,0832
2	192	174	187	118	122,856	0,0473
3	145	145	134	82	86,543	0,0600
Сумма				281	281,000	

Гипотезу о равенстве истинных отношений шансов можно проверить по таблице распределения хи-квадрат с $g-1$ степенью свободы с помощью статистики

$$\chi^2_{\text{homog}}=\sum_{i=1}^{g} W_i(\hat{x}_i)\,(X_i-\hat{x}_i)^2. \tag{10.35}$$

Для данных табл. 10.4

$$\chi^2_{\text{homog}} = 9{,}70, \tag{10.36}$$

что при двух степенях свободы означает статистически значимое различие ($p<0{,}01$) отношений шансов в трех группах табл. 10.1. Значение хи-квадрат (10.36) оказывается немного большим соответствующего значения (10.18), основанного на логарифме отношения шансов.

Опишем теперь критерий проверки значимости общего отношения шансов (т. е. его отличия от 1). Если истинное общее значение $\omega = 1$, то (10.31) становится линейным уравнением с единственным корнем

$$x_i = \frac{n_{i1}\, m_i}{n_{i.}}. \tag{10.37}$$

Соответствующее значение

$$W_i\ (x_i) = \frac{n_i^3}{n_{i1}\, n_{i2}\, m_i\, (n_{i.} - m_i)}. \tag{10.38}$$

При гипотезе $\omega = 1$ величина

$$\chi^2_{\text{assoc}} = \frac{\left(\left| \sum_{i=1}^{g} X_i - \sum_{i=1}^{g} x_i \right| - 0{,}5 \right)^2}{\sum_{i=1}^{g} \frac{1}{W_i\ (x_i)}} \tag{10.39}$$

распределена приближенно по хи-квадрат с 1 степенью свободы. Как указано в разд. 10.4, эта величина тесно связана со статистикой хи-квадрат Мантела—Ханзела.

Значения x_i и $W_i\ (x_i)$ при гипотезе, что $\omega = 1$, приведены в табл. 10.5.

Таблица 10.5

Значение для данных табл. 10.1 при истинности гипотезы «$\omega = 1$»

Исследование	n_{i1}	n_{i2}	m_i	X_i	x_i	$W_i\ (x_i)$	$1/W_i\ (x_i)$
1	105	105	115	81	57,500	0,0769	13,006
2	192	174	187	118	98,098	0,0438	22,809
3	145	145	134	82	67,000	0,0555	18,021
Сумма				281	222,598		53,836

Значение статистики (10.39) равно:

$$\chi^2_{\text{assoc}} = \frac{(|281 - 222{,}6| - 0{,}5)^2}{53{,}836} = 62{,}28. \qquad (10.40)$$

Значит, можно сделать вывод, что истинное типичное значение отношения шансов отличается от единицы. Величина (10.40) оказывается несколько больше соответствующего значения (10.27).

Приближенный $100(1-\alpha)\%$-ный доверительный интервал для истинного типичного отношения шансов (одинакового во всех группах по предположению) определяется следующим образом. Нижнюю доверительную границу ω_L можно найти из уравнений (10.41) — (10.43):

$$\frac{x_{iL}(n_{i2} - m_i + x_{iL})}{(n_{i1} - x_{iL})(m_i - x_{iL})} = \omega_L, \qquad (10.41)$$

где x_{iL} — корень в интервале (10.30),

$$W_i(x_{iL}) = \frac{1}{x_{iL}} + \frac{1}{n_{i1} - x_{iL}} + \frac{1}{m_i - x_{iL}} + \frac{1}{n_{i2} - m_i + x_{iL}} \qquad (10.42)$$

и

$$\frac{\left(\left(\sum_{i=1}^{g} X_i - \sum_{i=1}^{g} x_{iL}\right) - 0{,}5\right)^2}{\sum_{i=1}^{g} \frac{1}{W_i(x_{iL})}} = c^2_{\alpha/2}. \qquad (10.43)$$

Верхнюю границу ω_U находят аналогично, только поправка на непрерывность в (10.43) со значением $-0{,}5$ заменяется на $+0{,}5$

Как и при вычислении оценки типичного отношения шансов в (10.32) и (10.33), поиск верхней и нижней границ можно проводить методом проб и ошибок или с помощью итеративной вычислительной процедуры. В качестве начального приближения можно использовать границы (10.24), основанные на логарифме отношения шансов.

Значения величин в табл. 10.6 соответствуют нижней 95%-ной доверительной границе $\omega_L = 2{,}28$. Заметим, что

$$\frac{[(281 - 266{,}421) - 0{,}5]^2}{51{,}606} = 3{,}84, \qquad (10.44)$$

как и должно быть при уровне доверия 95%.

Таблица 10.6

Значения для данных табл. 10.1, соответствующие нижней 95%-ной доверительной границе $\omega_L = 2{,}28$

Исследование	n_{i1}	n_{i2}	m_i	X_i	x_{iL}	$W_i(x_{iL})$	$1/W_i(x_{iL})$
1	105	105	115	81	68,083	0,0803	12,453
2	192	174	187	118	116,672	0,0457	21,882
3	145	145	134	82	81,666	0,0579	17,271
Сумма				281	266,421		51,606

Значения в табл. 10.7 соответствуют верхней границе $\omega_U = 4{,}06$. Как и требуется,

$$\frac{(281 - 295{,}033 + 0{,}5)^2}{17{,}686} = 3{,}84. \qquad (10.45)$$

Таблица 10.7

Значения, соответствующие верхней 95%-ной доверительной границе $\omega_U = 4{,}06$ для данных табл. 10.1

Исследование	n_{i1}	n_{i2}	m_i	X_i	x_{iU}	$W_i(x_{iU})$	$1/W_i(x_{iU})$
1	105	105	115	81	74,985	0,0870	11,494
2	192	174	187	118	128,811	0,0494	20,243
3	145	145	134	82	91,237	0,0627	15,949
Сумма				281	295,033		47,686

Итак, приближенный 95%-ный доверительный интервал для предположительно типичного отношения шансов, построенный с помощью результатов Корнфилда, есть

$$2{,}28 \leqslant \omega \leqslant 4{,}06. \qquad (10.46)$$

Этот интервал немного шире и слегка смещен вправо относительно интервала (10.26).

10.4. Метод Мантела — Ханзела

Процедура, предложенная Мантелом и Ханзелом [Mante and Haenszel, 1959] и развитая Мантелом [Mantel, 1963], позволяет оценивать предположительно типичное значение отношения шансов и проверять значимость общей степени связи. Ис-

тересно, что критерий значимости связи основан непосредственно не на отношении шансов, а на другой мере связи. Радхакришна [Radhakrishna, 1965] показал корректность этого подхода.

Оценкой Мантела — Ханзела отношения шансов является

$$\bar{o}_{\text{МН}} = \frac{\sum_{i=1}^{g} \frac{n_{i1}\, n_{i2}}{n_{i.}}\, p_{i1}\,(1-p_{i2})}{\sum_{i=1}^{g} \frac{n_{i1}\, n_{i2}}{n_{i.}}\, p_{i2}\,(1-p_{i1})}. \qquad (10.47)$$

$\bar{o}_{\text{МН}}$ — взвешенное среднее отдельных отношений шансов по группам (см. задачу 10.3). Для данных табл. 10.1

$$\bar{o}_{\text{МН}} = \frac{87{,}516}{29{,}143} = 3{,}00, \qquad (10.48)$$

что немного больше оценки (10.25), полученной с помощью логарифма отношения шансов, и немного меньше оценки (10.34), полученной по методу Корнфилда — Гарта.

Критерий хи-квадрат Мантела — Ханзела проверки значимости общей меры связи основан на взвешенном среднем g разностей между пропорциями

$$\bar{d} = \sum_{i=1}^{g} \frac{n_{i1}\, n_{i2}}{n_{i.}}\,(p_{i1}-p_{i2}) \Big/ \sum_{i=1}^{g} \frac{n_{i1}\, n_{i2}}{n_{i}}. \qquad (10.49)$$

Статистика хи-квадрат Мантела — Ханзела задается выражением

$$\chi^2_{\text{МН}} = \frac{\left(\left| \sum_{i=1}^{g} \frac{n_{i1}\, n_{i2}}{n_{i.}}\,(p_{i1}-p_{i2}) \right| - 0{,}5 \right)^2}{\sum_{i=1}^{g} \frac{n_{i1}\, n_{i2}}{n_{i.}-1}\, \bar{p}_i\, \bar{q}_i} \qquad (10.50)$$

с 1 степенью свободы. Для данных табл. 10.1

$$\chi^2_{\text{МН}} = \frac{(58{,}374-0{,}5)^2}{54{,}024} = 62{,}00, \qquad (10.51)$$

что немного меньше значения (10.40) статистики хи-квадрат Корнфилда — Гарта. Выражения (10.50) и (10.39) будут совпадать, если в знаменателе (10.50) заменить $n_{i.}-1$ на $n_{i.}$.

Со статистикой Мантела — Ханзела тесно связаны статистика Кохрэна [Cochran, 1954]

$$\chi^2_C = \frac{\left(\sum_{i=1}^{g} \frac{n_{i1}\, n_{i2}}{n_{i.}}\,(p_{i1}-p_{i2}) \right)^2}{\sum_{i=1}^{g} \frac{n_{i1}\, n_{i2}}{n_{i.}}\, \bar{p}_i\, \bar{q}_i}. \qquad (10.52)$$

Выражение (10.50) отличается от выражения (10.52) не только тем, что содержит поправку на непрерывность, но и тем, что в знаменателе берется $n_{i.}-1$, а не $n_{i.}$. Последнее отличие более важно. Разница несущественна, если объемы выборок во всех группах велики, но значительна в противном случае.

Например, рассмотрим крайний случай, когда, как при исследовании по связанным парам, каждая выборка состоит из двух объектов — по одному из каждой выборки. Легко проверить, что статистики хи-квадрат Мак-Немара (8.3) и Мантела — Ханзела (10.50) совпадают. С другой стороны, статистика Кохрэна (10.52), если ввести в нее поправки на непрерывность, будет вдвое больше статистики Мак-Немара.

Мантел [Mantel, 1966] показал, как можно использовать критерий Мантела — Ханзела при сравнении независимых таблиц выживаемости. В [Mantel, 1977] описано, как можно модифицировать статистику Мантела — Ханзела, чтобы получить приближенный доверительный интервал для предположительно типичного отношения шансов. Эти процедуры слишком сложны, чтобы описывать их здесь. За подробностями мы отсылаем читателя к [Mantel, 1966, 1977, Mantel and Hankey, 1975].

По классическому правилу для того, чтобы с достаточной точностью считать, что статистика хи-квадрат для четырехклеточной таблицы имеет распределение хи-квадрат с 1 степенью свободы, объем выборок должен быть таким, чтобы все ожидаемые частоты имели значение не меньше пяти (см. разд. 2.2). Аналогичный критерий для статистики Мантела — Ханзела предложен Мантелом и Флейсом [Mantel and Fleiss, 1980]. По этому критерию каждая из четырех сумм ожидаемых частот,

$$\sum_{i=1}^{g} n_{i1}\bar{p}_i,\quad \sum_{i=1}^{g} n_{i2}\bar{p}_i,\quad \sum_{i=1}^{g} n_{i1}\bar{q}_i,\quad \sum_{i=1}^{g} n_{i2}\bar{q}_i,$$

должна отличаться не менее чем на 5 как от своего минимума, так и от своего максимума.

Значит, чтобы с уверенностью пользоваться для статистики (10.50) распределением хи-квадрат с 1 степенью свободы, вовсе не обязательно иметь большие маргинальные частоты. Число наблюдений в таблице может быть даже равно двум, как в случае связанных пар. Единственное, что нужно при этом — достаточно большое число таблиц, чтобы каждая сумма ожидаемых частот была велика.

10.5. Сравнение трех методов

Гарт [Gart, 1962, 1970], Одоров [Odoroff, 1970] и Мак-Кинлей [McKinlay, 1975в, 1978] сравнили три метода, описанные выше, а также методы Бёрча [Birch, 1964] и Гудмена [Goodman, 1969]. Следует различать два случая.

В первом случае число групп или слоев невелико, а число наблюдений в каждой группе значительно. Эта ситуация имеет место, когда сравниваемые выборки разбивают на ограниченное число слоев, а новые объекты относят к одному из существующих слоев, или когда анализируется несколько повторных выборок. В этом случае методы разд. 10.2, основанные на логарифме отношения шансов, лучше или лишь немного хуже других. Принимая во внимание их хорошую точность и относительную простоту, мы рекомендуем, если число слоев невелико, единообразно для каждого слоя проводить с помощью этих методов проверку основных гипотез об отношении шансов.

Во втором случае каждая группа или слой имеет небольшой размер, но их число велико. Так бывает, когда проводится расслоение сравниваемых выборок (обычно по окончании сбора данных) по большому числу факторов или когда применяется связывание и число объектов в связках, возможно, неодинаково (например, некоторые связки образованы парами объектов, другие — одним объектом из первой выборки и двумя — из второй и т. д.). или когда включение в исследование новых объектов означает создание новых групп или слоев.

В этом случае следует предпочесть оценку Мантела — Ханзела общего отношения шансов (10.47), критерий проверки его значимости по статистике Мантела — Ханзела с 1 степенью свободы (10.50) и доверительный интервал для общего отношения шансов по методу Корнфилда — Гарта (см. (10.41) — (10.43)). Проверка равенства отношений шансов во втором случае менее важна. В отличие от первого случая (малое число групп большого размера) методы, основанные на логарифме отношения шансов, при большом числе малых групп работают крайне плохо.

10.6. Альтернативы связыванию

Мак-Кинлей [McKinlay, 1975a] сделала ретроспективный обзор методов контроля факторов, вносящих смещение в нерандомизированных исследованиях (таких, как сравнительное проспективное или ретроспективное исследования), и также описала результаты статистических исследований по этим методам. В подготовленной ею библиографии — 165 работ, другие работы названы дополнительно Файнбергом в его комментариях к обзору Мак-Кинлей. Всего существуют три сравнительно простых метода контроля смещающих факторов: связывание (соответствующие методы анализа описаны в гл. 8), расслоение (соответствующие методы описываются в настоящей главе) и ковариационный, или регрессионный контроль (в данной книге не обсуждается, см. [Zubin, 1973]).

Предположим, например, что мы проводим ретроспективное исследование связи между курением и раком легких, учитывая возможность мешающего влияния таких факторов, как пол и возраст. Одним способом контроля этих мешающих факторов является связывание с каждым больным раком легких одного или нескольких контрольных индивидуумов такого же возраста и пола и применение методов разд. 8.1 или 8.3.

Другой способ состоит в том, чтобы перекрестным методом выбора извлечь выборку опытных объектов и выборку контрольных объектов, расслоить обе выборки по полу и возрасту, а затем сформировать раздельно по каждому слою четырехклеточную таблицу для сопоставления долей курящих среди опытных и среди контрольных объектов. Скажем, если имеется пять возрастных интервалов, то всего получится $g=10$ таблиц: пять для женщин и пять для мужчин. Можно считать, что полученное множество таблиц взято из g отдельных групп, и применять методы разд. 10.2—10.4.

Если из многих смещающих факторов учитываются два-три, то возможное влияние неконтролируемых (может быть, даже неизмеряемых) факторов можно оценить с помощью критериев, предложенных Броссом [Bross, 1966] и Шлесселманом [Schlesselman, 1978]. Если желателен одновременный контроль многих смещающих факторов (более трех), то в качестве основы для расслоения можно использовать «метку многомерного мешающего фактора» (multivariate confounder score) Миеттинена [Miettinen, 1976]. По мнению Миеттинена, достаточно иметь пять слоев. Однако при этом сначала надо применять многомерную процедуру типа дискриминантного анализа, чтобы определить, как вычислять составную метку.

Преимущество связывания — гарантия сходства двух выборок по факторам связывания; главный недостаток — практические затруднения при подборе контрольного объекта для каждого опытного объекта, если число последних велико. Другие недостатки связывания указаны в разд. 8.5.

Расслоение выборок после их извлечения хорошо тем, что заранее определять состав выборок не нужно. Другое их достоинство — возможность проверять постоянство связи в различных слоях. Недостатком расслоения является то, что число объектов одной из выборок в слое может оказаться малым по сравнению с числом объектов другой выборки, если размеры выборок не очень велики. От этого могут страдать мощность и точность результатов сравнения.

Кохрэн [Cochran, 1968] и Рубин [Rubin, 1973] исследовали эффективность связывания по сравнению с расслоением при контроле мешающих факторов в случае количественных измерений, Мак-Кинлей [McKinlay, 1975c] — в случае дихотомичес-

ких измерений. Опираясь на их результаты, можно рекомендовать связывание только для выборок умеренного размера, а перекрестное извлечение с последующим расслоением выборок — для выборок большого размера [1].

10.7. Методы, которые не следует применять

Критерий, описанный Флейсом

В первом издании этой книги описан критерий однородности, впервые предложенный Иэйтсом [Yates, 1959]. В этом критерии из суммы обычных статистик хи-квадрат с 1 степенью свободы для отдельных групп (без поправок на непрерывность) вычитается значение статистики хи-квадрат с 1 степенью свободы Кохрэна (10.52). Подход, описанный в разд. 10.1, применим здесь в следующей форме.

В качестве меры связи в i-й группе возьмем так называемую *стандартизованную разность* (standartized difference)

$$y_i = d_i = \frac{p_{i1} - p_{i2}}{\bar{p}_i \bar{q}_i} . \qquad (10.53)$$

Квадрат ее стандартной ошибки

$$[\text{s.e. } (d_i)]^2 = \frac{1}{\bar{p}_i \bar{q}_i} \left(\frac{n_{i.}}{n_{i1} n_{i2}} \right) , \qquad (10.54)$$

поэтому

$$w_i = \frac{\bar{p}_i \bar{q}_i n_{i1} n_{i2}}{n_{i.}} \qquad (10.55)$$

и

$$\chi_i^2 = w_i d_i^2 = \frac{(p_{i1} - p_{i2})^2}{\bar{p}_i \bar{q}_i (1/n_{i1} + 1/n_{i2})} , \qquad (10.56)$$

что совпадает с обычной статистикой хи-квадрат без поправки на непрерывность.

---

[1] Как мы видим, основное различие между связыванием и расслоением лежит в области планирования исследований: связывание относится к методам выбора II и III, расслоение — к методу выбора I. К сожалению, автор, противопоставляя эти способы контроля мешающих факторов, обращает мало внимания на их близость. Методы анализа расслоенных и связанных данных тесно связаны (см., например, примечания к разд. 8.1 и 8.3). Далее, можно легко представить формы контроля мешающих факторов, промежуточные по отношению к расслоению и связыванию. В некоторой степени связывание можно считать частным случаем расслоения, когда каждая связка является отдельным слоем.— *Примеч. пер.*

Средняя стандартизованная разность есть

$$\bar{d} = \frac{\sum_{i=1}^{g} \frac{(p_{i1}-p_{i2})\, n_{i1}\, n_{i2}}{n_{i.}}}{\sum_{i=1}^{g} \frac{\bar{p}_i\, \bar{q}_i\, n_{i1}\, n_{i2}}{n_{i.}}} \qquad (10.57)$$

Квадрат ее стандартной ошибки

$$[\text{s.e. } (\bar{d})]^2 = \frac{1}{\sum_{i=1}^{g} \frac{\bar{p}_i\, \bar{q}_i\, n_{i1}\, n_{i2}}{n_{i.}}}, \qquad (10.58)$$

поэтому статистикой хи-квадрат критерия значимости общей связи является

$$\chi^2_{\text{assoc}} = \frac{\bar{d}^2}{[\text{s.e. } (d)]^2} = \frac{\left(\sum_{i=1}^{g} \frac{(p_{i1}-p_{i2})\, n_{i1}\, n_{i2}}{n_{i.}} \right)}{\sum_{i=1}^{g} \frac{\bar{p}_i \bar{q}_i\, n_{i1}\, n_{i2}}{n_{i.}}}, \qquad (10.59)$$

что идентично статистике χ^2_C Кохрэна (10.52).

Ошибкой в первом издании было то, что величина

$$\chi^2_{\text{homog}} = \sum_{i=1}^{g} w_i\, d_i^2 - \chi^2_C \qquad (10.60)$$

с $g-1$ степенью свободы предлагалась в качестве статистики критерия однородности, корректного во всех случаях. Мантел и др. [Mantel et al., 1977] показали, что стандартизованная разность (10.53) и, следовательно, статистика (10.60) чувствительна к отношению объемов выборок, так же как и к истинному значению меры связи, и поэтому критерий, основанный на (10.60), может иногда быть неверным.

Таблица 10.8

Данные для проверки критерия однородности (10.60)

	Выборка 1		Выборка 2		Объединенная выборка	
Группа	n_{i1}	p_{i1}	n_{i2}	p_{i2}	$n_{i.}$	$\bar{p}_i$
$i=1$	230	0,87	50	0,20	280	0,75
$i=2$	40	0,25	810	0,0123	850	0,0235

Рассмотрим данные табл. 10.8, взятые из [Mantel et al., 1977]. Легко проверить, что отношение шансов в обеих группах равно 26,7. Тем не менее, как показано в табл. 10.9,

Таблица 10.9

Результаты вычислений по данным табл. 10.8

Группа	d_i	w_i	$w_i d_i$	$w_i d_i^2$
$i=1$	3,57	7,70	27,489	98,136
$i=2$	10,36	0,87	9,013	93,375
Сумма		8,57	36,502	191,511

две стандартизованные разности заметно различаются, и по критерию (10.60) мы получаем вывод о различии связи в двух группах:

$$\chi^2_{\text{homog}} = 191{,}51 - \frac{36{,}502^2}{8{,}57} = 36{,}04, \qquad (10.61)$$

что высоко значимо при $g-1=1$ степени свободы.

В [Mantel et al., 1977] продемонстрированы другие возможные аномалии, связанные со статистикой (10.60) (например, отношения шансов в g группах сильно различаются, а χ^2_{homog} тем не менее может равняться нулю). Статистика (10.60) нехороша тем, что стандартизованная разность (10.53), используемая для сравнения, зависит от размеров выборок n_{i1} и n_{i2} поскольку n_{i1} и n_{i2} влияют на значения $\bar{p}_i$ и $\bar{q}_i$. В связи с указанными недостатками критерий однородности, основанный на (10.60), применять не следует.

Метод суммирования хи-статистик

Схема совместного анализа нескольких четырехклеточных таблиц, описанная в разд. 10.1,— одна из наиболее распространенных, хотя она не всегда используется в явном виде. Давно известно, что метод, обычно называемый методом суммирования хи-статистик, имеет серьезные недостатки. Тем не менее его продолжают применять (например, [Finney, 1965]). Фактически этот метод использует в качестве меры связи величину

$$y_i = z_i = \frac{p_{i1} - p_{i2}}{\sqrt{\bar{p}_i \bar{q}_i (1/n_{i1} + 1/n_{i2})}}. \qquad (10.62)$$

7 Зак. 2383

Так как z_i является нормированной статистикой, то ее стандартная ошибка равна 1 и

$$w_i = \frac{1}{[\text{s.e. } (z_i)]^2} = 1. \tag{10.63}$$

Слово «хи» в названии метода происходит от того, что z_i — квадратный корень из (10.56) — статистики, распределенной по хи-квадрат, т. е. z_i — хи-статистика.

Когда y_i определена величиной (10.62),

$$\bar{y} = \frac{\sum_{i=1}^{g} z_i}{g} = \bar{z}, \tag{10.64}$$

а

$$\sum_{i=1}^{g} w_i = g \tag{10.65}$$

согласно (10.63). Из (10.9) следует, что

$$\chi^2_{\text{assoc}} = \frac{\left(\sum_{i=1}^{g} z_i\right)^2}{g} = g\bar{z}^2. \tag{10.66}$$

Этому подходу присущ серьезный дефект (см., например, [Pasternak and Mantel, 1966]). Рассмотрим по табл. 10.10 численный пример.

Таблица 10.10

Данные для проверки метода суммирования хи-статистик

Группа	Выборка 1		Выборка 2		Объединенная выборка		
	n_{i1}	p_{i1}	n_{i2}	p_{i2}	$n_{i.}$	$\bar{p}_i$	z_i
$i=1$	100	0,60	100	0,40	200	0,50	2,83
$i=2$	1000	0,60	1000	0,40	2000	0,50	8,94

Для группы 1

$$\chi_1^2 = z_1^2 = 8{,}00. \tag{10.67}$$

Для группы 2

$$\chi_2^2 = z_2^2 = 80{,}0. \tag{10.68}$$

Среднее значение z равно:

$$\bar{z} = 1/2\ (28{,}3 + 8{,}94) = 5{,}88, \tag{10.69}$$

что дает по (10.66)

$$\chi^2_{\text{assoc}} = 2 \cdot (5{,}88)^2 = 69{,}15 \qquad (10.70)$$

при 1 степени свободы. Это значение критерия наличия связи для совокупности неудовлетворительно тем, что оно меньше значения $\chi^2_2 = 80$ статистики для одной из групп. Связь, которая согласно данным табл. 10.10 имеется в группе 1, такая же, как и в группе 2, поэтому мы были вправе ожидать, что значимость вывода о наличии связи возрастет при добавлении информации по группе 1. В описанном методе этого не произошло[1].

Методы, в которых добавление информации, подтверждающей наличие связи, не приводит к повышению значимости вывода, применять не следует. Это утверждение в полной мере относится и к процедуре суммирования хи-статистик.

Метод сравнения суммарных наблюдаемых и ожидаемых частот

Следующий метод также характеризуется неадекватным изменением уровня значимости критерия при добавлении информации, подтверждающей наличие связи. Его можно было бы описать в рамках схемы разд. 10.1, но это затруднило бы его восприятие.

В этом методе сначала формируют четырехклеточную таблицу, суммируя соответственные частоты g отдельных таблиц. Пусть наблюдаемые частоты для двух групп $(g=2)$ заданы в табл. 10.11

Таблица 10.11

Данные, иллюстрирующие метод сравнения суммарных наблюдаемых и суммарных ожидаемых частот

	Группа 1				Группа 2			
	B	$\bar{B}$	Сумма		B	$\bar{B}$	Сумма
A	200	30	230	A	40	120	160
$\bar{A}$	10	40	50	$\bar{A}$	40	800	810
Сумма	210	70	280	Сумма	50	920	970
	$x_1^2 = 98{,}20$				$x_2^2 = 154{,}35$		

[1] Логика автора не очень аккуратна, поскольку критерий настроен на гипотезу об отсутствии связи, тем самым с увеличением числа выборок его мощность обязана падать, а это означает, что наличие связи труднее обнаружить.— *Примеч. ред.*

(статистики хи-квадрат вычислены без поправки на непрерывность). Связь между A и B, которой соответствует отношение шансов 26,7, одинакова в обеих группах. В табл. 10.12 приведены значения суммарных наблюдаемых частот.

Таблица 10.12

Сумма частот групп 1 и 2

	B	$\bar{B}$	Сумма
A	240	150	390
$\bar{A}$	20	840	860
Сумма	260	990	1250

На следующем шаге для каждой группы вычисляют частоты, ожидаемые при гипотезе об отсутствии связи. Ожидаемая частота равна произведению двух маргинальных частот, соответствующих данной клетке, поделенному на суммарное число наблюдений в таблице. Например, частота, ожидаемая в клетке (A, B) для группы 2, есть $160 \cdot 50/970 = 8{,}25$. Все ожидаемые частоты приведены в табл. 10.13.

Таблица 10.13

Ожидаемые частоты в группах 1 и 2

Группа 1				Группа 2			
	B	$\bar{B}$	Сумма		B	$\bar{B}$	Сумма
A	172,5	57,5	230	A	8,25	151,75	160
$\bar{A}$	37,5	12,5	50	$\bar{A}$	41,75	768,25	810
Сумма	210	70	280	Сумма	50	920	970

Затем по g таблицам ожидаемых частот формируют таблицу суммарных ожидаемых частот. Для данных нашего примера это табл. 10.14.

Наконец, вычисляют статистику хи-квадрат критерия значимости связи, суммируя по всем четырем клеткам квадраты разностей между суммарной наблюдаемой и суммарной ожидаемой частотами, поделенные на суммарную ожидаемую частоту. Из

Таблица 10.14

Суммы ожидаемых частот в группах 1 и 2

	B	$\bar{B}$	Сумма
A	180,75	209,25	390
$\bar{A}$	79,25	780,75	860
Сумма	260	990	1250

...дя 10.12 и 10.14 мы получим следующие значения для χ^2_{assoc}:

$$\chi^2_{assoc} = \frac{(240-180,75)^2}{180,75} + \frac{(150-209,25)^2}{209,25} + \frac{(20-79,25)^2}{79,25} + \\ + \frac{(840-780,75)^2}{780,75} = 84,99. \qquad (10.71)$$

...то значение меньше обоих значений исходных статистик ...квадрат, данных в табл. 10.11. Следовательно, метод сравнения суммарных наблюдаемых и суммарных ожидаемых частот обладает тем же недостатком, что и метод суммирования ...статистик, и, значит, он также не должен применяться.

Критерий хи-квадрат по суммарной таблице

...ледующий метод проверки значимости общей связи обладает недостатком, противоположным по природе недостатку двух предыдущих методов. Описать этот метод в терминах ...ода 10.1 нельзя. В нем просто требуется сформировать таблицу суммарных наблюдаемых частот, как для предыдущего метода, а затем вычислить по этой таблице статистику хи-квадрат.

Этот метод достаточно хорошо работает, когда соответствующие пропорции в группах близки, в частности, для данных ...дов 10.10—10.11. Однако этот случай является исключением. Рассмотрим табл. 10.15. В обеих группах, судя по данным, связь отсутствует.

Суммарные частоты даны в табл. 10.16, для которой значение статистики хи-квадрат, равное 5,01, указывает на наличие связи при уровне значимости 0,05. Совместный анализ двух таблиц с различными пропорциями и различными отношениями объемов выборок (n_{i1}/n_{i2}) привел к выводу о наличии исходно не существовавшей связи.

Таблица 10.15

Связь между A и B в двух группах

Группа 1				Группа 2			
	B	$\bar{B}$	Сумма		B	$\bar{B}$	Сумма
A	10	40	50	A	60	40	100
$\bar{A}$	20	80	100	$\bar{A}$	30	20	50
Сумма	30	120	150	Сумма	90	60	150
$\chi_1^2=0$				$\chi_2^2=0$			

Таблица 10.16

Сумма частот по группам 1 и 2

	B	$\bar{B}$	Сумма
A	70	80	150
$\bar{A}$	50	100	150
Сумма	120	180	300

Итак, по указанным причинам не следует применять описанные процедуры (см. также [Gart, 1962, Sheehe, 1966]). Это вынуждает нас к более сложным вычислениям, таким, как в 10.2—10.4. Такова цена за качество анализа.

Задачи

10.1. В разд. 10.2 было получено, что отношения шансов в трех группах согласно табл. 10.1 значимо различаются.

а) Отношения шансов o_2' и o_3' имеют близкие значения (см. табл. 10.2). Проверьте значимость их различия с помощью статистики

$$\frac{w_2' w_3'}{w_2' + w_3'} (L_2' - L_3')^2.$$

б) Значения o_2' и o_3' отличаются друг от друга меньше, чем от o_1'. Проверьте значимость различия среднего o_2' и o_3' от o_1'. (*Указание.* Среднее L_2' и L_3' равно $\bar{L}_{23} = (w_2' L_2' + w_3' L_3')/(w_2' + w_3')$.

$$\frac{w_1'(w_2'+w_3')}{w_1'+w_2'+w_3'}(L_1'-\bar{L}_{2,3})$$

и сравнивать со значениями в таблице распределения хи-квадрат с двумя степенями свободы, а не с одной, поскольку разбиение на группы диктуется данными.)

в) Сравните значения статистик в а) и б) со значением $\chi^2_{\text{гом ог}}$ в (10.18).

10.2. Примените методы разд. 10.2 к данным только из 2-й и 3-й групп табл. 10.1 и 10.2. В частности:

а) Вычислите средний логарифм отношения шансов. Найдите его стандартную ошибку. Значимо ли отличается средний логарифм отношения шансов от нуля?

б) Найдите приближенный 95%-ный интервал для логарифма отношения шансов.

в) Вычислите среднее отношение шансов. Найдите 95%-ный доверительный интервал для отношения шансов, соответствующий интервалу в б).

10.3. Хотя это и не совсем очевидно, $\bar{o}_{\text{мн}}$ (см. 10.47) является в действительности взвешенным средним g отдельных отношений шансов

$$o_i=\frac{p_{i1}(1-p_{i2})}{p_{i2}(1-p_{i1})}, \quad i=1, \ldots, g.$$

Покажите, что это верно, найдя совокупность весов $w_1, \ldots, w_g$, такую, что (10.17) равно:

$$\bar{o}_{\text{мн}}=\frac{\sum_{i=1}^{g} o_i w_i}{\sum_{i=1}^{g} w_i}.$$

10.4. Докажите, что при $n_{i1}=n_{i2}=1$, как при исследовании по связанным парам, хи-квадрат статистики Мантела — Ханзела (10.50) и Мак-Немара (8.3) совпадают.

ЛИТЕРАТУРА

Birch, M. W. (1964). The detection of partial association. I: The 2 × 2 case. *J. R. Stat. Soc., Ser B*, **26**, 313–324.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge, Mass.: M.I.T. Press.

Bross, I. D. J. (1966). Spurious effects from an extraneous variable. *J. Chronic Dis.*, **19**, 637–647.

Cochran, W. G. (1954). Some methods of strengthening the common χ^2 tests. *Biometrics*, **10**, 417–451.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, **24**, 295–313.

Cooper, J. E., Kendell, R. E., Gurland, B. J., Sharpe, L., Copeland, J. R. M., and Simon, R. (1972). *Psychiatric diagnosis in New York and London*. London: Oxford University Press.

Cornfield, J. (1956). A statistical problem arising from retrospective studies. Pp. 135–148 in J. Neyman (Ed.). *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, Vol. 4. Berkeley: University of California Press.

Cox, D. R. (1970). *The analysis of binary data*. London: Methuen.

Everitt, B. S. (1977). *The analysis of contingency tables*. New York: Halsted Press.

Fienberg, S. E. (1977). *The analysis of cross-classified categorical data*. Cambridge, Mass.: M.I.T. Press.

Finney, D. J. (1965). The design and logic of a monitor of drug use. *J. Chronic Dis.*, **18**, 77–98.

Gart, J. J. (1962). On the combination of relative risks. *Biometrics*, **18**, 601–610.

Gart, J. J. (1970). Point and interval estimation of the common odds ratio in the combination of 2 × 2 tables with fixed marginals. *Biometrika*, **57**, 471–475.

Gart, J. J. (1971). The comparison of proportions: A review of significance tests, confidence intervals and adjustments for stratification. *Rev. Int. Stat. Inst.*, 39, 16–37.

Goodman, L. A. (1969). On partitioning χ^2 and detecting partial association in three-way contingency tables. *J. R. Stat. Soc., Ser. B*, **31**, 486–498.

McKinlay, S. M. (1975a). The design and analysis of the observational study – A review. (With a Comment by S. E. Fienberg). *J. Am. Stat. Assoc.*, **70**, 503–523.

McKinlay, S. M. (1975b). The effect of bias on estimators of relative risk for pair-matched and stratified samples. *J. Am. Stat. Assoc.*, **70**, 859–864.

McKinlay, S. M. (1975a). A note on the chi-square test for pair-matched samples. *Biometrics*, **31**, 731–735.

McKinlay, S. M. (1978). The effect of non-zero second order interaction on combined estimators of the odds-ratio. *Biometrika*, **65**, 191–202.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *J. Am. Stat. Assoc.*, **58**, 690–700.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother. Rep.*, **50**, 163–170.

Mantel, N. (1977). Tests and limits for the common odds ratio of several 2 × 2 contingency tables: Methods in analogy with the Mantel-Haenszel procedure. *J. Stat. Plann. Inf.*, **1**, 179–189.

Mantel, N., Brown, C., and Byar, D. P. (1977). Tests for homogeneity of effect in an epidemiologic investigation. *Am. J. Epidemiol.*, **106**, 125–129.

Fleiss, J. L. and Fleiss, J. L. (1980). Minimum expected cell size requirements for the Mantel-Haenszel one-degree-of-freedom chi-square test and a related rapid procedure. *Am. J. Epidemiol.*, **112**, 129–134.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.*, **22**, 719–748.

Mantel, N. and Hankey, B. F. (1975). The odds ratios of a 2 × 2 contingency table. *Am. Stat.*, **29**, 143–145.

Miettinen, O. S. (1976). Stratification by a multivariate confounder score. *Am. J. Epidemiol.*, **104**, 609–620.

Norton, W. T. (1967). Small sample considerations in combining 2 × 2 tables. *Biometrics*, **23**, 349–356.

Odoroff, C. L. (1970). A comparison of minimum logit chi-square estimation and maximum likelihood estimation in 2 × 2 × 2 and 3 × 2 × 2 contingency tables: Tests for interaction. *J. Am. Stat. Assoc.*, **65**, 1617–1631.

Pasternack, B. S. and Mantel, N. (1966). A deficiency in the summation of chi procedure. *Biometrics*, **22**, 407–409.

Radhakrishna, S. (1965). Combination of results from several 2 × 2 contingency tables. *Biometrics*, **21**, 86–98.

Rubin, D. B. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, **29**, 185–203.

Schlesselman, J. J. (1978). Assessing effects of confounding variables. *Am. J. Epidemiol.*, **108**, 3–8.

Sheehe, P. R. (1966). Combination of log relative risk in retrospective studies of disease. *Am. J. Public Health*, **56**, 1745–1750.

Yates, F. (1955). The use of transformations and maximum likelihood in the analysis of quantal experiments involving two treatments. *Biometrika*, **42**, 382–403.

Глава 11
Ошибки классификации

До сих пор мы предполагали, что классификация объекта к одной из категорий (заболевшие и не заболевшие, имеющие и не имеющие фактор) происходит безошибочно. Это предположение порой бывает неправомерным. При любом способе сбора данных, будь то интервьюирование, анкетирование с помощью вопросника, получение сведений по официально зарегистрированным случаям, физические тесты или химические анализы, данные могут содержать ошибки. Непреднамеренно (неверное считывание записи, плохо расслышанный ответ) или вследствие неосознанных предубеждений больного, страдающего некоторым заболеванием, его могут отнести к больным с другим заболеванием, и наоборот. Сказанное в полной мере относится и к исходному фактору.

В методах выбора I и II ошибки классификации могут встречаться в обоих изучаемых факторах. В методе выбора III они содержатся только в данных по результирующему фактору (реакции на лечение).

В этой главе мы рассмотрим влияние ошибок классификации на результаты анализа, а в двух последующих опишем некоторые методы уменьшения и оценивания их величины. В разд. 11.1 подробно рассказывается об одном примере, в котором проявляются эффекты ошибочной классификации. В разд. 11.2 алгебраически анализируется воздействие ошибочной классификации на меру связи в случае, когда ошибка возникает только в одном факторе, а в разд. 11.3 — когда оба фактора наблюдаются с ошибками.

11.1. Пример эффектов ошибочной классификации

Дерн и др. [Dern et al., 1963] исследовали в Чикаго частоту дефицита глюкоза-6-фосфат дегидрогеназы (Г-6-ФД) в эритроцитах чернокожих пациентов мужского пола, больных шизофренией. Дефицит Г-6-ФД является нарушением метаболизма, сопряженным с полом, и обнаружен у 10—15% чернокожего мужского населения США.

Данные, полученные Дерном и др., сведены в табл. 11.1. Хи-квадрат для этих данных равен 7,95 и указывает на зависимость, значимую на уровне 0,01.

Таблица 11.1

Зависимость между дефицитом Г-6-ФД и подвидом шизофрении (исследования, проведенные в Чикаго)

Наличие Г-6-ФД	Диагноз		Сумма
	Кататония	Параноя	
Дефицит	15	6	21
Нет дефицита	57	99	156
Сумма	72	105	177

Сравниваемые пропорции — пропорции кататоников и параноиков с дефицитом Г-6-ФД — равны соответственно $p_C = 15/72 = 0{,}208$ и $p_P = 6/105 = 0{,}057$. Найдем значение отношения шансов:

$$o = \frac{15 \cdot 99}{6 \cdot 57} = 4{,}34. \qquad (11.1)$$

Фив и др. [Fieve et al., 1965] повторили такое исследование в пяти государственных клиниках Нью-Йорка. Результаты, полученные в четырех клиниках, приведены в табл. 11.2.

Таблица 11.2

Зависимость между дефицитом Г-6-ФД и подвидом шизофрении в четырех Нью-Йоркских государственных клиниках

Клиника	Кататония		Параноя		o
	N	Процент пациентов с дефицитом	N	Процент пациентов с дефицитом	
Central Islip	32	15,6	80	12,5	1,30
Pilgrim	78	16,7	76	6,6	2,84
Brooklyn	13	30,8	18	11,1	3,56
Kings Park	55	10,9	96	6,3	1,84

Четыре значения отношения шансов значимо не различаются. Среднее отношение шансов

$$\overline{o}_{\text{МН}} = 2{,}09 \qquad (11.2)$$

значимо отличается от единицы на уровне 0,05 (методы сравнения и объединения различных отношений шансов взяты из гл. 10). Результаты для этих четырех клиник подтверждают выводы Дерна и др. [Dern et al., 1963], однако выражение (11.2) указывает на более слабую связь.

В пятой клинике (в Рокленде) отношение шансов также значимо отличалось от единицы (табл. 11.3). Однако возникло следующее затруднение: показатель степени связи указывал на противоположную зависимость.

Таблица 11.3

Зависимость между дефицитом Г-6-ФД и подвидом шизофрении в государственной клинике Рокленда

Кататония		Паранойя		o
N	Процент пациентов с дефицитом	N	Процент пациентов с дефицитом	
28	7,1	29	24,1	0,24

Исследователи быстро связались с администрацией клиники в Рокленде и вздохнули с облегчением, когда узнали, что половина пациентов, больных шизофренией, не была обследована, поскольку была занята в других исследованиях. Пришлось повторить обследование всех чернокожих пациентов клиники мужского пола с диагнозом паранойя и кататония. Результаты второго обследования показаны в табл. 11.4. Отношение шансов снова значимо отличалось от единицы на уровне значимости 0,05, но зависимость противоположного направления была еще сильнее.

Таблица 11.4

Зависимость между дефицитом Г-6-ФД и подвидом шизофрении в государственной клинике Рокленда. Второе обследование

Кататония		Паранойя		o
N	Процент пациентов с дефицитом	N	Процент пациентов с дефицитом	
37	2,7	87	16,1	0,14

Итак, имелись свидетельства в пользу зависимости подвида шизофрении от дефицита фермента, но, к сожалению, в одном из случаев зависимость была противоположна по направлению зависимости в остальных случаях. Для полной неразберихи не хватало лишь большой выборки пациентов, указывающей на

отсутствие значимой разницы между параноиками и кататониками. Как раз такими оказались данные в выборке из 426 пациентов клиники Veterans Administration в Алабаме [Bowman et al., 1965], которые представлены в табл. 11.5. Для них $p_C = 10{,}4\%$, $p_P = 11{,}8\%$, $o = 0{,}87$, хи-квадрат равен 0,07.

Таблица 11.5

Зависимость между дефицитом Г-6-ФД и подвидом шизофрении в клинике Алабамы

Наличие Г-6-ФД	Диагноз		Сумма
	Кататония	Паранойя	
Дефицит	17	31	48
Нет дефицита	146	232	378
Сумма	163	263	426

Таким образом, в литературе приводились доказательства в пользу всех трех видов связи: положительной, отрицательной и нулевой. Эти противоречащие друг другу результаты сведены в табл. 11.6.

Таблица 11.6

Результаты для трех видов связи между дефицитом Г-6-ФД и подвидом шизофрении

Направление связи	Клиника	o
Превалирование кататонии	Чикаго	4,34
	Четыре клиники в Нью-Йорке	2,09
Превалирование паранойи	Рокленд	0,14
Отсутствие различия между кататонией и паранойей	Алабама	0,87

Пытаясь разобраться в такой путанице, исследователи в Нью-Йорке обратились сначала к методам обследования, использовавшимся во всех трех группах. Хотя различие и было обнаружено, оно оказалось недостаточно большим, чтобы объяснить расхождение. Во всяком случае, методы, применявшиеся в Рокленде и в четырех остальных нью-йоркских клиниках, не отличались.

Затем предположили, что различие результатов можно объяснить различием лекарств, назначавшихся пациентам, и по всем исследованиям вычислили отношение шансов отдельно для каждого из основных видов применявшихся лекарств (применяемые лекарства могут влиять на содержание Г-6-ФД). За несколькими исключениями, слишком редкими, чтобы сыграть существенную роль, отношения шансов для каждого вида в пределах одного исследования указывали на зависимость в направлении, совпадавшем с направлением для среднего отношения шансов в этом исследовании.

Каковы бы ни были эффекты, возникающие за счет различия лекарств и методов анализа крови, они «бледнеют» по сравнению с эффектами недостоверности диагноза психического заболевания. Накоплено значительное количество литературы, в которой отмечается, насколько недостоверно психиатрическое диагностирование [Zubin, 1967; Spitzer and Fleiss, 1974]. Возьмем, например, шизофрению. Оказывается, что для тех пациентов, которым один психиатр поставил диагноз шизофрения, второй психиатр поставит тот же диагноз, как правило, лишь примерно в 70% случаев. Далее, можно назвать цифру около 10% для тех пациентов, которым одним из психиатров был поставлен диагноз, отличный от шизофрении, тогда как вторым они будут диагностированы как больные шизофренией. Из редких публикаций о подвидах шизофрении следует, что достоверность диагноза для кататонической и параноидальной шизофрении еще меньше, чем шизофрении в целом.

Во всех исследованиях, посвященных зависимости подвида шизофрении от дефицита Г-6-ФД, истинность диагнозов, поставленных в данной клинике, сомнению не подвергалась, и попытки проверить точность диагностирования не предпринимались. Таким образом, по всей видимости, единственной существенной причиной противоречий между результатами трех исследований, как и между результатами в Рокленде и в остальных клиниках Нью-Йорка, явилась недостоверность диагнозов психических заболеваний.

В пользу различных подходов к диагностированию в пяти нью-йоркских клиниках говорит хотя бы тот факт, что существует заметная разница между пропорциями числа пациентов, отнесенных к кататоникам, среди пациентов, диагностированных как кататоники, либо как параноики. Эти пять клиник различаются по составу поступающих к ним больных, но не настолько сильно, чтобы обеспечить наблюдаемое различие пропорций числа больных кататонией. Задача 11.1 посвящена анализу различия между этими пропорциями.

Описанный пример взят из психиатрии, однако не следует думать, что неточность диагнозов является бичом лишь в пси-

астрии. Недостоверность присуща диагностированию детских состояний [Derryberry, 1938; Bakwin, 1945] и эмфиземы [Fletcher, 1952], расшифровке электрокардиограмм [Davies, 1958] и рентгенограмм [Yerushalmy, 1947; Cochrane and Garland, 1952], определению причин смерти [Markush et al., 1967]. Недостоверности диагнозов в других областях клинической медицины посвящены обзоры [Garland, 1960; Коган, 1975а, 1975b].

В действительности можно принять за аксиому, что ошибки при определении наличия или отсутствия заболевания либо состояния, а также при уточнении вида состояния неизбежны. Подобным же образом подвержено ошибкам и определение наличия исходного фактора (может быть, за исключением пола субъекта).

11.2. Алгебра ошибочной классификации

Существует некоторое недопонимание эффектов ошибочной классификации. Дело в том, что эти ошибки могут исказить показатель связи до такой степени, что, например, при сильной положительной связи он будет указывать на менее сильную положительную или даже отрицательную связь (и наоборот, для отрицательной связи), при отсутствии связи будет указывать на сильную связь. Этот факт расходится с давно сложившимся, но ошибочным мнением, что ошибочная классификация может лишь только уменьшить показатель степени связи [Newell, 1962].

Рассмотрим пример. Для простоты предположим, что заболевание пациента классифицируется точно и что ошибки возникают только при определении наличия или отсутствия изучаемого исходного фактора. Чтобы иметь дело с чем-то конкретным, предположим, что мы сравниваем женщин в возрасте от 55 до 64 лет, которым поставлен диагноз рак легких, и женщин такого же возраста с диагнозом рак молочной железы по тому, курили ли они когда-либо ранее или нет.

Мы предполагаем, что диагноз ставится точно, но факт, курила женщина или нет, определяется с ошибкой[1]. Ошибка при опросе может совершаться как по вине опрашиваемого, так и по вине опрашивающего. В первом случае (отвечает пациентка или ее родственник) возможными причинами ошибок являются:

---

[1] Автор негласно предполагает, что существует некая априори относительно классификация на курящих и не курящих по отношению к изучаемым больным и рассматривает ошибки в связи с этой «идеальной» классификацией. На самом деле модель формирования данных существенно сложней. *Примеч. ред.*

1. Вопрос неправильно понят.

2. При ответе, курила пациентка или не курил неумышленная ошибка.

3. Опрашиваемый преднамеренно ответил не вероятен ответ, что пациентка не курила, чем обра

Возможными причинами ошибок по вине опр- являются:

1. Ответ неправильно понят.

2. Совершена ошибка при записи ответа.

3. Ответ записывается в зависимости от типа Например, опрашивающий чересчур старается устр- можное смещение в сторону наличия зависимости. утверждение «Я изредка курила, когда была ребенко пишет как «Ранее не курила» для больной раком легк нее курила» для больной раком молочной железы.

В [Horwitz and Lysgaard-Hansen, 1975] перечисл- торые другие распространенные причины ошибок мендации по их учету и устранению (см. также мы приведем анализ эффектов ошибочной класси санный в [Keys and Kihlberg, 1963]. Эффекты сификации также изучались в [Rogot, 1961; Mots son, 1965; Assakul and Proctor, 1967; Koch, 1969, 1975; Copeland et al., 1977].

Рассмотрим сначала больных с диагнозом рак предположению, они выделены безошибочно). Пусть тинная пропорция больных, куривших когда-либо ранее, и, чит, $1-P_L$ — истинная пропорция не куривших больных. С значим через E_L показатель чувствительности тестир- определению факта «больная ранее курила», т. е. - того, что больная раком легких, курившая ранее, отн- некурившим, а через F_L — показатель альтернативность вероятность того, что не курившая ранее больная отнесена ривших. Пытаясь оценить истинную пропорцию куривш по собранным данным, мы в действительности оцениваем

$$p_L = (1-E_L)P_L + F_L(1-P_L). \qquad (11.3)$$

Фактически оцениваемая пропорция p_L может быть меньше, больше или равна истинной пропорции P_L в зависимости[1] от отношения величин E_L и F_L:

$$p_L > P_L, \text{ если } F_L/(E_L+F_L) > P_L,$$
$$p_L = P_L, \text{ если } F_L/(E_L+F_L) = P_L,$$
$$p_L < P_L, \text{ если } F_L/(E_L+F_L) < P_L.$$

[1] Из-за равенства $(p_L-P_L)=(E_L+F_L)\left(\frac{F_L}{E_L+F_L}-P_L\right)$, вытекающего из (11.3).— *Примеч. ред.*

При ошибках значениях E_L и F_I оценка будет завышена, если P_L меньше 0,5, и занижена, если P_L больше 0,5. Значит, даже если вероятности ошибок равны, ошибки не обязательно компенсируют друг друга.

Пусть теперь P_B — истинная пропорция больных раком молочной железы, куривших когда-либо ранее, а E_B и F_B — параметры чувствительности и альтернативности для больных раком молочной железы, аналогичные E_I и F_L. Тогда можно оценить лишь пропорцию тех, кто зарегистрирован как куривший среди больных раком молочной железы:

$$p_B = (1-E_B)P_B + F_B(1-P_B). \tag{11.4}$$

Анализ влияния ошибок на отношение шансов сложен [Diamond and Lilienfeld, 1962a, 1962b; Goldberg, 1975]. Графическое исследование влияния ошибок дано в [Copeland et al., 1977]. Поэтому мы предположим, что связь между курением и видом рокового заболевания измеряется простой разностью пропорций куривших[1]. Мы можем оценить не истинную разность

$$D = P_L - P_B, \tag{11.5}$$

а лишь $d = p_L - p_B$ — разность наблюдаемых пропорций. Как легко видеть, она сводится к

$$d = D + (F_I - F_B) + P_B(E_B + F_B) - P_L(E_L + F_L). \tag{11.6}$$

что показывает, что d — как правило, смещенная оценка D.

Фактически оцениваемая разность d может быть меньше, больше или равна истинной разности D. Она может даже иметь противоположный знак, т. е. может приводить к выводу о зависимости, обратной по отношению к истинной зависимости.

Такое обращение зависимости невозможно, в том случае, когда равны вероятности E_L и E_B, так что

$$1-E_L = 1-E_B = 1-E, \tag{11.7}$$

и вероятности F_L и F_B:

$$1-F_L = 1-F_B = 1-F. \tag{11.8}$$

Подставляя выражения (11.7) и (11.8) в (11.6) и проводя упрощения, получим, что разность между наблюдаемыми пропорциями равна:

$$d = D[1-(E+F)]. \tag{11.9}$$

---

[1] Качественно результаты выводов при такой характеризации связи практически те же, что и при отношениях шансов.— *Примеч. ред.*

Из равенства (11.9) следует, во-первых, что разность d, которую можно оценивать по наблюдаемым данным, не может равняться истинной разности D, если только обе вероятности E и F ошибочной классификации не равны нулю, а во-вторых, при естественном допущении, что и E, и F меньше $^1/_2$ (т. е. доля любой из этих ошибок меньше 50%), наблюдаемая разность имеет тот же знак, что и истинная, хотя и меньше ее по абсолютной величине, т. е. ближе к нулю. Как раз этот случай, рассмотренный Броссом в его классической работе [Bross, 1954], является одним из тех частных случаев, которые породили неверное мнение, что ошибки классификации всегда приводят лишь к сглаживанию различия между двумя пропорциями. Однако эту ситуацию (совпадение E_L и F_L и совпадение E_B и F_B) следует считать исключением (см., например, [Lilienfeld and Graham, 1958; Goldberg, 1975]).

Что касается отношения шансов, то эффекты также могут быть любыми. В ранее обсуждавшемся частном случае, когда $E_L = E_B < {}^1/_2$ и $F_L = F_B < {}^1/_2$, оценка степени связи для отношения шансов занижена, как и для разности долей. Точнее, если ω — истинное значение отношения шансов, а o — оценка по данным с ошибками классификации, то при $\omega > 1$ мы должны ожидать, что в результате получим $\omega > o > 1$, т. е. значение оценки отношения шансов будет превышать единицу, но не настолько, как ω.

11.3. Случай ошибочной классификации обоих факторов

Предыдущий раздел был посвящен случаю, когда ошибочно классифицируется только один из факторов. В [Keys and Kihlberg, 1963] приведено общее обсуждение ситуации, более близкой к реальности, когда при классификации происходят ошибки по обоим факторам. Приведенные ниже результаты и пример, связанные с вычислением отношения шансов, взяты из [Barron, 1977].

Пусть A и $\bar{A}$ — наличие и отсутствие одного из изучаемых факторов, а B и $\bar{B}$ — наличие и отсутствие другого фактора. Обозначим через $P(AB)$, $P(A\bar{B})$ и т. д. вероятности совместного выполнения событий при полном отсутствии ошибок классификации. Предположим, что

$$\omega = \frac{P(AB)\,P(\bar{A}\bar{B})}{P(A\bar{B})\,P(\bar{A}B)} \tag{11.10}$$

— истинное отношение шансов, связывающее события A и B.

Теперь допустим, что оба фактора подвержены ошибкам классификации с вероятностями правильной и ошибочной классификации, приведенными в табл. 11.7. Предполагается, что ошибки по каждому из факторов совершаются независимо.

Таблица 11.7

Вероятности правильной и неправильной классификации по факторам A и B

Истинное состояние	Классификация состояния		Истинное состояние	Классификация состояния	
	A	$\bar{A}$		B	$\bar{B}$
A	a_1	$1-a_1$	B	b_1	$1-b_1$
$\bar{A}$	a_2	$1-a_2$	$\bar{B}$	b_2	$1-b_2$

Обозначим через $p(AB)$, $p(A\bar{B})$ и т. д. вероятности совместного выполнения событий, когда классификация происходит с ошибками, определенными выше. В явном виде это можно представить как

$$p(AB) = a_1 b_1 P(AB) + a_1 b_2 P(A\bar{B}) + a_2 b_1 P(\bar{A}B) + a_2 b_2 P(\bar{A}\bar{B}); \tag{11.11}$$

$$p(A\bar{B}) = a_1(1-b_1)P(AB) + a_1(1-b_2)P(A\bar{B}) + a_2(1-b_1)P(AB) + a_2(1-b_2)P(\bar{A}\bar{B}); \tag{11.12}$$

$$p(\bar{A}B) = (1-a_1)b_1 P(AB) + (1-a_1)b_2 P(A\bar{B}) + (1-a_2)b_1 P(\bar{A}B) + (1-a_2)b_2 P(\bar{A}\bar{B}); \tag{11.13}$$

$$p(\bar{A}\bar{B}) = (1-a_1)(1-b_1)P(AB) + (1-a_1)(1-b_2)P(A\bar{B}) + (1-a_2)(1-b_1)P(\bar{A}B) + (1-a_2)(1-b_2)P(\bar{A}\bar{B}). \tag{11.14}$$

Наблюдаемое отношение шансов равно:

$$o = \frac{p(AB)\,p(\bar{A}\bar{B})}{p(A\bar{B})\,p(\bar{A}B)} \tag{11.15}$$

и может сильно отличаться от ω в (11.10).

Теперь рассмотрим пример. Пусть вероятности совместного выполнения событий для больных гипертонией (A) и раком

матки (B) при безошибочной диагностике имеют значения, приведенные в табл. 11.8.

Таблица 11.8

Вымышленные данные о совместных вероятностях заболевания гипертонией и раком матки (оба фактора наблюдаются с ошибками)

Гипертония	Рак	
	B	$\bar{B}$
A	0,122	0,060
$\bar{A}$	0,211	0,607

Далее, примем, что наличие гипертонии определяется правильно с вероятностью $a_1=0{,}90$, ее отсутствие — с вероятностью $1-a_2=0{,}98$, наличие и отсутствие рака матки — с вероятностями соответственно $b_1=0{,}95$ и $1-b_2=0{,}98$. Все эти (довольно высокие) значения были определены эмпирически (см. ссылки в [Barron, 1977]).

Исходя из (11.11) — (11.14) найдем наблюдаемые доли совместного выполнения событий:

$$p(AB)=0{,}11, \tag{11.16}$$

$$p(A\bar{B})=0{,}07, \tag{11.17}$$

$$p(\bar{A}B)=0{,}22, \tag{11.18}$$

$$p(\bar{A}\bar{B})=0{,}6. \tag{11.19}$$

Отсюда наблюдаемое отношение шансов равно:

$$o=\frac{0{,}11\cdot 0{,}60}{0{,}07\cdot 0{,}22}=4{,}29 \tag{11.20}$$

и на 25% меньше отношения шансов для точных значений величин в табл. 11.8:

$$\omega=\frac{0{,}122\cdot 0{,}607}{0{,}060\cdot 0{,}211}=5{,}85. \tag{11.21}$$

Задачи

11.1. Выше упоминалось различие пяти нью-йоркских государственных клиник в пропорциях числа пациентов, отнесенных к кататоникам, среди пациентов, диагностированных как кататоники или параноики.

а) Вычислите по приведенным ниже частотам указанные пропорции.

Клиника	Кататоники	Параноики	Сумма	Пропорция кататоников
Central Islip	32	80	112	p_1
Pilgrim	78	76	154	p_2
Brooklyn	13	18	31	p_3
Kings Park	55	96	151	p_4
Rockland	37	87	124	p_5
Сумма	215	357	572	$\bar{p}$

б) В (9.4) дана статистика хи-квадрат для сравнения нескольких пропорций. Вычислите ее значение для полученных пропорций.

в) Соотнесите вычисленное значение статистики со значениями табл. А.1 при четырех степенях свободы. На каком уровне значимости можно отвергнуть гипотезу равенства пропорции? Что можно сказать о подходах к дифференциальной диагностике кататонической и параноидальной шизофрении в этих пяти клиниках?

11.2. Допустим, что доля курящих женщин с диагнозом рак легких в возрасте от 55 до 64 лет равна $p_L = 0{,}50$, а вероятности ошибок составляют $E_L = 0{,}25$ и $F_L = 0{,}05$.

а) Вычислите значение (11.3) фактически оцениваемой пропорции p_L куривших женщин.

Допустим, что доля курящих женщин с диагнозом рак молочной железы в возрасте от 55 до 64 лет равна $p_B = 0{,}40$, а вероятности ошибок составляют $E_B = F_B = 0{,}10$.

б) Вычислите значение (11.4) фактически оцениваемой пропорции p_B куривших женщин.

в) Вычислите $P_L - P_B$, $p_L - p_B$. Сравните эти разности.

г) Вычислите отношение шансов по P_L и P_B, по p_L и p_B. Сравните полученные значения.

ЛИТЕРАТУРА

Assakul, K. and Proctor, C. H. (1967). Testing independence in two-way contingency tables with data subject to misclassification. *Psychometrika*, **32**, 67–76.

Bakwin, H. (1945). Pseudodoxia pediatrica. *New Engl. J. Med.*, **232**, 691–697.

Barron, B. A. (1977). The effects of misclassification on the estimation of relative risk. *Biometrics*, **33**, 414–418.

Bowman, J. E., Brewer, G. J., Frischer, H., Carter, J. L., Eisenstein, R. B., and Bayrakci, C. (1965). A re-evaluation of the relationship between glucose-6-phosphate dehydrogenase deficiency and the behavioral manifestations of schizophrenia. *J. Lab. Clin. Med.*, **65**, 222–227.

Bross, I. (1954). Misclassification in 2 × 2 tables. *Biometrics*, **10**, 478–486.

Cochrane, A. L. and Garland, L. H. (1952). Observer error in interpretation of chest films: International investigation. *Lancet*, **2**, 505–509.

Copeland, K. T., Checkoway, H., McMichael, A. J., and Holbrook, R. H. (1977). Bias due to misclassification in the estimation of relative risk. *Am. J. Epidemiol.*, **105**, 488–495.

Davies, L. G. (1958). Observer variation in reports on electrocardiograms. *Brit. Heart J.*, **20**, 153–161.

Dern, R. J., Glynn, M. F., and Brewer, G. J. (1963). Studies on the correlation of the genetically determined trait G-6-PD deficiency with behavioral manifestations in schizophrenia. *J. Lab. Clin. Med.*, **62**, 319–329.

Derryberry, M. (1938). Reliability of medical judgments on malnutrition. *Public Health Rep.*, **53**, 263–268.

Diamond, E. L. and Lilienfeld, A. M. (1962a). Effects of errors in classification and diagnosis in various types of epidemiological studies. *Am. J. Public Health*, **52**, 1137–1144.

Diamond, E. L. and Lilienfeld, A. M. (1962b). Misclassification errors in 2 × 2 tables with one margin fixed: Some further comments. *Am. J. Public Health*, **52**, 2106–2110.

Fieve, R. R., Brauninger, G., Fleiss, J. L., and Cohen, G. (1965). Glucose-6-phosphate dehydrogenase deficiency and schizophrenic behavior. *J. Psychiatr. Res.*, **3**, 255–262.

Fletcher, C. M. (1952). Clinical diagnosis of pulmonary emphysema–an experimental study. *Proc. R. Soc. Med.*, **45**, 577–584.

Garland, L. H. (1960). The problem of observer error. *Bull. N.Y. Acad. Med.*, **36**, 570–584.

Goldberg, J. D. (1975). The effects of misclassification on the bias in the difference between two proportions and the relative odds in the fourfold table. *J. Am. Stat. Assoc.*, **70**, 561–567.

Horwitz, O. and Lysgaard-Hansen, B. (1975). Medical observations and bias. *Am. J. Epidemiol.*, **101**, 391–399.

Keys, A. and Kihlberg, J. K. (1963). The effect of misclassification on estimated relative prevalence of a characteristic. *Am. J. Public Health*, **53**, 1656–1665.

Koch, G. G. (1969). The effect of non-sampling errors on measures of association in 2 × 2 contingency tables. *J. Am. Stat. Assoc.*, **64**, 852–863.

Koran, L. M. (1975a). The reliability of clinical methods, data and judgments, part 1. *New Engl. J. Med.*, **293**, 642–646.

Koran, L. M. (1975b). The reliability of clinical methods, data and judgments, part 2. *New Engl. J. Med.*, **293**, 695–701.

Lilienfeld, A. M. and Graham, S. (1958). Validity of determining circumcision status by questionnaire as related to epidemiological studies of cancer of the cervix. *J. Natl. Cancer Inst.*, **21**, 713–720.

Markush, R. E., Schaaf, W. E., and Seigel, D. G. (1967). The influence of the death certificate on the results of epidemiologic studies. *J. Natl. Med. Assoc.*, **59**, 105–113.

Mote, V. L., and Anderson, R. L. (1965). An investigation of the effect of misclassification on the properties of chi-square tests in the analysis of categorical data. *Biometrika*, **52**, 95–109.

Newell, D. J. (1962). Errors in the interpretation of errors in epidemiology. *Am. J. Public Health*, **52**, 1925–1928.

Ropot, E. (1961). A note on measurement errors and detecting real differences. *J. Am. Stat. Assoc.*, **56**, 314–319.

Spitzer, R. L. and Fleiss, J. L. (1974). A re-analysis of the reliability of psychiatric diagnosis. *Br. J. Psychiatry*, **125**, 341–347.

Yerushalmy, J. (1947). Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. *Public Health Rep.*, **62**, 1432–1449.

Zubin, J. (1967). Classification of the behavior disorders. Pp. 373–406 in P. R. Farnsworth, O. McNemar, and Q. McNemar (Eds.). *Annual review of psychology*. Palo Alto, Calif.: Annual Reviews.

Глава 12
Контроль ошибок классификации

В предыдущей главе мы рассмотрели некоторые эффекты, возникающие из-за ошибочной классификации. В разд. 12.1. этой главы мы обсудим статистические методы контроля ошибок. Алгебраические результаты для вероятностного контроля даются в разд. 12.2, а отдельные методы контроля ошибок непосредственно в эксперименте обсуждаются в разд. 12.3.

12.1. Статистический контроль ошибок

Часто исследователь располагает несколькими методами определения состояния пациента. Некоторые из них довольно дорогие, но в то же время надежные (т. е. слабо подверженные ошибкам), а другие — сравнительно дешевые, но зато ненадежные. Планируя обследование или сравнительные испытания, даже при умеренных объемах выборок, исследователь должен применять ненадежные методы, чтобы максимально снизить стоимость исследования [Rubin et al., 1956].

Если ученый ограничивается только недостоверными методами, он рискует получить смещенные оценки типа тех, которые были описаны в предыдущей главе. Однако, извлекая из полной выборки подвыборку, к которой затем применяются и неточный, и точный методы, исследователь сможет с относительно небольшими дополнительными затратами оценить вероятности ошибочной классификации и ввести поправки на смещение.

Рассмотрим в качестве примера классификацию объектов некоторой выборки по степени пристрастия к курению. Жертвуя надежностью в пользу простоты, можно положиться исключительно на опрос обследуемых. Напротив, если построить исследование на основе химических анализов концентрации тиоцианатов в моче, слюне и плазме обследуемых [Densen et al., 1967], то дешевизна будет принесена в жертву точности.

Допустим, надо оценить степень пристрастия к курению у каждой из N обследуемых в выборке госпитализированных больных с диагнозом рак легких. Пусть принято решение использовать только устный ответ каждой женщины о привычке к курению. Для простоты каждая пациентка будет характеризоваться как много курящая (например, выкуривающая в среднем за день 10 или более сигарет) или не много курящая. Пусть p_L обозначает пропорцию тех женщин, которые сказали, что они курят много.

Выбранный способ получения информации о степени пристрастия к курению недорог, но плох тем, что сильно подвержен возможным ошибкам. Поэтому предположим, что исследователь, решив оценить величину ошибки, которая происходит при классификации по устным ответам, взял подвыборку объема из всех тех же N пациенток с раком легких и протестировал концентрацию тиоцианатов в плазме. Положительный результат теста выделяет много курящих пациенток, отрицательный — мало курящих.

Разумеется, степень пристрастия к курению определяется по анализу крови тоже неточно, и не только из-за того, что дихотомия пациенток на много и мало курящих неточна, но и вследствие влияния случайных флуктуаций на результаты самого теста. Однако ввиду большей воспроизводимости этот тест можно считать стандартом по отношению к устному опросу.

Таблица 12.1

Пристрастие к курению, определенное на основе устного опроса и на основе анализа крови

Опрос	Анализ		
	Сильное	Несильное	Сумма
Сильное	n_{00}	n_{01}	n_0
Несильное	n_{10}	n_{11}	n_1

Пусть частоты в табл. 12.1 относятся к результатам классификации подвыборки из n больных раком легких по степени пристрастия к курению обоими способами. Обозначения взяты из [Tenenbein, 1970, 1971]. По этим данным с помощью величины n_{00}/n_0 можно оценить пропорцию числа женщин, отнесенных по результатам теста к категории много курящих среди много курящих по результатам опроса, а с помощью n_{10}/n_1 — пропорцию числа мало курящих женщин по результа-

там теста среди мало курящих по результатам опроса. Мы обозначили P_L общую пропорцию женщин, отнесенных при опросе к категории много курящих. Легко проверить, что оценкой общей пропорции женщин, отнесенных тестом к этой категории, является

$$P_L = \frac{n_{00}}{n_0} p_L + \frac{n_{10}}{n_1} (1-p_L). \quad (12.1)$$

Оценкой стандартной ошибки p_L является просто $\sqrt{p_L(1-p_L)/N}$, тогда как выражение для стандартной ошибки P_L более сложно:

$$\text{s.e.}(P_L) = \sqrt{\frac{P_L(1-P_L)}{N}\left(1+(1-K)\frac{N-n}{n}\right)}, \quad (12.2)$$

где

$$K = \frac{1-p_L}{p_L} \cdot \frac{(P_L - n_{10}/n_1)^2}{P_L(1-P_L)}. \quad (12.3)$$

Оценка (12.1) и ее стандартная ошибка (12.2) получены Тененбейном [Tenenbein, 1970, 1971], где даны также критерии для выбора подходящего значения n.

Деминг [Deming, 1977] предложил интересную идею, которая, по сути, состоит в приложении схемы двойного выбора Тененбейна к задаче выборочного обследования, и дал некоторые другие критерии определения подходящего n. Кьяккерини и Арнольд [Chiacchierini and Arnold, 1977] обобщили схему Тененбейна на случай, когда оба фактора подвержены ошибкам, а Хохберг [Hochberg, 1977] расширил ее до случая многомерных частотных таблиц. Другие подходы к оцениванию величин поправок рассмотрены в [Harper, 1964, Bryson, 1965, Press, 1968].

Проиллюстрируем описанные алгебраические результаты на следующем числовом примере. Пусть опрошено всего 200 женщин с раком легких, и 88 из них дали такие ответы, что их следует отнести к много курящим. Значит, наблюдаемая (но смещенная) доля много курящих среди пациенток с раком легких равна:

$$p_L = 0{,}44. \quad (12.4)$$

Теперь допустим, что 50 из 200 опрошенных пациенток были протестированы на уровень тиоцианатов серы и что в результате получена частотная табл. 12.2. Тогда значения поправок равны соответственно $n_{00}/n_0 = 18/20 = 0{,}90$ и $n_{10}/n_1 = 6/30 = 0{,}20$. Их подстановка в (12.1) дает в качестве лучшей оценки доли много курящих в этой группе значение

$$P = 0{,}90 \cdot 0{,}44 + 0{,}20 \cdot 0{,}56 = 0{,}51. \quad (12.5)$$

Значение оценки в (12.4) более чем на 10% меньше, чем в

Таблица 12.2

Оценка пристрастия к курению у 50 больных раком легких, определенная на основе устиого опроса и иа основе химического анализа крови

Опрос	Химический анализ		
	Сильное	Несильное	Сумма
Сильное	18	2	20
Несильное	6	24	30

Чтобы определить стандартную ошибку P_L, сначала вычислим значение K в (12.3):

$$K = \frac{0,56}{0,44} \cdot \frac{(0,51-0,20)^2}{0,51 \cdot 0,49} = 0,4894. \qquad (12.6)$$

Оценка стандартной ошибки P_L, данная в (12.2), равна:

$$\text{s.e.}\,(P_L) = \sqrt{\frac{0,51 \cdot 0,49}{200}\left(1 + 0,5106 \cdot \frac{150}{50}\right)} = 0,06. \qquad (12.7)$$

Если проводится сравнительное исследование, в котором, например, сравниваются доля много курящих среди пациенток с раком легких и среди пациенток с раком молочной железы, то для второй группы также следует извлечь подвыборку и провести обследование более точным методом (например, провести анализ крови в подвыборке пациенток с раком молочной железы). В задаче 12.1 приведены данные, с помощью которых требуется провести сравнение.

12.2. Вероятностный контроль ошибок

Информацию о величине ошибки нередко можно получить из внешних источников. В примере, описанном в разд. 11.3, доли правильного и неправильного распознавания гипертонии и рака матки были определены независимо от данных, по которым изучалась связь между двумя этими заболеваниями. В этом разделе мы используем обозначения из разд. 11.3 и результаты из [Barron, 1977].

Оценки $p(AB)$, $p(A\bar{B})$ и т. д. получают в результате проведенного обследования по данным с ошибками, а значения a_1, a_2, b_1 и b_2 (см. табл. 11.7) берут из другого источника. Тогда, обращая уравнения (11.11) — (11.14), можно получить значения для истинных вероятностей. Обозначим

$$p(A) = p(AB) + p(A\bar{B}), \tag{12.8}$$

$$p(B) = p(AB) + p(\bar{A}B). \tag{12.9}$$

Истинные вероятности равны:

$$P(AB) = \frac{p(AB) + a_2 b_2 - a_2 p(B) - b_2 p(A)}{(a_1 - a_2)(b_1 - b_2)}, \tag{12.10}$$

$$P(A\bar{B}) = \frac{-p(AB) - a_2 b_1 + a_2 p(B) + b_1 p(A)}{(a_1 - a_2)(b_1 - b_2)}, \tag{12.11}$$

$$P(\bar{A}B) = \frac{-p(AB) - a_1 b_2 + a_1 p(B) + b_2 p(A)}{(a_1 - a_2)(b_1 - b_2)}, \tag{12.12}$$

$$P(\bar{A}\bar{B}) = \frac{p(AB) + a_1 b_1 - a_1 p(B) - b_1 p(A)}{(a_1 - a_2)(b_1 - b_2)}. \tag{12.13}$$

Используя (см. (11.16)—(11.18)) значения $p(AB) = 0{,}11$, $p(A) = 0{,}11 + 0{,}07 = 0{,}18$, $p(B) = 0{,}11 + 0{,}22 = 0{,}33$ и $a_1 = 0{,}9$, $a_2 = 0{,}02$, $b_1 = 0{,}95$, $b_2 = 0{,}02$ для точных вероятностей получим значения

$$P(AB) = 0{,}122, \tag{12.14}$$

$$P(A\bar{B}) = 0{,}06, \tag{12.15}$$

$$P(\bar{A}B) = 0{,}211 \tag{12.16}$$

и

$$P(\bar{A}\bar{B}) = 0{,}607, \tag{12.17}$$

совпадающие со значениями в табл. 11.8.

Если имеются независимые оценки вероятностей правильной классификации, то точные вероятности можно вычислять по формулам (12.10)—(12.13). По этим скорректированным значениям рассчитывать величину требуемой меры степени связи, конечно, лучше, чем по вероятностям, полученным непосредственно из данных с ошибками. К сожалению, явные выражения для стандартных ошибок в этом случае еще не выведены.

12.3. Контроль ошибок в эксперименте

Намеченный план исследований почти всегда можно модифицировать так, чтобы возможная величина ошибки стала меньше. Здесь мы расскажем лишь о небольшом числе из мно-

жества существующих идей и методов. Один из методов основан на том, что надлежаще спланированные клинические испытания являются «дважды слепыми», т. е. при сравнении лекарственных препаратов или видов лечения, и пациенту, и сотруднику, оценивающему состояние пациента после лечения, неизвестен вид лечения, назначенный этому пациенту.

Эта идея— держать в неведении как больного, так и врача, оценивающего его состояние,— вполне осуществима в исследованиях, в которых наличие или отсутствие заболевания и наличие или отсутствие исходного фактора должны определяться практически одновременно (например, когда исследование проводится по пациентам, направленным в хирургическое отделение, и нет ни предшествующих данных, ни возможности проследить за пациентом в будущем). Одна из опасностей состоит в том, что диагност, зная, есть фактор или нет, способен предвзято судить о том, диагноз какого из заболеваний, включенных в исследование, поставлен больному. Чтобы контролировать ошибки такого вида, диагносту указывают, что он не должен интересоваться исходным фактором до тех пор, пока это не станет патогномоничным, противоестественным.

Вторая опасность — если лицу, которое решает, присутствует фактор или отсутствует, известен поставленный пациенту диагноз, то это может повлиять на его решение. В этом случае контроль состоит в том, чтобы не сообщать диагноз лицу, принимающему решение. Третья опасность заключается в том, что пациент может по-разному реагировать на лечение в зависимости от того, знает ли пациент, пусть предположительно, чем он болен. Выход здесь — держать в тайне от больного его диагноз в течение такого времени, которое достаточно, чтобы собрать всю нужную информацию. Какую долю правды можно скрывать от пациента и как долго — эти этические проблемы нужно решать отдельно (см. [Levin, 1954]).

До сих пор мы предполагали, что диагност и исследователь, собирающий информацию о сопутствующих факторах — не одно и то же лицо. Однако эти две роли разделимы не всегда, и в каждом исследовании надо быть готовым к тому, что они совмещены. Смещение, которое может при этом возникнуть, иллюстрируется исследованием, которое было посвящено изучению психических расстройств, сопряженных с системной красной волчанкой (СКВ).

Несколько лет назад стало увеличиваться число сообщений о том, что среди больных СКВ очень сильно распространены (как по числу случаев, так и по широте проявлений) психические расстройства. Изучая это явление, психиатры (Ganz et al., 1972] опрашивали по некоторой структурированной схеме больных в выборках с СКВ и с ревматоидным артритом (РА),

поступивших в четыре нью-йоркские клиники. Больные РА служили контрольной группой, поскольку эти два заболевания очень сходны и сообщений о психических расстройствах у больных РА было очень мало.

Были опрошены 68 больных СКВ и 36 больных РА. На основе опросов каждый пациент был охарактеризован по степени психических нарушений: сильные расстройства (например, очень высокая тревожность либо депрессия или галлюцинации и мании), умеренные (например, невысокая тревожность и депрессия), отсутствие психических отклонений. Было сделано все возможное, чтобы опрашивающий не знал диагноз пациента. В результате опросов были получены данные, отраженные в табл. 12.3.

Таблица 12.3

Психиатрическая симптоматичность в зависимости от диагноза
(результаты интервьюирования)

Симптомы	Системная красная волчанка		Ревматоидный артрит	
	N	%	*N*	%
Сильные	24	35	9	25
Умеренные	19	28	12	33
Отсутствуют	25	37	15	42
Сумма	68	100	36	100

Пропорции для двух диагностических категорий сходны. Статистика хи-квадрат с двумя степенями свободы для сравнения двух распределений равна 1,16, т. е. незначима при любом разумном уровне значимости. Итак, заключение, которое следует сделать на основе структурированных опросов, проводившихся сотрудниками, которым не был известен диагноз, состоит в том, что по психиатрической симптоматике больные СКВ отличаются от больных РА незначимо.

Еще одним шагом в исследовании стал анализ записей, сделанных лечащими врачами в тот же день, когда проводился опрос. Результаты опроса врачу не сообщались. Записи в истории болезни анализировались с помощью тех же критериев психиатрической симптоматичности, которые использовались при обработке результатов опросов. Анализ записей в каждой истории болезни проводился сотрудником, не опрашивавшим данного пациента. В итоге была составлена табл. 12.4.

Таблица 12.4

Психиатрическая симптоматичность в зависимости от диагноза — результаты по записям в историях болезни

Симптомы	Системная красная волчанка		Ревматоидный артрит	
	N	%	*N*	%
Сильные	5	7	0	0
Умеренные	15	22	2	6
Отсутствуют	48	71	34	94
Сумма	68	100	36	100

Теперь пропорции для двух диагностических категорий сильно различаются. Например, на основе данных историй болезни около трети пациентов с СКВ охарактеризовано как имеющие психические нарушения, в отличие от 6% пациентов с РА. Значение статистики хи-квадрат для табл. 12.4 равно 8,27 и указывает на различие, значимое на уровне 0,05.

Это различие, не обнаруженное по данным опросов, можно объяснить наличием положительной «обратной связи». Чем больше публикуется сообщений о высокой частоте психических расстройств у больных СКВ, тем больше клиницистов будет обращать внимание на эти расстройства и регистрировать их. Само по себе это вовсе не плохо. Однако научная ценность таких наблюдений весьма сомнительна, если нарушения у больных СКВ регистрируются более тщательно за счет внимания к психическим расстройствам у других больных.

Итак, ясно, что исследователь должен использовать такую информацию о сопутствующих факторах, которая получена лицами, которым неизвестно состояние пациента. Целесообразность проведения структурированного опроса или обследования — когда анализы и вопросы запланированы заранее и закодированы возможные варианты ответов и результатов анализов — может быть не столь очевидной.

Когда опрос проводится по установленному образцу, всем опрашиваемым задают одни и те же вопросы, касающиеся одних и тех же факторов. Таким путем устраняются различия в способе опроса и, значит, смещения, возникающие за счет того, что различным пациентам задают различные вопросы. Предварительная кодировка возможных ответов не только облегчает ввод данных и их машинную обработку, но и избавляет от необходимости разбираться и придавать четкий смысл устным ответам, иногда просто анекдотическим.

Структурированные опросы с успехом применяются в психиатрии [Spitzer et al., 1967, Wing et al., 1967, Spitzer et al, 1970] и при исследовании заболеваний дыхательных путей [Medical Research Council, 1966]. Необходимость таких процедур в психиатрической и бронхолегочной медицине обусловлена в основном различием между клинистами в отношении способов получения информации и ее последующей интерпретации.

Аналогичное положение свойственно почти любой области медицины, и нет убедительных доводов, по которым было бы невозможно распространить идею проведения структурированных опросов. Их применимость при заполнении истории болезни очевидна. Но эта идея настолько же полезна и при расшифровке рентгенограмм, электрокардиограмм (ЭКГ) и т. п. Важной причиной несогласованности врачей при постановке диагнозов сердечных заболеваний является различие в интерпретации ЭКГ кардиологами. Даже один и тот же кардиолог может интерпретировать одну электрокардиограмму по-разному в различных случаях. Естественно, многие из этих различий исчезнут, если кардиологи будут регистрировать отклонения от нормы, обнаруженные ими на каждом цикле ЭКГ, через заранее установленный код. Сказанное в такой же мере относится, конечно, и к распознаванию патологических изменений на рентгенограммах.

Запись информации в закодированном виде обеспечивает, кроме стремления к единообразию, достижения еще одной цели.

При подходящем способе кодирования данные можно охарактеризовать количественно, и, следовательно, обеспечить более объективную классификацию пациентов по степени тяжести заболевания по сравнению с клиническим выводом. Главное препятствие применению кодирования данных состоит в том, что, даже если кто-то найдет в себе силы и терпение разработать, проверить и переработать, если это будет необходимо, систему кодов, клиницисты, незнакомые с этой системой, относятся к ней без всякого энтузиазма. Однако отказ от предоставления данных в таком виде едва ли можно оправдать, учитывая, что кодирование данных приближает медицинские и эпидемиологические исследования к идеалу любого научного исследования, в котором все использовавшиеся критерии сообщаются и, значит, обеспечивается воспроизводимость исследования.

Задачи

12.1. В разделе 12.1 рассматривался способ коррекции смещения наблюдаемых пропорций. Теперь мы обсудим сравнение двух пропорций, наблюдаемых с ошибками.

а) Значение p_L, доли много курящих среди больных раком легких, полученное по устным ответам, составило 0,44. Допустим, что опрошено 200 женщин с раком молочной железы и 60 из них сказали, что они курят много. Вычислите пропорцию p_B много курящих среди больных раком молочной железы, а также значение отношения шансов для связи между видом рака и интенсивностью курения.

б) В разделе 12.1 дана оценка с поправками $P_L=0{,}51$ доли много курящих у больных раком легких. Поправки вычислялись по данным опросов и анализов у 50 пациенток. Допустим, что аналогичным образом для 50 опрошенных больных раком молочной железы проведено еще и химическое тестирование, давшее следующие результаты:

Опрос	Тест		Всего
	Много курит	Мало курит	
Много курит	18	0	18
Мало курит	2	30	32

Вычислите значения поправок n_{00}/n_0 и n_{10}/n_1. Вычислите с их помощью оценку

$$P_B=\frac{n_{00}}{n_0}p_B+\frac{n_{10}}{n_1}(1-p_B).$$

Каково значение отношения шансов, соответствующего оценкам долей с поправками? Сильнее или слабее связь между видом рака и курением по отношению шансов по сравнению с а)?

в) Теперь допустим, что результаты классификации подвыборки из 50 пациенток с раком молочной железы таковы:

Опрос	Тест		Всего
	Много курит	Мало курит	
Много курит	16	2	18
Мало курит	7	25	32

Вычислите значения поправок n_{00}/n_0 и n_{10}/n_1. Вычислите с их помощью новое значение P_B, а также соответствующее значение отношения шансов. Сильнее или слабее связь по этому значению, чем с а)?

12.2. Допустим, что значения поправок n_{00}/n_0 и n_{10}/n_1 для больных с одним заболеванием такие же, как и для больных с другим заболеванием. Покажите, как можно просто выразить P_L-P_B через p_L-p_B и разность $n_{00}/n_0-n_{10}/n_1$.

ЛИТЕРАТУРА

Barron, B. A. (1977). The effects of misclassification on the estimation of relative risk. *Biometrics*, **33**, 414–418.

Bryson, M. R. (1965). Errors of classification in a binomial population. *J. Am. Stat. Assoc.*, **60**, 217–224.

Chiacchierini, R. P. and Arnold, J. C. (1977). A two-sample test for independence in 2 × 2 contingency tables with both margins subject to misclassification. *J. Am. Stat. Assoc.*, **72**, 170–174.

Deming, W. E. (1977). An essay on screening, or on two-phase sampling, applied to surveys of a community. *Int. Stat. Rev.*, **45**, 29–37.

Densen, P. M., Davidow, B., Bass, H. E. and Jones, E. W. (1967). A chemical test for smoking exposure. *Arch. Environ. Health*, **14**, 865–874.

Ganz, V. H., Gurland, B. J., Deming, W. E., and Fisher, B. (1972). The study of the psychiatric symptoms of systemic lupus erythematosus: A biometric study. *Psychosom. Med.*, **34**, 199–206.

Harper, D. (1964). Misclassification in epidemiological surveys. *Am. J. Public Health*, **54**, 1882–1886.

Hochberg, Y. (1977). On the use of double sampling schemes in analyzing categorical data with misclassification errors. *J. Am. Stat. Assoc.*, **72**, 914–921.

Levin, M. L. (1954). Etiology of lung cancer: Present status. *N. Y. State J. Med.*, **54**, 769–777.

Medical Research Council (1966). Questionnaire on respiratory diseases. Dawlish, Devon, England: W. J. Holman, Ltd.

Press, S. J. (1968). Estimating from misclassified data. *J. Am. Stat. Assoc.*, **63**, 123–133.

Rubin, T., Rosenbaum, J., and Cobb, S. (1956), The use of interview data for the detection of association in field studies. *J. Chronic Dis.*, **4**, 253–266.

Spitzer, R. L., Endicott, J. Fleiss, J. L., and Cohen, J. (1970). Psychiatric Status Schedule: A technique for evaluating psychopathology and impairment in role functioning. *Arch. Gen. Psychoatry*, **23**, 41–55.

Spitzer, R. L., Fleiss, J. L., Endicott, J., and Cohen, J. (1967). Mental Status Schedule: Properties of factor analytically derived scales. *Arch. Gen. Psychiatry*, **16**, 479–493.

Tenenbein, A. (1970). A double sampling scheme for estimating from binomial data with misclassifications. *J. Am. Stat. Assoc.*, **65**, 1350–1361.

Tenenbein, A. (1971). A double sampling scheme for estimating from binomial data with misclassifications: Sample size determination. *Biometrics*, **27**, 935–944.

Wing, J. K., Birley, J. L. T., Cooper, J. E., Graham, P., and Isaacs, A. D. (1967). Reliability of a procedure for measuring and classifying "present psychiatric state." *Br. J. Psychiatry*, **113**, 499–515.

Глава 13

Определение степени согласованности экспертов

Статистические методы контроля ошибок, описанные в предыдущей главе, пригодны только в тех случаях, когда вероятности ошибок при классификации либо известны, благодаря каким-то априорным сведениям, либо поддаются оцениванию с помощью точной классификации подвыборки из исследуемой группы. Однако нередко факторы, являющиеся предметом исследования, таковы, что получить значения вероятности ошибочной классификации далеко не просто.

Чтобы все же оценить, до какой степени можно полагаться на результаты классификации объектов по данному фактору, естественно использовать множество объектов, каждый из которых классифицирован более одного раза, например, несколькими экспертами. Правда, при этом степень согласованности экспертов дает не более чем верхнюю границу точности классификации. Если согласованность экспертов высока, то возможно (но ни в коем случае не обязательно), что проводимая классификация отражает истинное состояние дел. Если же, напротив, согласованность экспертов низка, то анализ имеет мало смысла: бесполезно искать зависимость между классифицирующими и другими факторами, когда нельзя доверять даже результатам классификации по факторам, на основе которых строится анализ.

Эта глава посвящена измерению степени согласованности экспертов, проводящих классификацию на категоризованной шкале. В разд. 13.1 рассматривается случай, когда объект классифицируется дважды (двумя экспертами), в разд. 13.2 — когда объект классифицируется многими экспертами. Приложение полученных результатов в других задачах обсуждается в разд. 13.3.

13.1. Случай двух экспертов

Допустим, что два эксперта (диагноста) независимо классифицируют каждый из n объектов выборки. Шкала, на которой проводится классификация, образована k категориями. Рас-

смотрим гипотетический пример — табл. 13.1, в которой каждый элемент (кроме сумм) обозначает пропорцию объектов, отнесенных диагностом A к одной из $k=3$ категорий, и диагностом B — к одной из тех же категорий. Например, 5% всех объектов были отнесены диагностом A к категории неврозов и одновременно диагностом B — к категории психозов.

Таблица 13.1

Диагнозы двух экспертов по $n=100$ больным

Диагност A	Диагност B			
	Психозы	Неврозы	Органозы	Сумма
Психозы	0,75	0,01	0,04	0,80
Неврозы	0,05	0,04	0,01	0,10
Органозы	0	0	0,10	0,10
Сумма	0,80	0,05	0,15	1,00

Предположим, нам хочется определить степень согласованности диагностов как отдельно по каждой категории, так и в целом. Начнем анализ со сжатия исходной таблицы $k\times k$ в таблицу 2×2, в которой все категории, кроме изучаемой, объединены в одну категорию «все остальные». В табл. 13.2 представлены результаты сжатия в общем виде и конкретно — для неврозов по данным табл. 13.1. Нужно иметь в виду, что элементы a, b, c и d в таблице общего вида означают *пропорции* объектов, а не их число.

Простейшим и наиболее часто используемым индексом согласованности является суммарная пропорция согласия

$$p_o=a+d \qquad (13.1)$$

или ее вариант в виде $2p_o-1$. Статистика (13.1) была предложена в качестве индекса согласованности Холли и Гилфордом [Holley and Guilford, 1964] и Максвеллом [Maxwell, 1977]. Для неврозов суммарная пропорция согласия

$$p_o=0{,}04+0{,}89=0{,}93.$$

Это значение наряду с суммарными пропорциями согласия для двух других категорий приведено в первой колонке табл. 13.3. Судя по этим значениям, можно считать, что согласованность одинаково хороша для всех трех категорий, и согласованность по органическим расстройствам немного лучше согласованно-

Таблица 13.2

Таблица для определения согласованности по одной категории

Общий вид				По неврозам			
Диагност *A*	Диагност *B*			Диагност *A*	Диагност *B*		
	Данная категория	Все остальные	Сумма		Неврозы	Органозы и психозы	Сумма
Данная категория	*a*	*b*	p_1	Неврозы	0,04	0,06	0,10
Все остальные	*c*	*d*	q_1	Органозы и психозы	0,01	0,89	0,90
Сумма	p_2	q_2	1	Сумма	0,05	0,95	1,00

сти по неврозам, которая в свою очередь немного лучше согласованности по психозам.

Таблица 13.3

Значения индексов согласованности для данных из табл. 13.1

Категория	p_o	p_s	λ_r	p_s'	A	$\varkappa$
Психозы	0,90	0,94	0,88	0,75	0,84	0,69
Неврозы	0,93	0,53	0,06	0,96	0,75	0,50
Органозы	0,95	0,80	0,60	0,97	0,89	0,77

Допустим, что изучаемая категория относительно редкая. Тогда пропорция *d*, представляющая согласованность по отсутствию заболевания этой категории, скорее всего, будет велика и «издует» значение p_o. Было предложено множество индексов, которые основаны только на пропорциях *a*, *b* и *c*. Из всех этих индексов, пожалуй, только так называемая «пропорция частного согласия» (proportion of specific agreement)

$$p_s = \frac{2a}{2a+b+c} = \frac{a}{\bar{p}}, \qquad (13.2)$$

где $\bar{p}=(p_1+p_2)/2$, имеет здравую вероятностную интерпретацию. Случайно выберем одного из экспертов и сфокусируем внимание на выделенной категории. Тогда величина p_s есть условная вероятность того, что второй эксперт отнесет объект

к этой категории при условии, что выбранный эксперт также отнес данный объект к этой категории. Впервые этот индекс как меру сходства предложил Дайс [Dice, 1945].

Пропорция частного согласия по неврозам равна:

$$p_s = \frac{2 \cdot 0{,}04}{2 \cdot 0{,}04 + 0{,}06 + 0{,}01} = 0{,}53.$$

Значения p_s для всех трех категорий даны во второй колонке табл. 13.3. Выводы, основанные на p_o и на p_s, заметно отличаются. Теперь лучшей выглядит согласованность по психозам, значительно хуже нее согласованность по органозам и намного хуже — по неврозам.

Определим $\bar{q} = 1 - \bar{p}$, или

$$\bar{q} = \frac{1}{2}(q_1 + q_2) = d + \frac{b+c}{2}, \qquad (13.3)$$

и предположим, что $\bar{q} > \bar{p}$. Гудмен и Краскел [Goodman and Kruskal, 1954] предложили в качестве индекса согласованности

$$\lambda_r = \frac{(a+d) - \bar{q}}{1 - \bar{q}} = \frac{2a - (b+c)}{2a + (b+c)}, \qquad (13.4)$$

мотивируя свое предложение не столько с помощью понятия согласованности, сколько в терминах двух частот правильного предсказания категории объекта; когда предсказания даются при известном и при неизвестном результате одновременной классификации объекта двумя экспертами. Индекс λ_r принимает свое максимальное значение $+1$, когда имеется полная согласованность, и свое минимальное значение -1, когда $a = 0$, независимо от значения d (но не при $a + d = 0$, как считали Гудмен и Краскел).

Для неврозов

$$\lambda_r = \frac{2 \cdot 0{,}04 - (0{,}06 + 0{,}01)}{2 \cdot 0{,}04 + (0{,}06 + 0{,}01)} = 0{,}06.$$

Значения λ_r для нашего примера даны в третьей колонке табл. 13.3. Ввиду тождества

$$\lambda_r = 2p_s - 1 \qquad (13.5)$$

порядок значений обоих индексов для трех категорий одинаков.

В выражение для пропорции частного согласия не входит d. Если, наоборот, мы решим исключить a, то получим такой индекс:

$$p_s' = \frac{d}{\bar{q}} = \frac{2d}{2d + b + c}. \qquad (13.6)$$

Для неврозов

$$p_s' = \frac{2 \cdot 0{,}89}{2 \cdot 0{,}89 + 0{,}06 + 0{,}01} = 0{,}96.$$

Все и два других значения приведены в четвертой колонке. Из этих значений складывается еще одна картина, не похожая на предыдущие. Согласованность по органическим расстройствам (по отношению к их отсутствию) и по неврозам выглядит довольно хорошо и явно лучше согласованности по психозам.

Вместо того чтобы делать выбор между p_s и p_s', Рогот и Голдберг [Rogot and Goldberg, 1966] предложили использовать как метод согласованности просто их среднее

$$A = \frac{1}{2}(p_s + p_s') = \frac{a}{p_1 + p_2} + \frac{d}{q_1 + q_2}. \qquad (13.7)$$

Для неврозов

$$A = \frac{0{,}04}{0{,}10 + 0{,}05} + \frac{0{,}89}{0{,}90 + 0{,}95} = 0{,}75.$$

Как видно из соответствующей колонки табл. 13.3, индекс A упорядочивает три категории еще одним способом: согласованность по органическим расстройствам лучше всего, хуже согласованность по психозам и еще хуже — по неврозам.

Предполагались и другие индексы согласованности двух экспертов (например, [Fleiss, 1965, Armitage et al., 1966, Rogot and Goldberg, 1966, Bennett, 1972]), однако и так уже ясно, что для измерения степени согласованности экспертов совершенно недостаточно лишь полагаться на какой-либо произвольный индекс согласованности.

Мы сможем по-другому взглянуть на проблему, если задумаемся над тем, что, за исключением крайних случаев ($p_1 = 0$ либо $p_2 = q_1 = 0$), можно ожидать некоторой степени согласованности экспертов лишь за счет случайности (см. табл. 13.1). Например, если эксперт A использует одни критерии при классификации объекта по категориям некоторого фактора, а эксперт B — совершенно другие критерии, не зависящие от критериев первого эксперта, то вся наблюдаемая согласованность будет обусловлена случайными совпадениями.

По поводу необходимости учитывать согласованность, определяемую случайными совпадениями, при оценивании достоверности классификации высказывались различные мнения. Например, Рогот и Голдберг [Rogot and Goldberg, 1966] настаивают на сопоставлении наблюдаемой и ожидаемой согласованности, когда надо провести сравнение различных пар экспертов или различных типов объектов. Гудмен и Краскел [Goodman and Kruskal, 1954, p. 758], напротив, утверждают, что случай-

Таблица 13.4

Пропорции, ожидаемые при независимой классификации двумя экспертами
(для данных табл. 13.2)

Общий вид				Для неврозов			
Эксперт A	Эксперт B			Эксперт A	Эксперт B		
	Данная категория	Все остальные	Сумма		Неврозы	Органозы и психозы	Сумма
Данная категория	p_1p_2	p_1q_2	p_1	Неврозы	0,005	0,095	0,10
Все остальные	q_1p_2	q_1q_2	q_1	Органозы и психозы	0,045	0,855	0,90
Сумма	p_2	q_2	1	Сумма	0,05	0,95	1

ная согласованность не должна вызывать много забот, поскольку обычно можно считать, что наблюдаемая степень согласованности больше случайной (даже полагаясь на это предположение, следует проверять, является ли разница незначительной или существенной).

Эрмитедж и др. [Armitage et al., 1966, p. 102] имеют точку зрения, промежуточную по отношению к этим двум мнениям. Они принимают необходимость всегда учитывать согласованность, ожидаемую за счет случайных совпадений, когда сравниваются различные наборы данных, но утверждают, что способы учета случайности при измерении степени согласованности определяются слишком нечетко.

Тем не менее существует естественный способ учета случайности в расчетах. Рассмотрим любой индекс, принимающий значение 1 при полной согласованности. Пусть I_o обозначает наблюдаемое значение индекса (т. е. вычисленное по пропорциям табл. 13.2), а I_e — значение, ожидаемое при чисто случайных совпадениях (вычисленное по пропорциям табл. 13.4).

Прирост наблюдаемой согласованности по сравнению со случайной равен $I_o - I_e$, тогда как максимально возможный прирост есть $1 - I_e$. Отношения двух этих разностей есть «каппа»:

$$\hat{\varkappa} = \frac{I_o - I_e}{1 - I_e}. \qquad (13.8)$$

Коэффициент $\hat{\varkappa}$ — мера согласованности с требуемыми свойствами. При полной согласованности $\hat{\varkappa} = +1$. Если наблюдаемая

согласованность больше или равна случайной согласованности, $\varkappa \geq 0$ и в противном случае $-\varkappa < 0$. Минимальное значение $\varkappa$ зависит от маргинальных пропорций: если $I_p = 0{,}5$, то минимальное значение равно -1, а иначе минимум лежит между -1 и 0.

С помощью простых алгебраических выкладок можно проверить, что значение $\hat{\varkappa}$, вычисляемое в (13.8), *одинаково для всех индексов согласованности, определенных выше*, и равно:

$$\hat{\varkappa} = \frac{2\,(ad-bc)}{p_1q_2+p_2q_1}. \qquad (13.9)$$

Итак, вводя поправку на согласованность, ожидаемую за счет случайных совпадений, можно достичь важной цели — унификации различных подходов к определению степени согласованности.

Для неврозов

$$\hat{\varkappa} = \frac{2\,(0{,}01\cdot 0{,}89 - 0{,}06\cdot 0{,}01)}{0{,}10\cdot 0{,}95 + 0{,}05\cdot 0{,}90} = 0{,}50.$$

Это значение и значения $\varkappa$ для двух других заболеваний приведены в последней колонке табл. 13.3. Они близки к значениям для достоверности психиатрических диагнозов, полученных Спитцером и Флейсом [Spitzer and Fleiss, 1974] из научной литературы. Согласованность лучше всего по органическим расстройствам, менее хороша по психозам и хуже всего по неврозам.

Статистика (13.8) была впервые предложена Коэном (Cohen, 1960]. Ее варианты имеются в [Scott, 1955] и [Maxwell and Pilliner, 1968]. Все они интерпретировались как коэффициент внутриклассовой корреляции (см. [Ebel, 1951]), широко используемый в качестве меры согласованности решений различных экспертов в случае классификации по количественной шкале. Как показали Флейс [Fleiss, 1975] и Криппендорф [Krippendorff, 1970], только каппа идентична (с точностью до члена, содержащего $1/n$, где n — число объектов) варианту коэффициента внутриклассовой корреляции Бартко [Bartko, 1966], в котором в качестве источника нежелательного расхождения рассматривается различие между экспертами по результатам классификации к данной категории (т. е. разность p_1 и p_2).

Лэндис и Кох [Landis and Koch, 1977a] ориентировочно охарактеризовали интервалы значений каппа по соответствующей им степени согласованности. Во многих приложениях значения каппа, большие приблизительно 0,75, соответствуют прекрасной согласованности по сравнению со случайной, значения, меньшие примерно 0,40, указывают на слабый прирост согласо-

ванности по сравнению со случайной, а значения между 0,40 и 0,75 относятся к достаточно хорошей согласованности по сравнению со случайной.

Часто нужна общая мера согласованности по совокупности категорий. Можно ввести общее значение каппа, определив его как взвешенное среднее значений каппа для отдельных категорий, с весами, равными знаменателям в выражении для значений каппа (т. е. величинами $p_1q_2+p_2q_1$). Эквивалентное и более наглядное выражение можно получить, располагая данные в виде табл. 13.5.

Таблица 13.5

Пропорции совместной классификации двух экспертов при k категориях

Эксперт A	Эксперт B				
	1	2	...	k	Сумма
1	p_{11}	p_{12}	...	p_{1k}	$p_{1.}$
2	p_{21}	p_{22}	...	p_{2k}	$p_{2.}$
⋮					
k	p_{k1}	p_{k2}	...	p_{kk}	$p_{k.}$
Сумма	$p_{.1}$	$p_{.2}$	...	$p_{.k}$	1

Суммарная пропорция наблюдаемой согласованности есть

$$p_o = \sum_{i=1}^{k} p_{ii}, \tag{13.10}$$

суммарная пропорция случайной согласованности —

$$p_e = \sum_{i=1}^{k} p_{i.}p_{.i}. \tag{13.11}$$

Тогда общее значение каппа есть

$$\hat{\varkappa} = \frac{p_o - p_e}{1 - p_e}. \tag{13.12}$$

Для данных табл. 13.1

$$p_o = 0{,}75 + 0{,}4 + 0{,}10 = 0{,}89,$$

$$p_e = 0{,}80 \cdot 0{,}80 + 0{,}10 \cdot 0{,}05 + 0{,}10 \cdot 0{,}15 = 0{,}66$$

и

$$\hat{\varkappa} = \frac{0{,}89 - 0{,}66}{1 - 0{,}66} = 0{,}68.$$

Для проверки гипотезы о независимости классификаций различных экспертов (это означает, что истинная каппа равна нулю) подходит, как показали Флейс и др. [Fleiss et al., 1969], следующая оценка стандартной ошибки каппа:

$$\text{s.e.}_0(\hat{\kappa}) = \frac{1}{(1-p_e)\sqrt{n}} \sqrt{p_e + p_e^2 - \sum_{i=1}^{k} p_{i.} p_{.i} (p_{i.} + p_{.i})}, \quad (13.13)$$

где p_e определена в (13.11). Гипотеза проверяется против альтернативы, что согласованность лучше случайной, с помощью статистики[1]

$$z = \frac{\hat{\kappa}}{\text{s.e.}_0 \hat{\kappa}}, \quad (13.14)$$

сравниваемой со значениями стандартного нормального распределения. Гипотеза отвергается, если z достаточно велика (здесь предпочтительней не двухсторонний, а односторонний критерий).

Для текущих данных

$$\text{s.e.}_0(\hat{\kappa}) = \frac{1}{(1-0{,}66)\sqrt{100}} \sqrt{0{,}66 + 0{,}66^2 - 1{,}0285} = 0{,}076$$

и

$$z = \frac{0{,}68}{0{,}076} = 8{,}95.$$

Значит, общее значение каппа статистически высоко значимо, и эта величина ($\hat{\kappa} = 0{,}68$) соответствует хорошей согласованности.

Формулы (13.10)—(13.14) применимы и при числе категорий, равном двум[2]. Следовательно, их можно использовать при изучении достоверности классификации для каждой отдельной категории, как показано в табл. 13.6 на данных из табл. 13.1.

---

[1] Данную статистику можно применять для проверки гипотезы о независимости двух факторов (а не только экспертов) в таблице сопряженности $k \times k$, правда, при $k > 2$ — против более узкого класса альтернатив по сравнению с критерием хи-квадрат для двумерной таблицы сопряженности и остальными критериями (относительно случая $k = 2$ см. задачу 13.2). Естественно, $\hat{\kappa}$ в (13.12) можно интерпретировать как показатель степени связи, принимающий значения в интервале $[\gamma, 1]$, где $-1 \leqslant \gamma < 0$. — *Примеч. пер.*

[2] В частности, при $k = 2$ выражения (13.12) и (13.9) совпадают — *Примеч. пер.*

Таблица 13.6

Значения каппа для отдельных категорий и для совокупности категорий по табл. 13.1

Категория	p_o	p_e	$\hat{\kappa}$	s.e.$_\theta\,\hat{\kappa}$	z
Психозы	0,90	0,68	0,69	0,100	6,90
Неврозы	0,93	0,86	0,50	0,093	5,38
Органозы	0,95	0,78	0,77	0,097	7,94
По совокупности	0,89	0,66	0,68	0,076	8,95

Заметим, что общее значение каппа равняется сумме отдельных разностей $p_o - p_e$ (т. е. числителей в (13.12)), деленной на сумму отдельных разностей $1-p_e$ (т. е. знаменателей в (13.12)), например:

$$\hat{\kappa} = \frac{(0{,}90-0{,}68)+(0{,}93-0{,}86)+(0{,}95-0{,}78)}{(1-0{,}68)+(1-0{,}86)+(1-0{,}78)} = 0{,}68,$$

что подтверждает, что $\hat{\kappa}$ — это взвешенное среднее отдельных $\hat{\kappa}$.

Флейс, Коэн и Эверитт [Fleiss et al., 1969] показали, что при проверке гипотезы о равенстве истинной каппа (для одной категории или общей) некоторому значению κ, *отличному от нуля*, можно использовать следующую оценку стандартной ошибки $\hat{\kappa}$:

$$\text{s.e.}(\hat{\kappa}) = \frac{\sqrt{A+B-C}}{(1-p_e)\sqrt{n}}, \tag{13.15}$$

где

$$A = \sum_{i=1}^{k} p_{ii}\,[1-(p_{i.}+p_{.i})(1-\hat{\kappa})]^2, \tag{13.16}$$

$$B = (1-\hat{\kappa})^2 \sum_{i\neq j}\sum p_{ij}(p_{.i}+p_{j.})^2, \tag{13.17}$$

$$C = [\hat{\kappa} - p_e(1-\hat{\kappa})]^2. \tag{13.18}$$

Гипотеза о том, что истинная каппа равна κ, отвергается, если отношение

$$z = \frac{|\hat{\kappa}-\kappa|}{\text{s.e.}(\hat{\kappa})} \tag{13.19}$$

вполне превышает критическое значение для нормального распределения. Приближенный $100(1-\alpha)\%$-ный доверительный интервал для $\varkappa$ есть

$$\widehat{\varkappa}-c_{\alpha/2}\,\text{s.e.}\,(\widehat{\varkappa}) \leqslant \varkappa \leqslant \widehat{\varkappa}+c_{\alpha/2}\,\text{s.e.}\,(\widehat{\varkappa}). \qquad (13.20)$$

Проверим по данным табл. 13.1 гипотезу о том, что истинное общее значение каппа равно 0,80. Величины (13.16)—(13.18), входящие в (13.15), принимают значения

$$A=0{,}75[1-(0{,}80+0{,}80)\cdot(1-0{,}68)]^2+0{,}04[1-(0{,}10+0{,}05)(1-0{,}68)]^2+0{,}10[1-(0{,}10+0{,}15)(1-0{,}68)]^2=0{,}2995;$$

$$B=(1-0{,}68)^2[0{,}01\cdot(0{,}80+0{,}10)^2+0{,}04\cdot(0{,}80+0{,}10)^2+0{,}05\cdot(0{,}05+0{,}80)^2+0{,}01\cdot(0{,}05+0{,}10)^2+0\cdot(0{,}15+0{,}80)^2+0\cdot(0{,}15+0{,}10)^2]=0{,}0079;$$

$$C=(0{,}68-0{,}66(1-0{,}68))^2=0{,}2198.$$

В итоге

$$\text{s.e.}\,(\widehat{\bar{\varkappa}})=\frac{\sqrt{0{,}2995+0{,}0079-0{,}2198}}{(1-0{,}66)\sqrt{100}}=0{,}087$$

и

$$z=\frac{|0{,}68-0{,}80|}{0{,}087}=1{,}38,$$

так что гипотеза «$\bar{\varkappa}=0{,}80$» не отвергается.

Допустим теперь, что мы хотим сопоставить и объединить $g\geqslant 2$ независимых оценок каппа. Здесь применимы методы анализа из разд. 10.1. Пусть $V_m(\widehat{\varkappa}_m)$ обозначает квадрат стандартной ошибки m-й оценки $\widehat{\varkappa}_m$, т. е. квадрат (13.15). Объединенная оценка значения $\varkappa$, предположительно одинакового во всех g группах, есть

$$\widehat{\varkappa}_{\text{overall}}=\frac{\sum_{m=1}^{g}\dfrac{\widehat{\varkappa}_m}{V_m(\widehat{\varkappa}_m)}}{\sum_{m=1}^{g}\dfrac{1}{V_m(\widehat{\varkappa}_m)}} \qquad (13.21)$$

Чтобы проверить гипотезу о равенстве g истинных значений каппа, можно соотнести величину

$$\chi^2_{\text{equal}}=\sum_{m=1}^{g}\frac{(\widehat{\varkappa}_m-\widehat{\varkappa}_{\text{overall}})^2}{V_m(\widehat{\varkappa}_m)} \qquad (13.22)$$

с критическим значением распределения хи-квадрат с $g-1$ степенями свободы. Гипотеза отвергается при значимо большой величине (13.22). Границы приближенного 100 $(1-\alpha)$%-ного доверительного интервала для одинакового по предположению истинного значения $\varkappa$ определяется как

$$\widehat{\varkappa}_{\text{overall}} \pm c_{\alpha/2} \sqrt{\frac{1}{\sum_{m=1}^{g} \frac{1}{V_m(\widehat{\varkappa}_m)}}}. \qquad (13.23)$$

Коэн обобщил в [Cohen, 1968] (см. также [Spitzer et al., 1967]) предложенную им меру согласованности экспертов — каппа — на случай, когда можно количественно охарактеризовать относительную важность любой возможной рассогласованности. Пусть на основе теоретических или клинических результатов независимо от анализируемых данных всем k^2 парам категорий назначены веса согласованности w_{ij} $(i=1, \ldots, k;\ j=1, \ldots, k)$ (см. [Cicchetti, 1976]). Веса ограничены интервалом $0 \leqslant w_{ij} \leqslant 1$, причем

$$w_{ii} = 1 \qquad (13.24)$$

(т. е. полной согласованности соответствует максимальный вес),

$$0 \leqslant w_{ij} < 1 \quad \text{при } i \neq j \qquad (13.25)$$

(т. е. любая несогласованность имеет вес, меньший максимального) и

$$w_{ij} = w_{ji} \qquad (13.26)$$

(т. е. пара экспертов считается симметричной).

Взвешенная пропорция наблюдаемой согласованности есть

$$p_{o(w)} = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij}\, p_{ij}, \qquad (13.27)$$

где расположение пропорций имеет вид табл. 13.5, взвешенная пропорция случайной согласованности —

$$p_{e(w)} = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij}\, p_{i.}\, p_{.j}. \qquad (13.28)$$

Тогда взвешенная каппа определяется выражением

$$\widehat{\varkappa}_w = \frac{p_{o(w)} - p_{e(w)}}{1 - p_{e(w)}}. \qquad (13.29)$$

Если $w_{ij} = 0$ для всех $i \neq j$ (т. е. всякая рассогласованность одинаково серьезна), то взвешенная каппа будет совпадать с общей каппа (13.12).

Величину взвешенной каппа можно интерпретировать аналогично невзвешенной каппа: $\hat{\varkappa}_w \geqslant 0{,}75$ означает в большинстве случаев прекрасную согласованность, $\hat{\varkappa}_w \leqslant 0{,}40$ — плохую согласованность.

Предположим, что k категорий упорядочены. Проведем двухфакторный дисперсионный анализ данных, которые будут представлять результаты классификации, если взять номера $1, \ldots, k$ как количественные значения. Бартко [Bartko, 1966] выводит для этого случая формулу внутриклассового коэффициента корреляции. Как показали Флейс и Коэн [Fleiss and Cohen, 1973], коэффициент внутриклассовой корреляции совпадает с точностью до члена, зависящего от $1/n$, с взвешенной каппа, когда весами являются

$$w_{ij} = 1 - \frac{(i-j)^2}{(k-1)^2}. \tag{13.30}$$

Чиккетти и Элиссон [Cicchetti and Alison, 1971] независимо от Коэна [Cohen, 1968] предложили статистику для определения согласованности экспертов, которая формально является взвешенной каппа. Соответствующие веса —

$$w_{ij} = 1 - \frac{|i-j|}{k-1}. \tag{13.31}$$

Выборочное распределение взвешенной каппа выведено в [Fleiss, Cohen and Everitt, 1969]. Этот результат подтверждается в работах [Cicchetti and Fleiss, 1977, Landis and Koch, 1977a, Fleiss and Cicchetti, 1978, Hubert, 1978]. Для проверки гипотезы о нулевом истинном значении взвешенной каппа подходит следующая оценка стандартной ошибки $\varkappa_w$,

$$\text{s.e.}_0(\varkappa_w) = \frac{1}{(1-p_{e(w)})\sqrt{n}} \sqrt{\sum_{i=1}^{k}\sum_{j=1}^{k} p_{i.}p_{.j}[w_{ij} - (\bar{w}_{i.} + \bar{w}_{.j})]^2 - p_{e(w)}^2}, \tag{13.32}$$

где

$$\bar{w}_{i.} = \sum_{j=1}^{k} p_{.j} w_{ij}; \tag{13.33}$$

$$\bar{w}_{.j} = \sum_{i=1}^{k} p_{i.} w_{ij}. \tag{13.34}$$

Чтобы проверить гипотезу, значение

$$z = \frac{\hat{\varkappa}_w}{\text{s.e.}_0(\hat{\varkappa}_w)}. \tag{13.35}$$

сравнивают с критическим значением стандартного нормального распределения.

Для проверки гипотезы о равенстве истинного значения взвешенной каппа некоторому *ненулевому* значению $\varkappa_w$ допустимо использовать оценку стандартной ошибки $\widehat{\varkappa}_w$:

$$\text{s.e.}(\widehat{\varkappa}_w)=\frac{1}{(1-p_{e(w)})\sqrt{n}}\times$$

$$\times\sqrt{\sum_{i=1}^{k}\sum_{j=1}^{k}p_{ij}[w_{ij}-(\bar{w}_{i\cdot}+\bar{w}_{\cdot j})(1-\widehat{\varkappa}_w)]^2-[\widehat{\varkappa}_w-p_{e(w)}(1-\widehat{\varkappa}_w)]^2}. \tag{13.36}$$

Гипотеза отвергается, если значение отношения

$$z=\frac{|\widehat{\varkappa}_w-\varkappa_w|}{\text{s.e.}(\widehat{\varkappa}_w)} \tag{13.37}$$

больше критического значения для стандартного нормального распределения.

Можно показать (см. задачу 13.4), что выражения (13.13) и (13.15) для стандартных ошибок невзвешенных каппа — частные случаи выражений (13.32) и (13.36) для стандартных ошибок взвешенных каппа, когда $w_{ii}=1$ для всех i и $w_{ij}=0$ для всех $i\neq j$.

В некоторых работах ([Light, 1971, Landis and Koch, 1977a]) были предприняты попытки обобщить каппа на случай, когда каждый эксперт из неизменного множества экспертов классифицирует все объекты. Однако для этой задачи получено пока мало результатов, которые к тому же нельзя описать в виде, пригодном для общего применения. В следующем разделе мы рассмотрим сходную, но более простую задачу, в которой различные объекты классифицируются различными экспертами.

13.2. Случай нескольких экспертов

Предположим, проведено исследование n объектов, в котором i-ый объект классифицировали m_i экспертов. Будем считать, что состав группы экспертов может меняться от объекта к объекту. Сначала будем предполагать, что $k=2$, т. е. что классификация состоит в отнесении объекта к одной из двух категорий. Случай $k>2$ будет рассматриваться в этом разделе позже. Обозначим x_i число экспертов, отнесших i-й объект к первой, «положительной», произвольно выбранной категории. Значит, ко второй («отрицательной») категории i-й объект отнесен m_i-x_i раз.

Каппа-статистику мы выведем с помощью тождеств, связывающих коэффициент внутриклассовой корреляции и каппа. Начнем с однофакторного дисперсионного анализа данных, которые мы получим, присвоив «положительным» классификациям значение 1, а «отрицательным» — 0. Именно таким путем шли Лэндис и Кох [Landis and Koch, 1977b] (за исключением того, что число степеней свободы у них было не n, как взято ниже, а $n-1$).

Определим общую пропорцию положительных классификаций

$$p = \frac{\sum_{i=1}^{n} x_i}{n\bar{m}}, \tag{13.38}$$

где

$$\bar{m} = \frac{\sum_{i=1}^{n} m_i}{n} \tag{13.39}$$

— среднее число экспертов на объект. Если число объектов велико (скажем, $n \geqslant 20$), то средний межобъектный квадрат (mean square between subjects) приближенно равен:

$$\mathrm{BMS} = 1/n \sum_{i=1}^{n} \frac{(x_i - m_i\,\bar{p})^2}{m_i}, \tag{13.40}$$

а средний внутриобъектный квадрат (mean square within subjects) равен:

$$\mathrm{WMS} = \frac{1}{n\,(\bar{m}-1)} \sum_{i=1}^{n} \frac{x_i\,(m_i - x_i)}{m_i}. \tag{13.41}$$

Формально оценкой внутриклассового коэффициента корреляции является [1]

$$r = \frac{\mathrm{BMS} - \mathrm{WMS}}{\mathrm{BMS} + (m_0 - 1)\,\mathrm{WMS}}, \tag{13.42}$$

---

[1] В этом разделе используется модель со случайными эффектами одномерного однофакторного дисперсионного анализа, в которой предполагается, что $y_{ij} = \mu + a_i + \varepsilon_{ij}$, $i = 1, \ldots, n$, $j = 1, \ldots, m_i$, где y_{ij} — наблюдаемая случайная величина; μ — генеральное среднее; a_{il} и ε_{ij} — некоррелированные случайные величины; $E(a_i) = 0$, $\mathrm{Var}\,(a_i) = \sigma_a^2$, $E(\varepsilon_{ij}) = 0$, $\mathrm{Var}\,(\varepsilon_{ij}) = \sigma_e^2$. Коэффициент внутриклассовой корреляции равен по определению $\rho = \sigma_a^2/(\sigma_a^2 + \sigma_e^2)$ и, конечно, неотрицателен. В данном случае это определение используется формально, так что оценки (13.42) и (13.44), как показано далее, могут принимать отрицательные значения. — *Примеч. пер.*

где

$$m_0 = \bar{m} - \frac{\sum_{i=1}^{n} (m_i - \bar{m})^2}{n(n-1)\,\bar{m}}. \qquad (13.43)$$

Если n очень велико, значения m_0 и $\bar{m}$ будут близки. Заменив m_0 в (13.42) на $\bar{m}$, получим окончательное выражение для коэффициента внутриклассовой корреляции, т. е. для каппа:

$$\hat{\varkappa} = \frac{\text{BMS} - \text{WMS}}{\text{BMS} + (\bar{m}-1)\,\text{WMS}} = 1 - \frac{\sum_{i=1}^{n} \frac{x_i(m_i - x_i)}{m_i}}{n(\bar{m}-1)\,\bar{p}\,\bar{q}}, \qquad (13.44)$$

где $\bar{q} = 1 - \bar{p}$.

Статистика $\hat{\varkappa}$ имеет следующие свойства. Если пропорция положительных классификаций постоянна от объекта к объекту, т. е. $x_i/m_i = \bar{p}$ для всех i, и $\bar{p}$ не равняется 0 или 1, то несогласованность имеется внутри объектов, но не между объектами. Легко видеть, что в этом случае $\varkappa$ принимает свое минимальное значение $-1/(\bar{m}-1)$.

Если пропорции x_i/m_i ведут себя в точности как биномиальные пропорции с параметрами m_i и общей вероятностью p, то между объектами сходства столько же, сколько внутри объектов. В этом случае $\hat{\varkappa}$ равна нулю.

Если любая пропорция x_i/m_i принимает одно из двух значений 0 или 1, то имеется идеальная согласованность внутри объектов. В этом случае $\hat{\varkappa}$ принимает значение 1.

Рассмотрим гипотетические данные для $n = 25$ объектов, приведенные в табл. 13.7. Для этих данных среднее число экспертов на объект равно:

$$\bar{m} = \frac{81}{25} = 3{,}24,$$

суммарная пропорция положительных классификаций —

$$\bar{p} = \frac{46}{25 \cdot 3.24} = 0{,}568,$$

а значение суммы из (13.40) равно:

$$\sum_{i=1}^{25} \frac{x_i(m_i - x_i)}{m_i} = 6{,}30.$$

Каппа (13.44) равняется:

$$\hat{\varkappa} = 1 - \frac{6{,}30}{25\,(3{,}24-1)\cdot 0{,}568 \cdot 0{,}432} = 0{,}54,$$

что указывает на умеренную согласованность экспертов.

Таблица 13.7

Классификация $n=25$ объектов. Объекты классифицируются различными группами экспертов

Объект (i)	Число экспертов (m_i)	Число „положительных" классификаций (x_i)	Объект (i)	Число экспертов (m_i)	Число „положительных" классификаций (x_i)
1	2	2	15	2	0
2	2	0	16	2	2
3	3	2	17	3	1
4	4	3	18	2	1
5	3	3	19	4	1
6	4	1	20	5	4
7	3	0	21	3	2
8	5	0	22	4	0
9	2	0	23	3	0
10	4	4	24	3	3
11	5	5	25	2	2
12	3	3			
13	4	4	Сумма	81	46
14	4	3			

Флейс и Казик [Fleiss and Cuzick, 1979] вывели стандартную ошибку $\varkappa$, которую можно применять для проверки гипотезы о нулевом истинном значении каппа. Пусть $\overline{m}_H$ — *гармоническое среднее* число экспертов на объект, т. е.

$$\overline{m}_H = n \Big/ \sum_{i=1}^{n} \frac{1}{m_i} . \tag{13.45}$$

Стандартная ошибка $\varkappa$ приближенно равна:

$$\mathrm{se}(\hat{\varkappa}) = \frac{1}{(\overline{m}-1)\sqrt{n\,\overline{m}_H}} \sqrt{2(\overline{m}_H - 1) + \frac{(\overline{m}-\overline{m}_H)(1-4\,\overline{p}\,\overline{q})}{\overline{m}\,\overline{p}\,\overline{q}}} . \tag{13.46}$$

Гипотезу проверяют, сравнивая

$$z=\frac{\hat{\varkappa}}{\text{s. e.}_0(\hat{\varkappa})} \tag{13.47}$$

с критическим значением стандартного нормального распределения.

Для данных табл. 13.7

$$\bar{m}_H=\frac{25}{8{,}5167}=2{,}935,$$

$$\text{s.e.}_0(\varkappa)=\frac{1}{(3{,}24-1)\sqrt{25\cdot 2{,}935}}\times$$

$$\times\sqrt{2(2{,}935-1)+\frac{(3{,}24-2{,}935)\,(1-4\cdot 0{,}568\cdot 0{,}432)}{3{,}24\cdot 0{,}568\cdot 0{,}432}}=0{,}103.$$

Значение (13.47)

$$z=\frac{0{,}54}{0{,}103}=5{,}24$$

указывает, что $\hat{\varkappa}$ значимо отличается от нуля.

Теперь допустим, что число категорий $k\geqslant 2$. Обозначим через $\bar{p}_j$ общую пропорцию классификаций в j-ю категорию, через $\hat{\varkappa}_j$ — значение каппа для этой категории, $j=1, \ldots, k$. Лэндис и Кох [Landis and Koch, 1977] предложили брать в качестве общей меры согласованности экспертов взвешенное среднее

$$\hat{\hat{\varkappa}}=\frac{\sum_{j=1}^{k}\bar{p}_j\bar{q}_j\hat{\varkappa}_j}{\sum_{j=1}^{k}\bar{p}_j\bar{q}_j}, \tag{13.48}$$

где $\bar{q}_j=1-\bar{p}_j$.

Выражение стандартной ошибки $\hat{\hat{\varkappa}}$ для проверки гипотезы о равенстве истинного значения нулю в случае различного числа экспертов на объект еще предстоит вывести. Однако известны простые выражения для $\hat{\varkappa}_j$, $\hat{\hat{\varkappa}}$ и их стандартных ошибок при условии одинакового числа $m_i=m$ экспертов. Если x_{ij} — число классификаций i-го объекта ($i=1, \ldots, n$) в j-ю категорию ($j=1, \ldots, k$), то, поскольку

$$\sum_{j=1}^{k}x_{ij}=m \tag{13.49}$$

ля всех i, значение $\hat{\kappa}_j$ равно:

$$\hat{\kappa}_j = 1 - \frac{\sum_{i=1}^{n} x_{ij}\,(m - x_{ij})}{nm\,(m-1)\,\bar{p}_j\,\bar{q}_j}\,, \qquad (13.50)$$

значение $\hat{\hat{\kappa}}$ —

$$\hat{\hat{\kappa}} = 1 - \frac{nm^2 - \sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}^2}{nm(m-1)\sum_{j=1}^{k} \bar{p}_j \bar{q}_j}\,. \qquad (13.51)$$

лгебраически эквивалентные выражения для этих статистик казаны в [Fleiss, 1971], где явно показано, что в эти меры ходят поправки на случайную согласованность.

Таблица 13.8

Классификация десяти объектов пятью экспертами по трем категориям

	Число классификаций в j-ю категорию			
Объект	$j=1$	$j=2$	$j=3$	$\sum_{j=1}^{3} x_{ij}^2$
1	1	4	0	17
2	2	0	3	13
3	0	0	5	25
4	4	0	1	17
5	3	0	2	13
6	1	4	0	17
7	5	0	0	25
8	0	4	1	17
9	1	0	4	17
10	3	0	2	13
Сумма	20	12	18	174

Табл. 13.8 содержит иллюстративные данные, которые представляют результаты классификации каждого из $n=10$ объектов $m=5$ экспертами по $k=3$ категориям.

Общие пропорции равны: $\bar{p}_1=20/50=0{,}40$, $\bar{p}_2=12/50=0{,}24$, $\bar{p}_3=18/50=0{,}36$. Числитель в (13.50) для первой категории равен:

$$\sum_{i=1}^{n} x_{i1}\,(5-x_{i1}) = 1\cdot(5-1)+2\cdot(5-2)+\ldots+3\cdot(5-3)=34,$$ значит,

$$\hat{\kappa}_1 = 1-\frac{34}{10\cdot 5\cdot 4\cdot 0{,}40\cdot 0{,}60} = 0{,}29.$$

Аналогично $\hat{\kappa}_2=0{,}67$, $\hat{\kappa}_3=0{,}35$.

Общее значение $\hat{\bar{\kappa}}$ равняется согласно (13.51)

$$\hat{\bar{\kappa}} = 1-\frac{10\cdot 25-184}{10\cdot 5\cdot 4\cdot(0{,}40\cdot 0{,}60+0{,}24\cdot 0{,}76+0{,}36\cdot 0{,}64)} = 0{,}42,$$

или согласно (13.48) —

$$\hat{\bar{\kappa}} = \frac{(0{,}40\cdot 0{,}60)\cdot 0{,}29+(0{,}24\cdot 0{,}76)\cdot 0{,}67+(0{,}36\cdot 0{,}64)\cdot 0{,}35}{0{,}40\cdot 0{,}60+0{,}24\cdot 0{,}76+0{,}36\cdot 0{,}64} = 0{,}42.$$

Флейс, Ни и Лэндис [Fleiss, Nee and Landis, 1979] вывели следующие приближенные формулы для стандартных ошибок $\hat{\bar{\kappa}}$ и $\hat{\kappa}_j$, которые можно использовать для проверки гипотезы о равенстве соответствующего истинного значения нулю:

$$\text{s.e.}_0(\hat{\bar{\kappa}}) = \frac{\sqrt{2}}{\sum_{j=1}^{k}\bar{p}_j\bar{q}_j\sqrt{nm(m-1)}}\times$$

$$\times\sqrt{\left(\sum_{j=1}^{k}\bar{p}_j\bar{q}_j\right)^2-\sum_{j=1}^{k}\bar{p}_j\bar{q}_j(\bar{q}_j-\bar{p}_j)}, \qquad (13.52)$$

$$\text{s.e.}_0(\hat{\kappa}_j) = \sqrt{\frac{2}{nm(m-1)}}. \qquad (13.53)$$

Обратите внимание, что $\text{s.e.}_0(\hat{\kappa}_j)$ не зависит от $\bar{p}_j$ и $\bar{q}_j$. Кроме того, легко проверить, что (13.53) — частный случай выражения (13.46), когда все m_i равны, поскольку при этом $\bar{m}=\bar{m}_H=m$.

Для данных табл. 13.8

$$\sum_{j=1}^{3}\bar{p}_j\bar{q}_j = 0{,}40\cdot 0{,}60+0{,}24\cdot 0{,}76+0{,}36\cdot 0{,}64 = 0{,}653,$$

$$\sum_{j=1}^{3}\bar{p}_j\bar{q}_j(\bar{q}_j-\bar{p}_j) = 0{,}40\cdot 0{,}60\cdot(0{.}60-0{,}40)+0{,}24\cdot 0{,}76\cdot(0{,}76-0{,}24)+$$
$$+0{,}36\cdot 0{,}64\,(0{,}64-0{,}36) = 0{,}2074,$$

поэтому

$$\text{s.e.}_0(\hat{\bar{\kappa}}) = \frac{\sqrt{2}}{0{,}653\sqrt{10\cdot5\cdot4}}\sqrt{0{,}653^2 - 0{,}2074} = 0{,}072.$$

Поскольку

$$z = \frac{\hat{\bar{\kappa}}}{\text{s.e.}_0(\hat{\bar{\kappa}})} = \frac{0{,}42}{0{,}072} = 5{,}83,$$

то общее значение каппа значимо отличается от нуля (хотя ее величина указывает на весьма посредственную согласованность).

Согласно (13.53) приближенная стандартная ошибка любой из трех $\hat{\kappa}_j$ равна:

$$\text{s.e.}_0(\hat{\kappa}_j) = \sqrt{\frac{2}{10\cdot5\cdot4}} = 0{,}10.$$

Все каппа значимо ($p<0{,}01$) отличаются от нуля, хотя лишь κ_2 достигает значения, соответствующего довольно высокой согласованности.

Лэндис и Кох [Landis and Koch, 1977b] описывают, как вычислять стандартные ошибки $\hat{\kappa}_j$ и $\hat{\bar{\kappa}}$, которые можно использовать в случае ненулевых истинных значений. Их методы и результаты слишком сложны, чтобы быть представленными в данной книге.

13.3. Дальнейшие приложения

Хотя каппа-статистики первоначально разрабатывались для измерения степени согласованности экспертов (и именно в таком контексте описывались в двух предыдущих разделах), область их применений значительно шире. Они полезны при определении по категоризованным данным таких качеств, как «сходство», «соответствие», «кластерная структура». Ниже приводятся некоторые примеры приложений.

1. При исследовании сопутствующих или решающих факторов подростковой наркомании может оказаться интересным определение соответствия между отношением к наркотикам одного из родителей (того же пола, что и подросток) и отношением к наркотикам лучшего друга подростка. Здесь можно использовать как взвешенную, так и невзвешенную каппа (разд.

13.1), считая экспертом A — родителя, экспертом B — друга подростка.

2. Допустим, что в городе установлено m станций, следящих за уровнем загрязнения воздуха, и что на протяжении n дней каждая станция сообщает, превышает ли уровень содержания некоторого вредного вещества (например, двуокиси серы) официально установленный порог. Чтобы определить, насколько хорошо (или плохо) согласуются показания нескольких станций слежения, можно применить вариант каппа, описанный в разд. 13.2.

3. Допустим, проводится изучение роли наследственных факторов в развитии юношеской гипертонии, к которому привлечено n семей. Пусть в i-й семье m_i братьев и сестер. Чтобы описать степень наследственного сцепления по заболеванию, можно использовать вариант каппа, описанный в разд. 13.2.

4. Многие из индексов согласованности, указанных в разд. 13.1, используются в численной таксономии ([Sneath and Socal, 1973]) для определения степени сходства различных объектов в анализе (кстати, p_s (13.2) был предложен Дайсом [Dice, 1945] именно для этой цели). Предположим, что два объекта — люди, языки или что-нибудь еще — сравниваются по тому, обладают они или нет каждым из n дихотомических признаков. Пропорциями a, b, c, d в табл. 13.2 будут тогда пропорция числа признаков, которыми обладают оба объекта, пропорция числа признаков, которыми обладает первый объект, но не обладает второй, и т. д. Поправки на случайное сходство в этом случае так же важны, как и поправки на случайную согласованность при определении степени согласованности экспертов.

5. Исследования, в которых с каждым опытным (экспериментальным) объектом связывалось несколько контрольных объектов, обсуждались в разд. 8.3. Если контрольные объекты были хорошо подобраны в каждой связке, сходство реакций контрольных объектов в одной связке должно быть сильнее, чем в разных связках. Чтобы выяснить, насколько хорошо было проведено связывание, можно применить вариант каппа из разд. 13.2.

Как легко заметить, каппа в качестве меры согласованности или сходства симметрична по отношению к экспертам или объектам. Однако нередко один или несколько экспертов могут считаться стандартом для остальных (например, два из $m=5$ экспертов могут быть старше или опытнее; в примере 2 одна из станций слежения за уровнем загрязнения воздуха может быть оснащена более точными приборами). В таком случае каппа уже не применима, и надо использовать методы, описанные в [Light, 1971, Williams, 1976, Wackerley et al., 1978].

Задачи

13.1. Докажите, что учет случайной согласованности в любом из индексов (13.1), (13.2), (13.4), (13.6) и (13.7) с помощью формулы (13.8) приводит к одному и тому же выражению (13.9).

13.2. Докажите, что при $k=2$ квадрат отношения (13.14) совпадает со стандартной статистикой хи-квадрат без поправки на непрерывность.

13.3. Допустим, при проведении $g=3$ независимых исследований по достоверности классификации получены следующие результаты:

Исследование 1 ($n=20$)			Исследование 2 ($n=20$)			Исследование 3 ($n=30$)		
Эксперт A	Эксперт B		Эксперт C	Эксперт D		Эксперт E	Эксперт F	
	+	−		+	−		+	−
+	0,60	0,05	+	0,75	0,10	+	0,50	0,20
−	0,20	0,15	−	0,05	0,10	−	0,10	0,20

а) Вычислите значения каппа для этих исследований, их стандартные ошибки (13.15) и общее значение каппа (13.21).

б) Значимо ли различие трех значений каппа? (Вычислите значение статистики (13.22) и обратитесь к таблице распределения хи-квадрат с двумя степенями свободы.)

в) Найдите с помощью (13.23) приближенный 95%-ный доверительный интервал для общего значения каппа.

13.4. Продемонстрируйте совпадение выражений (13.13) и (13.22) для стандартных ошибок при $w_{ji}=1$ для всех i и $w_{ij}=0$ для $i\neq j$. Докажите также тождество выражений (13.15) и (13.36) при этой системе весов.

13.5. Докажите, что при $k=2$ (13.52) и (13.53) совпадают.

ЛИТЕРАТУРА

Armitage, P., Blendis, L. M., and Smyllie, H. C. (1966). The measurement of observer disagreement in the recording of signs. *J. R. Stat. Soc., Ser. A*, **129**, 98–109.

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychol. Rep.*, **19**, 3–11.

Bennett, B. M. (1972). Measures for clinicians' disagreements over signs. *Biometrics*, **28** 607–612.

Cicchetti, D. V. (1976). Assessing inter-rater reliability for rating scales: Resolving some basic issues. *Br. J. Psychiatry*, **129**, 452–456.

Cicchetti, D. V. and Allison, T. (1971). A new procedure for assessing reliability of scoring EEG sleep recordings. *Am. J. EEG Technol.*, **11**, 101–109.

Cicchetti, D. V. and Fleiss, J. L. (1977). Comparison of the null distributions of weighted kappa and the C ordinal statistic. *Appl. Psychol. Meas.*, **1**, 195–201.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.*, **20**, 37–46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.*, **70**, 213–220.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, **26**, 297–302.

Ebel, R. L. (1951). Estimation of the reliability of ratings. *Psychometrika*, **16**. 407–424.

Fleiss. J. L. (1965). Estimating the accuracy of dichotomous judgments. *Psychometrika*, **30**, 469–479.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychol. Bull.*, **76**, 378–382.

Fleiss, J. L. (1975). Measuring agreement between two judges on the presence or absence of a trait. *Biometrics*, **31**, 651–659.

Fleiss, J. L. and Cicchetti, D. V. (1978). Inference about weighted kappa in the non-null case. *Appl. Psychol. Meas.*, **2**, 113–117.

Fleiss, J. L. and Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ. Psychol. Meas.*, **33**, 613–619.

Fleiss, J. L., Cohen, J., and Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychol. Bull.*, **72**, 323–327.

Fleiss, J. L. and Cuzick, J. (1979). The reliability of dichotomous judgments: Unequal numbers of judges per subject. *Appl. Psychol. Meas.*, **3**, 537–542.

Fleiss, J. L., Nee, J. C. M., and Landis, J. R. (1979). The large sample variance of kappa in the case of different sets of raters. *Psychol. Bull.*, **86**, 974–977.

Goodman, L. A. and Kruskal, W. H. (1954). Measures of association for cross classifications. *J. Am. Stat. Assoc.*, **49**, 732–764.

Holley, J. W. and Guilford, J. P. (1964). A note on the *G* index of agreement. *Educ. Psychol. Meas.*, **32**, 281–288.

Hubert, L. J. (1978). A general formula for the variance of Cohen's weighted kappa. *Psychol. Bull.*, **85**, 183–184.

Krippendorff, K. (1970). Bivariate agreement coefficients for reliability of data. Pp. 139–150 in E. F. Borgatta (Ed.). *Sociological methodology 1970*. San Francisco: Jossey-Bass.

Landis, J. R. and Koch, G. G. (1977a). The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–174.

Landis, J. R. and Koch, G. G. (1977b). A one-way components of veriance model for categorical data. *Biometrics*, **33**, 671–679.

Light, R. J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. *Phychol. Bull.*, 76, 365–377.

Maxwell, A. E. (1977). Coefficients of agreement between observers and their interpretation. *Br. J. Psychiatry*, **130**, 79–83.

Maxwell, A. E. and Pilliner, A. E. G. (1968). Deriving coefficients of reliability and agreement for ratings. *Br. J. Math. Stat. Psychol.*, **21**, 105–116.

Bogot, F. and Goldberg, I. D. (1966). A proposed index for measuring agreement in test-retest studies. *J. Chronic Dis.*, 19, 991–1006.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quart.*, 19, 321–325.

Sneath, P. H. A. and Sokal, R. R. (1973). *Numerical Taxonomy*. San Francisco: W. H. Freeman.

Spitzer, R. L., Cohen, J., Fleiss, J. L., and Endicott, J. (1967). Quantification of agreement in psychiatric diagnosis. *Arch. Gen. Psychiatry*, 17, 83–87.

Spitzer, R. L. and Fleiss, J. L. (1974). A reanalysis of the reliability of psychiatric diagnosis. *Br. J. Psychiatry*, **125**, 341–347.

Wackerley, D. D., McClave, J. T., and Rao, P. V. (1978). Measuring nominal scale agreement between a judge and a known standard. *Psychometrika*, **43**, 213–223.

Williams, G. W. (1976). Comparing the joint agreement of several raters with another rater. *Biometrics*, 32, 619–627.

Глава 14
Стандартизация долей

Одна из проблем, наиболее часто возникающих в эпидемиологии и биомедицинских приложениях статистики, — это сравнение долей наличия некоторого фактора или события в разных популяциях или в одной популяции, но в различные периоды времени. Если бы популяции были сходны по факторам, взаимодействующим с исследуемым событием (например, по полу, возрасту, расе, семейному положению), то сравнение общих, или «грубых», долей (overall, total или crude rates) не вызывало бы особых затруднений.

Однако, если популяции различаются по отношению к таким факторам, прямое сравнение может приводить к ошибочным выводам. Эту задачу сравнения можно формализовать алгебраически следующим образом. Пусть $p_1, \ldots, p_I$ — пропорции объектов одной из сравниваемых популяций в различных слоях (возрастных интервалах, социально-экономических группах и т. д.), причем всего в популяции I слоев. Естественно, $\sum_{i=1}^{I} p_i = 1$. Если c_i — доля наступления события для i-го слоя этой популяции (*частная доля* i-го слоя), то *грубая* (или общая, средняя) доля в популяции —

$$c = \sum_{i=1}^{I} c_i p_i. \tag{14.1}$$

Если распределение объектов второй популяции по I слоям описывается пропорциями $P_1, \ldots, P_I$ ($\sum P_i = 1$), а C_i — частотная доля i-го слоя второй популяции, то грубая доля в этой популяции равна:

$$C = \sum_{i=1}^{I} C_i P_i. \tag{14.2}$$

Разность двух грубых долей,

$$d = c - C, \tag{14.3}$$

как легко проверить, равняется:

$$d = \sum_{i=1}^{I} \frac{p_i + P_i}{2} (c_i - C_i) + \sum_{i=1}^{I} \frac{c_i + C_i}{2} (p_i - P_i) \qquad (14.4)$$

(см. [Kitagawa, 1955; Hemphill and Ament, 1970]). Миеттинен [Miettinen, 1972] привел аналогичные выкладки для отношения двух грубых долей.

Итак, разность двух грубых долей распадается на две компоненты. Первая составляющая

$$d_1 = \sum_{i=1}^{I} \frac{p_i + P_i}{2} (c_i - C_i) \qquad (14.5)$$

обобщенно описывает различие между двумя наборами частных долей, которое, собственно, и представляет основной интерес. В то же время вторая компонента

$$d_2 = \sum_{i=1}^{I} \frac{c_i + C_i}{2} (p_i - P_i) \qquad (14.6)$$

соответствует разнице между распределением объектов в популяциях по слоям, обычно мало или совсем неинтересной.

Разложение (14.4) позволяет высказать следующие утверждения:

1. Если распределения в популяциях одинаковы, т. е. $p_1 = P_1, \ldots, p_I = P_I$, то $d_2 = 0$ и разность грубых долей действительно обобщенно выражает различие между двумя наборами частных долей.

2. Если два множества частных долей совпадают, т. е. $c_1 = C_1, \ldots, c_I = C_I$, то $d_1 = 0$ и разность грубых долей описывает лишь различие между популяциями по распределению объектов по слоям, которое, как правило, не важно.

3. При любом фиксированном значении d_1, обозначающем различие между двумя наборами частных долей, наблюдаемая разность общих долей двух популяций, выражаемая через d, может оставаться неизменной, увеличиваться, уменьшаться и даже изменять знак в зависимости от распределений объектов в популяциях. Компонента d_2 дает положительный прирост, если в первой популяции пропорции объектов больше по сравнению со второй популяцией в тех слоях, в которых частные доли велики (т. е. при $p_i > P_i$ в тех слоях, где $(c_i + C_i)/2$ велико), и отрицательный прирост — в противном случае. Именно такой механизм действует, когда грубая доля в первой популяции меньше, чем во второй, несмотря на то, что все частные доли первой популяции больше соответствующих долей второй популяции.

В разд. 14.1 представлены некоторые доводы в пользу стандартизации, а также даны предостережения от ее необдуманного применения. В разд. 14.2 описывается непрямой метод стандартизации, а в разд. 14.3 приведен пример, показывающий, что применение этого метода может привести к ошибочным результатам. Разд. 14.4 посвящен прямому методу стандартизации, а в разд. 14.5 рассматриваются некоторые другие методы. В разд. 14.6 обсуждаются способы стандартизации по двум взаимодействующим факторам.

14.1. Стандартизация: за и против

Цель стандартизации (или выравнивания долей) — предупреждать при сравнении двух или более множеств частных долей аномалии, описанные выше в пунктах 2 и 3. Стандартизация, сама по себе, не способна заменить сравнение отдельных частных долей. Она используется для того, чтобы в целом описать изучаемую популяцию (в терминах заболеваемости, смертности и т. п., в зависимости от того, к чему относятся частные доли).

Вулси [Woolsey, 1959, p. 60] указал, что «особый интерес представляют именно частные доли, поскольку вариации (исследуемого явления) в различных классах популяций можно подробно и точно проследить только с помощью анализа частных долей». Элвибэк [Elveback, 1966] также подчеркивает важность анализа частных долей и резко критикует применение стандартизованных долей.

Один из доводов против стандартизации состоит в следующем: нельзя с помощью какого-либо отдельно взятого метода стандартизации обнаружить, что различие между частными долями меняется от слоя к слою. Наоборот, выравнивание будет маскировать эти различия. Можно привести пример из [Kitagawa, 1966], где сравниваются частные по отношению к возрасту доли смертей среди белых мужчин, живущих в столичных округах Соединенных Штатов, и соответствующие частные доли среди мужчин, живущих в нестоличных округах. Для возраста до 40 лет первые частные доли меньше вторых, после 40 лет справедливо обратное соотношение. Сравнение по какому-либо одиночному показателю не сможет вскрыть это различие. Чтобы не потерять информацию, здесь потребуется сравнение по меньшей мере двух показателей.

Другим примером служат данные из [El-Badry, 1969], где указано, что в Цейлоне, Индии и Пакистане во многих возрастных категориях смертность у мужчин ниже, чем у женщин. Одиночные суммарные индексы для мужчин и женщин могут замаскировать это явление и, следовательно, скрыть на-

правление дальнейших исследований. Долл и Кук [Doll and Cook, 1967] описывают другие случаи неадекватности одиночного обобщающего индекса при описании долей событий, зависимых от возраста и пола.

Несмотря на то, что сравнение по какому-либо показателю не заменяет непосредственный анализ частных долей, можно указать несколько доводов в пользу стандартизации.

1. Две популяции легче сравнивать по одиночному суммарному показателю, чем по полным наборам частных долей.

2. Если какой-нибудь слой представлен малым числом объектов, соответствующие частные доли могут быть слишком неточными и недостоверными, чтобы использовать их для сравнения.

3. Для малых популяций и некоторых групп вычислить частные доли порой бывает невозможно. Такая ситуация может встретиться при изучении определенных профессиональных групп или при изучении популяций в географических районах, специально выделенных для отдельного исследования. В таких случаях можно располагать данными лишь о суммарном числе событий (например, смертей), но не о числе событий в отдельных слоях.

Другие доводы в пользу стандартизации приведены в [Woolsey, 1959; Kalton, 1968; Cochran, 1968], где также изучается влияние на результат изменения числа слоев I. В [Mansner and Bahn, 1974, p. 138] изящно резюмируются достоинства и недостатки, присущие анализу по грубым, частным или стандартизованным долям.

14.2. Непрямая стандартизация

Второй и третий доводы в пользу стандартизации (недостоверность или даже невозможность вычисления некоторых частных долей), служат основанием, пожалуй, для наиболее распространенного метода стандартизации — так называемого непрямого метода (indirect standardization). Для его реализации необходимы следующие данные:

1. Грубая доля c для исследуемой популяции.

2. Распределение объектов по различным слоям этой популяции — $p_1, \ldots, p_I$.

3. Множество частных долей выбранной стандартной популяции — $c_{S1}, \ldots, c_{SI}$.

4. Грубая доля c_S для стандартной популяции.

Непрямую стандартизацию начинают с вычисления общей доли для исследуемой популяции при условии, что набор част-

ных долей в этой популяции такой же, как в стандартной популяции. Эта грубая доля [1] равна:

$$c' = \sum_{i=1}^{I} c_{Si} p_i . \qquad (14.7)$$

После этого получают непрямо стандартизованную долю (indirect adjusted rate)

$$c_{\text{indirect}} = c_S \frac{c}{c'} . \qquad (14.8)$$

Она равна грубой доле для стандартной популяции c_S, умноженной на отношение наблюдаемой грубой доли исследуемой популяции c к грубой доле c', которая наблюдалась бы, если бы исследуемая популяция имела набор частных долей стандартной популяции.

Для примера возьмем данные из [Stark and Mantel, 1966]. В штате Мичиган за период от 1950 до 1964 г. 731177 матерей родили первого ребенка (в этих и последующих данных учитываются только дети, родившиеся живыми). У 412 из этих детей было обнаружено заболевание монголизм (болезнь Дауна), что в результате дает значение грубой доли, равной $c = 56{,}3$ случая монголизма на 100000 первенцев. За тот же 15-летний период родилось 442811 детей, которые были пятыми, шестыми и т. д. У 740 из них был найден монголизм, что дает грубую долю, равную $C = 167{,}1$ случая на 100000 детей.

Непосредственное сравнение этих двух долей не вполне допустимо, поскольку возраст матери связан зависимостью как с числом детей у нее, так и с наличием монголизма у ребенка. Значит, надо применять такой метод стандартизации, в котором выравнивалось бы различие между распределениями по возрасту для матерей с первым ребенком и для матерей с пятым, шестым и т. д. Непрямой метод демонстрируется в табл. 14.1.

В качестве стандартной популяции были выбраны все дети, родившиеся в Мичигане за 1950—1964 гг. Грубая доля монголизма в этом штате в целом составила $c_S = 89{,}5$ на 100000 детей, а частные по возрасту матери доли указаны во второй колонке таблицы. Распределения возраста матерей, родивших первого ребенка, и матерей, родивших пятого и шестого и т. д. ребенка, приведены в третьей и пятой колонках. Значения, вычисленные по формуле (14.7), даны в последней строке таблицы.

[1] В разд. 14.3 и 14.6 она называется ожидаемой грубой долей. — *Примеч. пер.*

Таблица 14.1

Пример непрямой стандартизации

Возраст матери	Частные доли c_{Si} на 100000 детей по Мичигану в целом	Ребенок			
		первый		пятый, шестой и т. д.	
		p_i	$c_{Si}\,p_i$	P_i	$c_{Si}\,P_i$
Менее 20	42,5	0,315	13,4	0,001	0,0
20–24	42,5	0,451	19,2	0,069	2,9
25–29	52,3	0,157	8,2	0,279	14,6
30–34	87,7	0,054	4,7	0,339	29,7
35–39	264,0	0,019	5,0	0,235	62,0
40 и более	864,4	0,004	3,5	0,078	67,4
Сумма			$c'=54$		$C'=176,6$

Итак, нам даны грубые доли случаев монголизма на 100000 детей:

$$c=56,3 \qquad (14.9)$$

— для первенцев и

$$C=167,1 \qquad (14.10)$$

— для пятых, шестых и т. д. детей.

Используя частные по возрасту матери доли для Мичигана в целом, мы вычисляем по (14.7) грубые доли:

$$c'=54,0$$

— для первенцев и

$$C'=176,6$$

— для пятых, шестых и т. д. детей. Грубая доля для штата в целом равна:

$$c_S=89,5 \text{ случая монголизма на 100000 детей.}$$

Подставляя это значение в (14.8), находим непрямо стандартизованные доли: для первых детей —

$$c_{\text{indirect}}=89,5\cdot\frac{56,3}{54,0}=93,3, \qquad (14.11)$$

для пятых, шестых и т. д.—

$$C_{\text{indirect}} = 89{,}5 \cdot \frac{167{,}1}{176{,}6} = 84{,}7. \qquad (14.12)$$

Сравнивая исходные грубые доли (14.9) и (14.10), мы видим, что риск монголизма у детей, родившихся пятыми, шестыми и т. д., возрастает в три раза по сравнению с первенцами. С другой стороны, применяя для сравнения стандартизованные доли (14.11) и (14.12), мы сделаем вывод, что большого различия в риске монголизма нет.

Наблюдаемое в соответствии с грубыми долями увеличение риска для детей, родившихся позже, видимо, объясняется различием в распределении возраста матери. К старшим возрастным категориям, где частные доли монголизма выше, относится большее число матерей с последующими детьми, чем матерей с первым ребенком. После выравнивания различий по возрасту оказывается, что, если вообще есть хотя бы какое-нибудь различие, оно таково, что доли монголизма у первых детей выше, чем у пятых, шестых и т. д.

14.3. Одно свойство непрямой стандартизации

Рассмотрим вымышленные данные табл. 14.2, описывающие зависимость смертности (на 1000 человек) от пола в двух группах.

Таблица 4.2

Частные по отношению к полу доли смертности в двух группах

Пол	Группа 1		Группа 2	
	p_i	доля/1000	P_i	доля/1000
Мужчина	0,60	2,0	0,80	2,0
Женщина	0,40	1,0	0,20	1,0

Множества частных по полу долей совпадают, но различное распределение объектов по полу в этих группах приводит к неравенству грубых долей. Для группы 1 грубая доля равна:

$$c = 2{,}0 \cdot 0{,}60 + 1{,}0 \cdot 0{,}40 = 1{,}6 \text{ (смертей на 1000 чел.)}, \qquad (14.13)$$

для группы 2 —

$$C = 2{,}0 \cdot 0{,}80 + 1{,}0 \cdot 0{,}20 = 1{,}8 \text{ (смертей на 1000 чел.)}. \qquad (14.14)$$

Теперь допустим, что на самом деле получение частных долей потребовало бы больших затрат, так что фактически мы располагаем лишь данными о распределениях по полу в этих группах и о двух грубых долях. Значит, чтобы сравнить две группы, надо применить непрямую стандартизацию, используя данные о некоторой стандартной популяции. Предположим, что в стандартной популяции грубая доля равна:

$$c_S = 1{,}5 \text{ (смертей на 1000 чел.)}, \tag{14.15}$$

а частные по отношению к полу доли составляют:

$$c_{S1} = 2{,}2 \text{ (смертей на 1000 мужчин)}, \tag{14.16}$$

$$c_{S2} = 0{,}9 \text{ (смертей на 1000 женщин)}. \tag{14.17}$$

Найдем теперь ожидаемую грубую долю в группе 1:

$$c' = 2{,}2 \cdot 0{,}60 + 0{,}9 \cdot 0{,}40 = 1{,}68 \text{ (смертей на 1000 чел.)}, \tag{14.18}$$

что дает непрямо стандартизованную долю:

$$c_{\text{indirect}} = 1{,}5 \cdot \frac{1{,}6}{1{,}68} = 1{,}43 \text{ (смертей на 1000 чел.)}. \tag{14.19}$$

Соответствующие величины для второй группы равны:

$$C' = 2{,}2 \cdot 0{,}80 + 0{,}9 \cdot 0{,}20 = 1{,}94 \text{ (смертей на 1000 чел.)}, \tag{14.20}$$

$$C_{\text{indirect}} = 1{,}5 \cdot \frac{1{,}8}{1{,}94} = 1{,}39 \text{ (смертей на 1000 чел.)}. \tag{14.21}$$

Стандартизованные доли (14.19) и (14.21) ближе друг к другу, чем исходные грубые доли (14.13) и (14.14), и более точно отражают равенство частных долей в табл. 14.2, хотя тот факт, что совпадение двух наборов частных долей не привело к точному равенству стандартизованных долей, вызывает легкое беспокойство. Это свойство характерно для непрямой стандартизации. Ниже мы рассмотрим прямую стандартизацию, для которой подобное явление невозможно. Однако в большинстве случаев, как и в этом примере, искажение невелико. Более того, такое искажение исключено, когда стандартная популяция является объединением двух исследуемых популяций.

Непрямая стандартизация, очевидно, не может полностью учесть различия в составе популяции. Поэтому, пытаясь описать различия между группами с помощью непрямо стандартизованных долей, следует иметь в виду, что различие между значениями стандартизованных долей зависят не только от различия множеств частных долей, но и от различия в составах популяций. Критику непрямой стандартизации можно найти еще в [Yule, 1934; Kilpatrick, 1963]. Бреслоу и Дэй [Breslow and Day, 1975], однако, предлагают математическую модель для част-

ных долей, в которой непрямая стандартизация адекватна (в этой модели предполагается, что каждая частная доля является произведением двух множителей, один из которых описывает слой, а другой — популяцию).

14.4. Прямая стандартизация

Вторым, часто применяемым методом стандартизации, является так называемый прямой метод (direct standardization). Прямая стандартизация может использоваться только тогда, когда известны наборы частных долей в исследуемых популяциях. Для реализации этого метода необходимо располагать следующими данными:

1. Набор частных долей исследуемой популяции, $c_1, c_2, \ldots, c_I$.

2. Распределение по слоям объектов стандартной популяции, $p_{S1}, \ldots, p_{SI}$.

Прямо стандартизованная доля (direct adjusted rate) — это просто

$$c_{\text{direct}} = \sum_{i=1}^{I} c_i p_{Si} . \qquad (14.22)$$

Слово «прямо» означает, что в данном случае мы работаем непосредственно с частными долями исследуемой популяции в отличие от непрямой стандартизации, рассмотренной в предыдущих разделах.

Чтобы проиллюстрировать прямой метод, вновь обратимся к примеру, из разд. 14.2. В табл. 14.3 приводится распределение возраста матерей для детей, родившихся в штате Мичиган с 1950 по 1964 г. (стандартное распределение), и частные доли по отношению к возрасту матери заболевания монголизмом для первенцев и для детей, родившихся пятыми, шестыми и т. д. (данные взяты из [Stark and Mantel, 1966]).

Заключение, которое будет сделано при сравнении прямо стандартизованных долей, совпадает с заключением, принятым по непрямо стандартизованным долям: доли монголизма для двух рассматриваемых групп детей примерно одинаковы. Такое согласие выводов, сделанных после прямой и непрямой стандартизации, обычно, хотя и не всегда, обязательно.

Принимая во внимание постоянство соотношения между частными долями табл. 14.3 (в пяти из шести категориях возраста матери риск монголизма для первого ребенка немного выше, чем для пятого, шестого и т. д.), можно уверенно сказать, что нет никаких убедительных причин проводить стандартизацию. Частные доли говорят сами за себя. Единственным оправдани-

Таблица 14.3

Пример прямой стандартизации
(доли приведены в расчете на 100000 чел.)

Возраст матери	Распределение по Мичигану в целом, p_{Si}	Ребенок			
		первый		пятый, шестой и т. д.	
		c_i	$c_i\, p_{Si}$	C_i	$C_i\, p_{Si}$
Менее 20	0,113	46,5	5,3	0	0,0
20—24	0,330	42,8	14,1	26,1	8,6
25—29	0,278	52,2	14,5	51,0	14,2
30—34	0,173	101,3	17,5	74,7	12,9
35—39	0,084	274,5	23,1	251,7	21,1
40 и более	0,022	819,1	18,0	857,8	18,9
Сумма			$c_{\text{direct}}=92,5$		$C_{\text{direct}}=75,7$

ом стандартизации в таком случае служит довод, приведенный в конце разд. 14.1,— с отдельным обобщающим индексом легче работать, чем с полным набором частных долей.

Прямая стандартизация имеет важнейшее преимущество перед непрямой стандартизацией. Если в каждом слое частные доли двух групп равны, то вне зависимости от того, какая популяция была выбрана в качестве стандартной, прямо стандартизованные доли будут совпадать. Рассмотрим, например, частные доли из табл. 14.2. Прямо стандартизованные доли для групп 1 и 2 совпадают и равны:

$$c_{\text{direct}}=2{,}0p_{S1}+1{,}0p_{S2}\,. \qquad (14.23)$$

Прямая стандартизация обладает более общим свойством. Если во всех слоях выполняется неравенство какого-либо вида между частными долями одной и другой популяции, то прямо стандартизованные доли будут связаны неравенством того же вида, т. е., например, если все частные доли первой группы больше соответствующих частных долей второй группы, то прямо стандартизованная доля в группе 1 будет больше, чем в группе 2, независимо от распределения стандартной популяции.

Конечно, на самом деле это свойство прямой стандартизации довольно тривиально, поскольку стандартизованные доли не добавляют ничего нового в картину, описываемую непосредственно частными долями. Здесь стандартизованная доля служит просто удобным суммарным показателем.

Стандартизованная доля, независимо от используемого метода стандартизации, имеет смысл лишь по отношению к другой стандартизованной доле, полученной точно таким же способом. Ее значение, отдельно взятое, не представляет интереса. Например, для долей из табл. 14.2 прямо стандартизованная доля принимает значения 1,25, 1,50 и 1,75, когда стандартное распределение по полу — соответственно (0,25; 0,75), (0,50; 0,50) и (0,75; 0,25). Две прямо стандартизованные доли для этих данных совпадают при любом стандарте, хотя их величины сильно зависят от распределения в стандартной популяции. Шпигельман и Маркс [Spiegelman and Marks, 1966] в экспериментальном исследовании показали, что при прямой стандартизации долей смертности выбор стандартной популяции обычно мало влияет на различие между стандартизованными долями и изменяет лишь их величины.

Если в соотношениях между частными долями сравниваемых групп в различных слоях не прослеживается какая-нибудь систематичность, то ценность любого метода стандартизации становится сомнительной. В задаче 14.1 требуется провести сравнение двух групп, у которых в одних слоях частные доли первой группы меньше, чем частные доли второй группы, а в других слоях, наоборот,— больше. Оказывается, что в зависимости от того, в каких слоях концентрируется стандартная популяция, любая из двух групп может иметь большую стандартизованную долю. Можно даже так подобрать стандартную популяцию, что стандартизованные доли в этих группах будут равны. Заметим еще, что обычные методы стандартизации будут скрывать наличие перекрытия по частным долям, если оно есть.

Компромиссное решение состоит в следующем: всю совокупность слоев нужно разбить на несколько множеств соседних слоев, чтобы частные доли двух групп во всех слоях данного множества были связаны одним соотношением, а затем вычислить для каждого такого множества стандартизованную долю. Этот способ продемонстрирован в задаче 14.1, где достаточно разбить совокупность слоев на два множества. Не всегда можно легко и естественно провести разбиение такого типа. Тем не менее, конечно, лучше оперировать несколькими стандартизованными долями, а не одной общей стандартизованной долей, которая дает, скорее, искаженное, чем обобщающее описание данных.

14.5. Другие обобщающие индексы

В [Woolsey, 1959; Kitagawa, 1964] приведены различные подходы к стандартизации долей, а в [Chiang, 1961; Keyfitz, 1966] даны выражения для стандартных ошибок.

Мы опишем сначала варианты двух методов выравнивания, рассмотренных выше. Вариантом непрямо стандартизованной доли можно считать величину

$$\mathrm{SMR} = \frac{c}{c'} = \frac{c_{\mathrm{indirect}}}{c_S}, \qquad (14.24)$$

являющуюся просто отношением наблюдаемой грубой доли к ожидаемой (c_s — доля в стандартной популяции). Если изучаемое явление — смертность, то этот индекс называют *стандартизованным отношением смертностей* (standardized mortality ratio, или standard mortality figure), откуда и взято обозначение. Величину SMR можно вычислять как для общей смертности, так и для смертности по определенной причине. Куппер и др. [Kupper et al., 1978] показали, как при некоторых умеренных ограничениях можно строить выводы о SMR для смертности по определенной причине, основываясь на анализе пропорциональных долей смертности (т. е. отношений числа смертей по отдельной причине к общему числу смертей). Гейл [Gail, 1978] описывает методы анализа вариации SMR в различных популяциях.

Соответствующий вариант для прямой стандартизации

$$\mathrm{CMF} = \frac{c_{\mathrm{direct}}}{c_S} \qquad (14.25)$$

в приложении к смертности называют *сравнительным значением смертности* (comparative mortality figure).

Существуют и другие методы выравнивания, однако применяемые реже тех, которые описаны нами выше. В одном из них просто вычисляется среднее грубой и прямо стандартизованной долей:

$$\mathrm{CMR} = \tfrac{1}{2}(c + c_{\mathrm{direct}}) = \sum_{i=1}^{I} \tfrac{1}{2}(p_{Si} + p_i)\, c_i. \qquad (14.26)$$

Для смертности этот индекс называют *сравнительной долей смертности* (comparative mortality rate). Его редкое применение — следствие затруднений в его интерпретации.

Два следующих индекса рассчитаны на использование долей, частных по отношению к возрасту, и фактически приписывают равный вес всем возрастам. Пусть n_i обозначает длину i-го возрастного интервала в годах (т. е. если первый возраст-

ной интервал — 0—4 года, то $n_1=5$ лет). Первый индекс [Yule, 1934] равен:

$$\mathrm{EADR}=\frac{\sum_{i=1}^{I} n_i c_i}{\sum_{i=1}^{I} n_i} \tag{14.27}$$

и в приложении к анализу смертности называется *эквивалентной средней долей смертей* (equivalent average death rate). Его можно рассматривать как прямо стандартизованную долю, если предположить, что число людей в различном возрасте одинаково.

Второй индекс [Yerushalmy, 1951; Elveback, 1966] равен:

$$\mathrm{MI}=\frac{\sum_{i=1}^{I} n_i \frac{c_i}{c_{Si}}}{\sum_{i=1}^{I} n_i} \tag{14.28}$$

и в приложении к смертности называется *индексом смертности* (mortality index). Индекс смертности — взвешенное среднее отношений частных долей, когда весом является длина возрастного интервала. Полезность последних двух индексов довольно ограничена, поскольку неясно, насколько допустимо присваивать равную важность каждому возрастному интервалу.

С индексом (14.28) сходен индекс

$$\mathrm{RMI}=\sum_{i=1}^{I} p_i \frac{c_i}{c_{Si}}, \tag{14.29}$$

называемый для смертности *относительным индексом смертности* (relative mortality index). Относительный индекс смертности — также взвешенное среднее отношений частных долей сравниваемой и стандартной популяций, но здесь веса соответствуют действительному распределению данной популяции по возрасту.

Эквивалентным ему выражением является

$$\mathrm{RMI}=\frac{\sum_{i=1}^{I} \frac{e_i}{c_{Si}}}{N}, \tag{14.30}$$

где e_i — наблюдаемое число смертей (или, в общем случае, событий) в i-м слое, а N — суммарное *число* людей в данной популяции. Таким образом, чтобы вычислить относительный

индекс смертности, о данной популяции надо знать лишь ее полный размер и число событий в каждом слое.

Этот индекс можно использовать, например, при сравнительном анализе данных для периода между переписями населения, когда распределение популяции по возрасту неизвестно, а распределение числа смертей по возрасту можно определить на основании записей актов гражданского состояния.

14.6. Выравнивание по двум факторам

В табл. 14.4 содержатся доли случаев заболевания монголизмом, частные как по отношению к возрасту матери, так и по отношению к числу ранее родившихся у этой матери детей. Мы опишем два метода, которые рассчитаны на случай, когда

Таблица 14.4

Распределение числа обнаруженных случаев заболеваний монголизмом и полного числа новорожденных в зависимости от возраста матери и числа родившихся у нее детей, Мичиган, 1950—1964 гг.

(в числителе — число новорожденных, знаменателе — число случаев монголизма) *

Возраст матери	Ребенок					Сумма
	1	2	3	4	5+	
Менее 20	107 / 230 061	25 / 72 202	3 / 15 050	1 / 2 293	0 / 327	136 / 319 933
20—24	141 / 329 449	150 / 326 701	71 / 175 702	26 / 68 800	8 / 30 666	396 / 931 318
25—29	60 / 114 920	110 / 208 667	114 / 207 084	64 / 132 424	63 / 123 419	411 / 786 511
30—34	40 / 39 487	84 / 83 228	103 / 117 300	89 / 98 301	112 / 149 919	428 / 488 235
35—39	39 / 14 208	82 / 28 466	108 / 45 026	137 / 46 075	262 / 104 088	628 / 237 863
40 и более	25 / 3 052	39 / 5 375	75 / 8 660	96 / 9 834	295 / 34 392	530 / 61 313
Сумма	412 / 731 177	490 / 721 639	474 / 568 819	413 / 357 727	740 / 442 811	2 529 / 2 823 173

* Данные взяты из [Stark and Mantel, 1966].

два фактора связаны с каким-то явлением и друг с другом и когда мы хотим определить влияние каждого из этих факторов по отдельности. Первый метод основан на прямой, второй — на непрямой стандартизации.

Одновременная прямая стандартизация. В табл. 14.5 приведены частные, грубые и прямо стандартизованные доли.

Таблица 14.5

Доля случаев заболевания монголизмом (число больных монголизмом на 100000 новорожденных) в зависимости от возраста матери и «номера» ребенка

Возраст матери	Ребенок					Грубая доля	Стандартизованная доля*
	1	2	3	4	5+		
Менее 20	46,5	34,6	19,9	43,6	0	42,5	30,4
20—24	42,8	45,9	40,4	37,8	26,1	42,5	39,9
25—29	52,2	52,7	55,1	48,3	51,0	52,3	52,2
30—34	101,3	100,9	87,8	90,5	74,7	87,7	92,9
35—39	274,5	288,1	239,9	297,3	251,7	264,0	270,3
40 и более	819,1	725,6	866,1	976,2	857,8	864,4	830,4
Грубая доля	56,3	67,6	83,3	115,5	167,1	89,5	
Стандартизованная доля**	92,3	91,2	85,1	92,7	75,5		88,0***

* Последняя колонка содержит частные по возрасту матери доли, прямо стандартизованные по числу детей (в качестве стандартного взято распределение по «номеру» ребенка в полной выборке).

** Последняя строка содержит частные по «номеру» ребенка доли, прямо стандартизованные по возрасту матери (в качестве стандартного взято распределение по возрасту матери в полной выборке).

*** Всегда, когда в качестве стандартного распределения берется распределение в полной выборке, общие доли, основанные на двух множествах прямо стандартизованных долей, будут равны друг другу но не обязательно будут равны полной грубой доле.

В пределах любой категории возраста матери нет заметных различий между долями монголизма, частными по «номеру» ребенка у матери. По-видимому, возрастание грубой доли с увеличением числа детей, родившихся ранее у данной матери, является отражением зависимости между возрастом матери и

«номером» ребенка, а отнюдь не прямой связи между последним и риском монголизма.

С другой стороны, видно, что риск монголизма сильно связан с возрастом матери. В пределах каждой категории, соответствующей числу детей у матери, прослеживается четкое увеличение случаев монголизма с увеличением возраста матери.

Прямо стандартизованные доли, приведенные в последней строке и в последней колонке табл. 14.5, служат лишь для того, чтобы подытожить информацию, содержащуюся в 30 частных долях: на риск монголизма сильное влияние оказывает возраст матери, а не число детей. По сути, прямая стандартизация оказалась уместной только из-за систематичности соотношений между частными долями в возрастных категориях (довольно слабая вариабельность частных по «номеру» ребенка долей) и в категориях, соответствующих «номеру» ребенка (заметное увеличение риска с увеличением возраста матери).

Одновременная непрямая стандартизация. Если известны все частные доли (причем все они получены по выборкам, размер которых обеспечивает достаточную точность), то можно применять только что описанный метод одновременной прямой стандартизации. Если же частные доли неизвестны или получены по сравнительно малым выборкам, можно использовать метод Мантела и Старка [Mantel and Stark, 1968], основанный на непрямой стандартизации.

Может случиться, что нет данных, нужных, чтобы вычислить доли, частные одновременно по отношению к возрасту матери и к числу детей у нее. Вместо этого, например, мы располагаем лишь данными табл. 14.6 и грубыми долями для возраста матери и для числа детей у нее.

В отличие от предыдущего случая, когда для обобщенного описания использовались прямо стандартизованные доли, в настоящей ситуации, когда частные доли неизвестны, мы вынуждены применять непрямо стандартизованные доли (см. последнюю строку и последнюю колонку табл. 14.6).

Можно выделить два неприятных свойства непрямой стандартизации. Во-первых, она не дает равных общих долей. Во-вторых, значения непрямо стандартизованных долей могут выходить из диапазона значений частных долей в данной категории (сравните, например, непрямо стандартизованную долю для первой категории возраста матери, равную 62,7, с частными долями в табл. 14.5, принимающими значения от 0 до 46,5).

Мантел и Старк [Mantel and Stark, 1968] в качестве средства контроля этой и, возможно, других аномалий предлагают следующую процедуру.

1. Сначала, используя в качестве стандартного набор грубых долей, частных по возрасту матери, вычислить непрямо стан-

Таблица 14.6

Распределение числа новорожденных в зависимости от возраста матери и «номера» ребенка, полные грубые доли и непрямые стандартизованные доли

Возраст матери	Ребенок					Грубая доля	Стандартизованная доля*
	1	2	3	4	5+		
Менее 20	230061	72202	15050	2293	327	42,5	62,7
20—24	329449	326701	175702	68800	30666	42,5	51,9
25—29	114920	208667	207081	132424	123419	52,3	49,9
30—34	39487	83228	117300	98301	149919	87,7	70,9
35—39	14208	28466	45026	46075	104088	264,0	192,6
40 и более	3052	5375	8660	9834	34392	864,4	582,9
Грубая доля	56,3	67,6	83,3	115,5	167,1	89,5	79,2***
Стандартизованная доля**	93,0	92,7	87,3	94,3	84,8	90,7***	

* Последняя колонка содержит частные по возрасту матери доли, непрямо стандартизованные по числу детей (в качестве стандартного взят набор частных по «номеру» ребенка долей в полной выборке). Так, например,

$$c_{20-24\ (\text{indirect})} = 51,9 = 89,5 \frac{42,5 \cdot 9,31318}{56,3 \cdot 3,29449 + \ldots + 167,1 \cdot 0,30666}.$$

** Последняя строка содержит частные по «номеру» ребенка доли, непрямо стандартизованные по возрасту матери (в качестве стандартного взят набор частных по возрасту матери долей в полной выборке). Так, например,

$$c_{2\ (\text{indirect})} = 92,7 = 89,5 \cdot \frac{67,6 \cdot 7,24639}{42,5 \cdot 0,72202 + \ldots + 864,4 \cdot 0,05375}.$$

*** Общие доли, основанные на двух наборах непрямо стандартизованных долей, почти никогда не будут равны ни друг другу, ни полной грубой доле.

дартизованные доли, частные по «номеру» ребенка. В табл. 14.6 они даны в последней строке.

2. Затем, взяв в качестве стандартного *полученный набор стандартизованных долей,* вычислить непрямо стандартизованные доли, частные по возрасту матери. Умножьте полученные

стандартизованные доли на отношение ожидаемой полной грубой доли (90,7) к наблюдаемой полной грубой доле (89,5):

Возраст матери	менее 20	20—24	25—29	30—34	35—39	40 и более
Стандартизованная доля	41,6	42,0	52,4	89,0	270,3	892,9

Например, число 41,6 было получено следующим образом:

$$\frac{42,5}{92,7}\cdot 89,5\cdot\frac{90,7}{89,5}=41,6,$$

где 42,5 — грубая доля в первой возрастной категории; 90,7 — ожидаемая полная грубая доля [1], а

$$92,7=\frac{93,0\cdot 2,3+92,7\cdot 0,722+\ldots+84,8\cdot 0,00327}{3,2}.$$

3. Продолжать вычисления шагов 1 и 2, каждый раз используя набор стандартизованных долей, найденных для одного фактора на предыдущем шаге, чтобы получить новый набор для другого фактора. Каждую долю нового набора умножить на отношение ожидаемой полной грубой доли, полученной по предыдущему набору, к наблюдаемой полной грубой доле.

4. Завершить процесс, когда наборы долей, вычисленные на новом шаге, совпадут с найденными ранее. В данном случае для этого потребовалось четыре цикла.

5. Наборами долей для возраста матери, которые мы получим в результате, будут:

Возраст матери	менее 20	20—24	25—29	30—34	35—39	40 и более
Стандартизованная доля	41,0	41,8	52,6	89,7	273,3	904,5

и для «номера» ребенка:

Ребенок	1	2	3	4	5
Стандартизованная доля	92,5	93,7	87,3	93,6	83,6

6. Два набора долей, полученных таким образом, характеризуются тем, что при использовании любого из них в качестве стандартного набора будет получен второй набор. Кроме того, в данном примере оказалось, что каждая доля находится в диапазоне значений частных долей, хотя это не всегда имеет место в процедуре стандартизации Мантела — Старка. Наконец, стан-

[1] Напомним, что ожидаемой грубой долей называется величина (14.7). Здесь в качестве стандартизованных частных долей используются непрямо стандартизованные доли, а распределения по каждому фактору вычисляются по данным об общем числе новорожденных, т. е. по данным табл. 14.4 или 11.6. — *Примеч. пер.*

дартизация в этой структуре дает одинаковые ожидаемые полные доли (в примере 91,2) для обоих факторов. Это свойство не выполнялось бы, если бы мы на каждом шагу не умножали полученные доли на отношение полной ожидаемой доли, основанной на наборе долей, который вычислен на предыдущем шаге, к полной наблюдаемой доле.

Выводы по полученным стандартизованным долям такие же, как и раньше: сильное влияние возраста матери и слабое влияние числа детей. Заметим, однако, что в данном случае мы вовсе не использовали частные доли.

Процедура Мантела — Старка обладает свойством, что она дает одинаковые результаты независимо от набора долей, с которого начинают вычисления, хотя число шагов зависит от начального набора. Другой подход к задаче разделения эффектов двух коррелирующих факторов основан на логлинейной или логистической регрессионных моделях [Berry, 1970; Breslow and Day, 1975; Gail, 1978]. Арифметические вычисления в этом подходе сложнее, чем в процедуре Мантела — Старка.

Задачи

14.1. Данные взяты из табл. 2 [Discher and Feinberg, 1969].

Частные (по возрасту) доли заболеваний легких у мужчин, занятых в обрабатывающей промышленности и в службах сервиса

Возрастной интервал	Рабочие		Служащие сервиса	
	число	процент нарушений	число	процент нарушений
20—29	403	2,2	256	4,8
30—39	688	3,2	525	3,2
40—49	683	2,2	599	2,8
50—59	539	6,9	453	6,6
60+	133	12,8	155	9,0

а) Какой вывод вы предпочли бы сделать, сравнивая по частным относительно возраста долям два рода деятельности?

б) Как вы думаете, что можно получить, проводя выравнивание по возрасту, а затем сравнивая полученные стандартизованные доли? Какая информация, по вашему мнению, будет потеряна?

в) Рассмотрите следующие стандартные распределения:

Возрастной интервал	Стандартные распределения		
	1	2	3
20—29	0,25	0,05	0,07
30—39	0,25	0,05	0,75
40—49	0,30	0,10	0,06
50—59	0,10	0,40	0,06
60+	0,10	0,40	0,06

ервое стандартное распределение концентрируется в возрастных груп-
моложе 50 лет, второе — в возрастных группах 50 лет и старше, тре-
ье — в возрастной группе 30—39 лет. Проведите по всем трем стандартным
спределениям прямую стандартизацию и укажите в каждом случае, како-
я величина и характер различия между двумя стандартизованными до-
лями.

г) Используйте теперь в качестве стандарта распределение по возрасту
объединенной исходной выборке. Проведите прямую стандартизацию для
возрастного интервала 20—49 лет и отдельно — для интервала 50 лет и
старше. Каковы ваши выводы после сравнения двух стандартизованных до-
лей в каждой из двух возрастных групп? Согласуются ли теперь они с ис-
ходными данными?

ЛИТЕРАТУРА

Berry, G. (1970). Parametric analysis of disease incidences in multiway tables. *Biometrics*, **26**, 572–579.

Breslow, N. L. and Day, N. E. (1975). Indirect standardization and multiplicative models for rates, with reference to the age adjustment of cancer incidence and relative frequency data. *J. Chronic Dis.*, 28, 289–303.

Chiang, C. L. (1961). Standard error of the age-adjusted death rate. U.S. Department of Health, Education and Welfare: Vital Statistics- Special Reports, **47**, 271–285.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, **24**, 295–313.

Discher, D. P. and Feinberg, H. C. (1969). Screening for chronic pulmonary disease: Survey of 10,000 industrial workers. *Am. J. Public Health*, **59**, 1857–1867.

Doll, R. and Cook, P. (1967). Summarizing indices for comparison of cancer incidence data. *Int. J. Cancer*, **2**, 269–279.

El-Badry, M. A. (1969). Higher female than male mortality in some countries of south Asia: A digest. *J. Am. Stat. Assoc.*, **64**, 1234–1244.

Elveback, L. R. (1966). Discussion of "Indices of mortality and tests of their statistical significance." *Hum. Biol.*, **38**, 322–324.

Gail, M. (1978). The analysis of heterogeneity for indirect standardized mortality ratios. *J. R. Stat. Soc., Ser. A*, **141**, 224–234.

Hemphill, F. M. and Ament, R. P. (1970). Quantitative assessment of subcategory contributions to observed change in relative frequencies. Paper read at annual meeting of American Public Health Association, Houston.

Kalton, G. (1968). Standardization: A technique to control for extraneous variables. *Appl. Stat.*, 17, 118–136.

Keyfitz, N. (1966). Sampling variance of standardized mortality rates. *Hum. Biol.*, **38**, 309–317.

Kilpatrick, S. J. (1963). Mortality comparisons in socio-economic groups. *Appl. Stat.*, **12**, 65–86.

Kitagawa, E. M. (1955). Components of a difference between two rates. *J. Am. Stat. Assoc.*, 50, 1168–1194.

Kitagawa, E. M. (1964). Standardized comparisons in population research. *Demography*, **1**, 296–315.

Kitagawa, E. M. (1966). Theoretical considerations in the selection of a mortality index, and some empirical comparisons. *Hum. Biol.*, **38**, 293–308.

Kupper, L. L., McMichael, A. J., Symons, M. J., and Most, B. M. (1978). On the utility of proportional mortality analysis. *J. Chronic Dis.*, **31**, 15–22.

Mantel, N. and Stark, C. R. (1968). Computation of indirect-adjusted rates in the presence of confounding. *Biometrics*, **24**, 997–1005.

Mausner, J. S. and Bahn, A. K. (1974). *Epidemiology: An introductory text*. Philadelphia: W. B. Saunders.

Miettinen, O. S. (1972). Components of the crude risk ratio. *Am. J. Epidemiol.*, **96**, 168–172.

Spiegelman, M. and Marks, H. H. (1966). Empirical testing of standards for the age adjustment of death rates by the direct method. *Hum. Biol.*, **38**, 280–292.

Stark, C. R. and Mantel, N. (1966). Effects of maternal age and birth order on the risk of mongolism and leukemia. *J. Natl. Cancer Inst.*, 37, 687–698.

Woolsey, T. D. (1959). Adjusted death rates and other indices of mortality. Chapter 4 in F. E. Linder and R. D. Grove. *Vital statistics rates in the United States, 1900–1940.* Washington, D. C.: U. S. Government Printing Office.

Yerushalmy, J. (1951). A mortality index for use in place of the age-adjusted death rate. *Am. J. Public Health,* **41**, 907–922.

Yule, G. U. (1934). On some points relating to vital statistics, more especially statistics of occupation mortality. *J. R. Stat. Soc.*, **97**, 1–72.

Приложения

Таблица А.

Критические значения распределения хи-квадрат

Степени свободы	Уровень значимости				
	.10	.05	.025	.01	
1	2.71	3.84	5.02	6.63	
2	4.61	5.99	7.38	9.21	10.60
3	6.25	7.81	9.35	11.34	12.84
4	7.78	9.49	11.14	13.28	14.86
5	9.24	11.07	12.83	15.09	16.75
6	10.64	12.59	14.45	16.81	18.55
7	12.02	14.07	16.01	18.48	20.28
8	13.36	15.51	17.53	20.09	21.96
9	14.68	16.92	19.02	21.67	23.59
10	15.99	18.31	20.48	23.21	25.19
11	17.28	19.68	21.92	24.72	26.76
12	18.55	21.03	23.34	26.22	28.30
13	19.81	22.36	24.74	27.69	29.82
14	21.06	23.68	26.12	29.14	31.32
15	22.31	25.00	27.49	30.58	32.80
16	23.54	26.30	28.85	32.00	34.27
17	24.77	27.59	30.19	33.41	35.72
18	25.99	28.87	31.53	34.81	37.16
19	27.20	30.14	32.85	36.19	38.58
20	28.41	31.41	34.17	37.57	40.00
25	34.38	37.65	40.65	44.31	46.93
30	40.26	43.77	46.98	50.89	53.67
40	51.80	55.76	59.34	63.69	66.77
60	74.40	79.08	83.30	88.38	91.95
100	118.50	124.34	129.56	135.81	140.17

Приводится с сокращениями по табл. 8 "Biometrika tables 2nd edition," Edited by E. S. Pearson and H. O. Hartley. Camb.. Cambridge, England, 1958.

Таблица А.2

Критические значения нормального распределения

P – площадь хвостов под нормальной кривой ниже $-Z$ и выше $+Z$. При проверке значимости P является уровнем значимости, соответствующим полученному значению Z.

Площадь под нормальной кривой справа от Z равна $1 - P/2$, если Z отрицательно, и $P/2$, если Z положительно.

Допустим, надо найти такое значение Z, при котором площадь под нормальной кривой справа от Z была равна $1 - B$. Если $1 - B$ больше 0,50, возьмите значение Z, соответствующее $P = 2B$, и припишите знак минус этому значению. Если $1 - B$ меньше 0.50. возьмите значение Z, соответствующее $P = 2\,(1 - B)$.

Z	P	Z	P	Z	P
0.0	1.0000	1.2	0.2301	2.4	0.0164
0.1	0.9203	1.282	0.20	2.5	0.0124
0.126	0.90	1.3	0.1936	2.576	0.01
0.2	0.8415	1.4	0.1615	2.6	0.0093
0.3	0.7642	1.440	0.15	2.7	0.0069
0.385	0.70	1.5	0.1336	2.8	0.0051
0.4	0.6892	1.6	0.1096	2.813	0.005
0.5	0.6171	1.645	0.10	2.9	0.0037
0.524	0.60	1.7	0.0891	3.0	0.0027
0.6	0.5485	1.8	0.0719	3.090	0.002
0.674	0.50	1.9	0.0574	3.1	0.0019
0.7	0.4839	1.960	0.05	3.2	0.0014
0.8	0.4237	2.0	0.0455	3.3	0.0010
0.842	0.40	2.1	0.0357	3.4	0.0007
0.9	0.3681	2.2	0.0278	3.5	0.0005
1.0	0.3173	2.242	0.025	3.6	0.0003
1.036	0.30	2.3	0.0214	3.7	0.0002
1.1	0.2713	2.326	0.02	3.8	0.0001

Взято с сокращениями из табл. 1 и 4 "Biometrika tables for statisticians, Vol. I, 2nd edition", Edited by E. S. Pearson and H. O. Hartley, Cambridge University Press, Cambridge, England, 1958.

Таблица А.3

Размер выборки (на одну группу) для критерия проверки различия пропорций

$P_1 = 0.05$

P_2	Уровень значимости	Мощность								
		0.99	0.95	0.90	0.85	0.80	0.75	0.70	0.65	0.50
0.10	0.01	1368	1025	863	762	686	624	572	525	407
	0.02	1235	911	760	665	595	538	489	446	339
	0.05	1054	758	621	536	474	423	381	344	252
	0.10	910	637	513	437	381	336	299	267	188
	0.20	758	512	402	336	288	250	219	192	128
0.15	0.01	447	337	285	253	228	209	192	177	139
	0.02	404	300	252	221	199	180	165	151	117
	0.05	345	251	207	179	160	143	130	118	88
	0.10	299	212	172	147	130	115	103	93	67
	0.20	250	171	136	115	99	87	77	68	47
0.20	0.01	241	183	155	138	125	115	106	98	77
	0.02	218	163	137	121	109	99	91	84	65
	0.05	187	136	113	99	88	79	72	66	50
	0.10	162	115	94	81	72	64	58	52	38
	0.20	135	94	75	64	55	49	44	39	28
0.25	0.01	157	120	102	91	83	76	70	65	52
	0.02	142	107	90	80	72	66	61	56	44
	0.05	122	90	75	65	58	53	48	44	34
	0.10	106	76	62	54	48	43	39	35	26
	0.20	88	62	50	42	37	33	29	26	19
0.30	0.01	113	87	74	66	60	55	51	48	38
	0.02	102	77	66	58	53	48	44	41	33
	0.05	88	65	54	48	43	39	35	32	25
	0.10	76	55	45	39	35	32	29	26	20
	0.20	64	45	36	31	27	24	22	20	14
0.35	0.01	86	66	57	51	46	43	40	37	30
	0.02	78	59	50	45	41	37	34	32	25
	0.05	67	50	42	37	33	30	28	25	20
	0.10	58	42	35	30	27	25	22	20	16
	0.20	48	34	28	24	21	19	17	16	12
0.40	0.01	68	53	45	41	37	34	32	30	24
	0.02	62	47	40	36	33	30	28	26	21
	0.05	53	39	33	29	27	24	22	21	16
	0.10	46	34	28	24	22	20	18	17	13
	0.20	38	27	22	19	17	15	14	13	10
0.45	0.01	55	43	37	33	31	28	26	25	20
	0.02	50	38	33	29	27	25	23	21	17
	0.05	43	32	27	24	22	20	18	17	14
	0.10	37	27	23	20	18	16	15	14	11
	0.20	31	22	18	16	14	13	12	11	8
0.50	0.01	46	36	31	28	26	24	22	21	17
	0.02	41	32	27	25	23	21	19	18	15

Продолжение

P_2	Уровень значимости	Мощность								
		0.99	0.95	0.90	0.85	0.80	0.75	0.70	0.65	0.50
	0.05	35	27	23	20	18	17	16	14	12
	0.10	31	23	19	17	15	14	13	12	9
	0.20	26	19	15	13	12	11	10	9	7
0.55	0.01	38	30	26	24	22	20	19	18	15
	0.02	35	27	23	21	19	18	17	16	13
	0.05	30	23	19	17	16	14	13	12	10
	0.10	26	19	16	14	13	12	11	10	8
	0.20	21	16	13	11	10	9	9	8	6
0.60	0.01	32	26	22	20	19	18	16	16	13
	0.02	29	23	20	18	17	15	14	13	11
	0.05	25	19	16	15	14	12	12	11	9
	0.10	22	16	14	12	11	10	9	9	7
	0.20	18	13	11	10	9	8	7	7	5
0.65	0.01	28	22	19	18	16	15	14	14	11
	0.02	25	20	17	16	14	13	13	12	10
	0.05	21	16	14	13	12	11	10	9	8
	0.10	18	14	12	11	10	9	8	8	6
	0.20	15	11	10	9	8	7	7	6	5
0.70	0.01	23	19	17	15	14	13	13	12	10
	0.02	21	17	15	14	13	12	11	10	9
	0.05	18	14	12	11	10	10	9	8	7
	0.10	16	12	10	9	9	8	7	7	6
	0.20	13	10	8	7	7	6	6	5	4
0.75	0.01	20	16	15	13	13	12	11	11	9
	0.02	18	15	13	12	11	10	10	9	8
	0.05	15	12	11	10	9	8	8	7	6
	0.10	13	10	9	8	7	7	7	6	5
	0.20	11	8	7	6	6	5	5	5	4
0.80	0.01	17	14	13	12	11	10	10	9	8
	0.02	15	13	11	10	10	9	9	8	7
	0.05	13	10	9	8	8	7	7	7	6
	0.10	11	9	8	7	7	6	6	5	5
	0.20	9	7	6	6	5	5	5	4	4
0.85	0.01	15	12	11	10	10	9	9	8	7
	0.02	13	11	10	9	8	8	8	7	6
	0.05	11	9	8	7	7	7	6	6	5
	0.10	9	8	7	6	6	5	5	5	4
	0.20	8	6	5	5	5	4	4	4	3
0.90	0.01	12	10	9	9	8	8	8	7	7
	0.02	11	9	8	8	7	7	7	7	6
	0.05	9	8	7	6	6	6	6	5	5
	0.10	8	6	6	5	5	5	5	4	4
	0.20	6	5	5	4	4	4	4	3	3
0.95	0.01	10	9	8	8	7	7	7	7	6

Продолжени

P_2	Уровень значимости	Мощность								
		0.99	0.95	0.90	0.85	0.80	0.75	0.70	0.65	0.50
	0.02	9	8	7	7	7	6	6	6	5
	0.05	8	6	6	6	5	5	5	5	4
	0.10	6	5	5	5	4	4	4	4	4
	0.20	5	4	4	4	4	3	3	3	3
					$P_1 = 0.10$					
0.15	0.01	2137	1595	1340	1179	1060	963	880	806	620
	0.02	1928	1416	1176	1027	916	826	749	682	513
	0.05	1642	1174	957	823	725	646	579	520	375
	0.10	1415	984	787	667	579	509	450	399	275
	0.20	1175	787	613	508	433	373	324	281	182
0.20	0.01	627	471	397	351	316	288	264	243	189
	0.02	566	419	349	306	274	248	226	206	157
	0.05	483	348	286	247	219	196	176	159	117
	0.10	417	293	236	201	176	156	139	124	88
	0.20	347	236	185	155	133	116	102	89	60
0.25	0.01	316	238	202	179	162	147	136	125	98
	0.02	285	212	178	156	140	127	116	107	82
	0.05	244	177	146	127	113	101	91	83	62
	0.10	211	149	121	104	91	81	73	65	47
	0.20	176	120	96	81	70	61	54	48	33
0.30	0.01	196	149	126	112	102	93	86	79	63
	0.02	178	133	112	98	89	81	74	68	53
	0.05	152	111	92	80	71	64	58	53	40
	0.10	131	94	76	66	58	52	47	42	31
	0.20	110	76	61	51	45	39	35	31	22
0.35	0.01	136	104	88	79	72	66	61	56	45
	0.02	123	93	78	69	62	57	52	48	38
	0.05	105	77	64	56	50	45	41	38	29
	0.10	91	65	54	46	41	37	33	30	22
	0.20	76	53	43	36	32	28	25	22	16
0.40	0.01	101	77	66	59	54	49	46	42	34
	0.02	91	69	58	52	47	43	39	36	29
	0.05	78	58	48	42	38	34	31	29	22
	0.10	68	49	40	35	31	28	25	23	17
	0.20	56	40	32	27	24	21	19	17	13
0.45	0.01	78	60	51	46	42	39	36	33	27
	0.02	71	54	46	41	37	34	31	29	23
	0.05	60	45	38	33	30	27	25	23	18
	0.10	52	38	31	27	24	22	20	18	14
	0.20	44	31	25	22	19	17	15	14	10
0.50	0.01	62	48	41	37	34	31	29	27	22
	0.02	56	43	37	33	30	27	25	23	19
	0.05	48	36	30	27	24	22	20	19	15
	0.10	42	30	25	22	20	18	16	15	12
	0.20	35	25	20	18	16	14	13	11	9
0.55	0.01	51	39	34	31	28	26	24	23	19

Продолжение

P_2	Уровень значимости	Мощность								
		0.99	0.95	0.90	0.85	0.80	0.75	0.70	0.65	0.50
	0.02	46	35	30	27	25	23	21	20	16
	0.05	39	29	25	22	20	18	17	16	12
	0.10	34	25	21	18	16	15	14	13	10
	0.20	28	20	17	15	13	12	11	10	7
0.60	0.01	42	33	28	26	24	22	20	19	16
	0.02	38	29	25	23	21	19	18	17	14
	0.05	32	24	21	18	17	15	14	13	11
	0.10	28	21	17	15	14	13	12	11	8
	0.20	23	17	14	12	11	10	9	8	6
0.65	0.01	35	27	24	22	20	19	17	16	14
	0.02	31	24	21	19	18	16	15	14	12
	0.05	27	20	18	16	14	13	12	11	9
	0.10	23	17	15	13	12	11	10	9	7
	0.20	19	14	12	10	9	8	8	7	6
0.70	0.01	29	23	20	19	17	16	15	14	12
	0.02	26	21	18	16	15	14	13	12	10
	0.05	22	17	15	13	12	11	11	10	8
	0.10	19	15	13	11	10	9	9	8	7
	0.20	16	12	10	9	8	7	7	6	5
0.75	0.01	25	20	17	16	15	14	13	12	11
	0.02	22	18	15	14	13	12	11	11	9
	0.05	19	15	13	12	11	10	9	9	7
	0.10	16	13	11	10	9	8	8	7	6
	0.20	14	10	9	8	7	6	6	6	4
0.80	0.01	21	17	15	14	13	12	11	11	9
	0.02	19	15	13	12	11	11	10	9	8
	0.05	16	13	11	10	9	9	8	8	6
	0.10	14	11	9	8	8	7	7	6	5
	0.20	11	9	7	7	6	6	5	5	4
0.85	0.01	18	14	13	12	11	11	10	10	8
	0.02	16	13	11	11	10	9	9	8	7
	0.05	13	11	9	9	8	8	7	7	6
	0.10	11	9	8	7	7	6	6	6	5
	0.20	10	7	6	6	5	5	5	4	4
0.90	0.01	15	12	11	10	10	9	9	8	7
	0.02	13	11	10	9	9	8	8	7	6
	0.05	11	9	8	7	7	7	6	6	5
	0.10	10	8	7	6	6	5	5	5	4
	0.20	8	6	6	5	5	4	4	4	3
0.95	0.01	12	10	9	9	8	8	8	7	7
	0.02	11	9	8	8	7	7	7	7	6
	0.05	9	8	7	6	6	6	6	5	5
	0.10	8	6	6	5	5	5	5	4	4
	0.20	6	5	5	4	4	4	4	3	3

Продолжение

$P_1 = 0.15$

P_2	Уровень значимости	Мощность								
		0.99	0.95	0.90	0.85	0.80	0.75	0.70	0.65	0.50
0.20	0.01	2810	2094	1756	1545	1388	1259	1149	1052	806
	0.02	2534	1858	1541	1343	1198	1078	977	888	664
	0.05	2157	1538	1252	1075	945	840	751	674	483
	0.10	1856	1287	1027	868	753	660	582	515	351
	0.20	1539	1027	797	659	559	480	415	360	228
0.25	0.01	783	586	494	435	392	357	326	300	232
	0.02	707	521	434	380	340	307	279	254	193
	0.05	603	433	354	305	270	241	216	195	142
	0.10	520	363	292	248	216	191	169	151	106
	0.20	432	291	228	190	163	141	123	108	71
0.30	0.01	380	286	241	213	193	176	161	148	116
	0.02	343	254	213	187	167	152	138	126	97
	0.05	293	212	174	151	134	120	108	98	72
	0.10	253	178	144	123	108	95	85	76	54
	0.20	211	143	113	95	82	71	83	55	38
0.35	0.01	229	173	147	130	118	108	99	91	72
	0.02	207	154	130	114	102	93	85	78	60
	0.05	177	129	106	92	82	74	67	61	45
	0.10	153	109	88	76	67	59	53	48	35
	0.20	128	88	70	59	51	45	39	35	24
0.40	0.01	155	118	100	89	81	74	68	63	50
	0.02	140	105	89	78	70	64	59	54	42
	0.05	120	88	73	63	57	51	46	42	32
	0.10	104	74	60	52	46	41	37	33	25
	0.20	87	60	48	41	35	31	28	25	18
0.45	0.01	113	86	73	65	60	55	50	47	37
	0.02	102	77	65	57	52	47	44	40	32
	0.05	87	64	53	47	42	38	34	32	24
	0.10	75	54	44	39	34	31	28	25	19
	0.20	63	44	35	30	26	23	21	19	14
0.50	0.01	86	66	56	50	46	42	39	36	29
	0.02	78	59	50	44	40	37	34	31	25
	0.05	66	49	41	36	32	29	27	25	19
	0.10	57	42	34	30	26	24	22	20	15
	0.20	48	34	27	23	21	18	16	15	11
0.55	0.01	67	52	45	40	37	34	31	29	24
	0.02	61	46	39	35	32	29	27	25	20
	0.05	52	39	33	29	26	24	22	20	16
	0.10	45	33	27	24	21	19	17	16	12
	0.20	38	27	22	19	17	15	13	12	9
0.60	0.01	54	42	36	33	30	28	26	24	20
	0.02	49	37	32	29	26	24	22	21	17
	0.05	42	31	26	23	21	19	18	16	13
	0.10	36	27	22	19	17	16	14	13	10
	0.20	30	22	18	15	14	12	11	10	8

Продолжение

f_2	Уровень значимости	Мощность								
		0.99	0.95	0.90	0.85	0.80	0.75	0.70	0.65	0.50
0.65	0.01	44	34	30	27	25	23	21	20	16
	0.02	40	31	26	24	22	20	19	17	14
	0.05	34	26	22	19	18	16	15	14	11
	0.10	29	22	18	16	15	13	12	11	9
	0.20	25	18	15	13	11	10	9	9	7
0.70	0.01	36	29	25	23	21	19	18	17	14
	0.02	33	26	22	20	18	17	16	15	12
	0.05	28	21	18	16	15	14	13	12	9
	0.10	24	18	15	14	12	11	10	10	8
	0.20	20	15	12	11	10	9	8	7	6
0.75	0.01	30	24	21	19	18	16	15	15	12
	0.02	27	21	19	17	16	14	13	13	11
	0.05	23	18	15	14	13	12	11	10	8
	0.10	20	15	13	11	10	10	9	8	7
	0.20	17	12	10	9	8	8	7	6	5
0.80	0.01	25	20	18	16	15	14	13	13	11
	0.02	23	18	16	14	13	12	12	11	9
	0.05	19	15	13	12	11	10	9	9	7
	0.10	17	13	11	10	9	8	8	7	6
	0.20	14	10	9	8	7	7	6	6	4
0.85	0.01	21	17	15	14	13	12	12	11	9
	0.02	19	15	13	12	11	11	10	10	8
	0.05	16	13	11	10	9	9	8	8	6
	0.10	14	11	9	8	8	7	7	6	5
	0.20	12	9	8	7	6	6	5	5	4
0.90	0.01	18	14	13	12	11	11	10	10	8
	0.02	16	13	11	11	10	9	9	8	7
	0.05	13	11	9	9	8	8	7	7	6
	0.10	11	9	8	7	7	6	6	6	5
	0.20	10	7	6	6	5	5	5	4	4
0.95	0.01	15	12	11	10	10	9	9	8	7
	0.02	13	11	10	9	8	8	8	7	6
	0.05	11	9	8	7	7	7	6	6	5
	0.10	9	8	7	6	6	5	5	5	4
	0.20	8	6	5	5	5	4	4	4	3
					$P_1 = 0.20$					
0.25	0.01	3386	2522	2114	1858	1668	1512	1379	1262	965
	0.02	3053	2236	1853	1615	1438	1294	1172	1064	794
	0.05	2597	1850	1504	1290	1134	1007	900	806	575
	0.10	2235	1547	1233	1041	901	789	695	614	417
	0.20	1852	1232	955	788	668	572	494	426	268
0.30	0.01	915	685	576	507	456	415	379	348	268
	0.02	826	608	506	442	395	356	324	295	222
	0.05	704	504	412	355	313	279	250	225	163
	0.10	607	423	339	288	250	220	195	174	121
	0.20	504	339	265	220	188	162	141	123	80
0.35	0.01	433	325	274	242	219	199	183	168	131
	0.02	391	289	242	212	190	172	156	143	109

Продолжение

P_2	Уровень значимости	Мощность								
		0.99	0.95	0.90	0.85	0.80	0.75	0.70	0.65	0.50
	0.05	334	240	197	171	151	135	122	110	81
	0.10	288	202	163	139	122	107	96	85	61
	0.20	240	162	128	107	92	80	70	62	41
0.40	0.01	256	193	164	145	131	120	110	101	79
	0.02	232	172	144	127	114	103	94	86	66
	0.05	198	143	118	102	91	82	74	67	50
	0.10	171	121	98	84	74	65	58	52	38
	0.20	142	97	77	65	56	49	43	38	26
0.45	0.01	171	129	110	98	88	81	74	69	54
	0.02	154	115	97	85	77	70	64	59	46
	0.05	132	96	79	69	62	56	50	46	35
	0.10	114	81	66	57	50	45	40	36	26
	0.20	95	65	52	44	38	34	30	27	19
0.50	0.01	122	93	79	71	64	59	54	50	40
	0.02	110	83	70	62	56	51	47	43	34
	0.05	94	69	57	50	45	41	37	34	26
	0.10	82	58	48	41	37	33	29	27	20
	0.20	68	47	38	32	28	25	22	20	14
0.55	0.01	92	70	60	54	49	45	41	38	31
	0.02	83	63	53	47	43	39	36	33	26
	0.05	71	52	44	38	34	31	28	26	20
	0.10	61	44	36	32	28	25	23	21	16
	0.20	51	36	29	25	22	19	17	15	11
0.60	0.01	71	55	47	42	38	35	33	31	25
	0.02	64	49	42	37	34	31	28	26	21
	0.05	55	41	34	30	27	25	23	21	16
	0.10	47	35	29	25	22	20	18	17	13
	0.20	40	28	23	20	17	15	14	13	9
0.65	0.01	56	44	38	34	31	29	27	25	20
	0.02	51	39	33	30	27	25	23	21	17
	0.05	44	33	27	24	22	20	18	17	13
	0.10	38	28	23	20	18	16	15	14	11
	0.20	31	22	18	16	14	13	11	10	8
0.70	0.01	46	36	31	28	25	24	22	21	17
	0.02	41	32	27	24	22	21	19	18	14
	0.05	35	26	22	20	18	17	15	14	11
	0.10	30	22	19	17	15	14	12	11	9
	0.20	25	18	15	13	12	11	10	9	7
0.75	0.01	37	29	25	23	21	20	18	17	14
	0.02	34	26	23	20	19	17	16	15	12
	0.05	29	22	19	17	15	14	13	12	10
	0.10	25	18	16	14	12	11	10	10	8
	0.20	21	15	12	11	10	9	8	7	6
0.80	0.01	31	24	21	19	18	17	16	15	12

Прооолжение

..ь ости	Мощность								
	0.99	0.95	0.90	0.85	0.80	0.75	0.70	0.65	0.50
.02	28	22	19	17	16	15	14	13	11
.05	23	18	16	14	13	12	11	10	8
.10	20	15	13	12	11	10	9	8	7
.20	17	12	10	9	8	8	7	6	5
.01	25	20	18	16	15	14	13	13	11
.02	23	18	16	14	13	12	12	11	9
.05	19	15	13	12	11	10	9	9	7
.10	17	13	11	10	9	8	8	7	6
.20	14	10	9	8	7	7	6	6	4
.01	21	17	15	14	13	12	11	11	9
.02	19	15	13	12	11	11	10	9	8
.05	16	13	11	10	9	9	8	8	6
.10	14	11	9	8	8	7	7	6	5
.20	11	9	7	7	6	6	5	5	4
.01	17	14	13	12	11	10	10	9	8
.02	15	13	11	10	10	9	9	8	7
.05	13	10	9	8	8	7	7	7	6
.10	11	9	8	7	7	6	6	5	5
.20	9	7	6	6	5	5	5	4	4
	$P_1 = 0.25$								
.01	3867	2878	2411	2119	1902	1723	1572	1438	1098
.02	3486	2552	2114	1841	1639	1474	1334	1211	902
.05	2965	2110	1714	1470	1291	1145	1023	916	652
.10	2550	1764	1404	1185	1025	897	789	696	471
.20	2112	1404	1087	895	758	649	559	482	301
.01	1023	765	643	566	509	462	423	387	298
.02	923	679	564	493	440	397	360	328	247
.05	786	563	459	395	348	310	278	250	181
.10	678	472	378	320	278	245	217	192	133
.20	562	377	294	244	208	179	156	136	88
.01	476	357	301	266	240	218	200	183	142
.02	430	317	265	232	207	188	171	156	118
.05	366	264	216	187	165	147	133	120	88
.10	316	221	178	152	133	117	104	93	65
.20	263	178	140	117	100	87	76	67	44
.01	278	209	177	156	141	129	118	109	85
.02	251	186	156	137	123	111	101	93	71
.05	214	155	127	110	98	88	79	72	53
.10	185	130	105	90	79	70	63	56	40
.20	154	105	83	70	60	52	46	41	28
.01	182	138	117	104	94	86	79	73	57
.02	165	123	103	91	82	74	68	62	48
.05	141	102	85	74	65	59	53	48	36
.10	121	86	70	60	53	47	42	38	28
.20	101	70	55	47	41	36	31	28	20
.01	129	98	83	74	67	62	57	53	42
.02	117	87	74	65	59	53	49	45	35
.05	99	73	60	53	47	43	39	35	27
.10	86	61	50	43	38	34	31	28	21

Прооолжение

P_2	Уровень значимости	Мощность								
		0.99	0.95	0.90	0.85	0.80	0.75	0.70	0.65	0.50
	0.20	72	50	40	34	29	26	23	21	15
0.60	0.01	96	73	62	56	51	47	43	40	32
	0.02	86	65	55	49	44	40	37	34	27
	0.05	74	54	45	40	36	32	29	27	21
	0.10	64	46	38	33	29	26	24	21	16
	0.20	53	37	30	26	22	20	18	16	12
0.65	0.01	73	56	48	43	39	36	34	31	25
	0.02	66	50	43	38	34	32	29	27	21
	0.05	57	42	35	31	28	25	23	21	17
	0.10	49	36	29	26	23	21	19	17	13
	0.20	41	29	23	20	18	16	14	13	9
0.70	0.01	58	45	38	34	32	29	27	25	21
	0.02	52	40	34	30	28	25	23	22	17
	0.05	44	33	28	25	22	20	19	17	14
	0.10	38	28	23	20	18	17	15	14	11
	0.20	32	23	19	16	14	13	12	10	8
0.75	0.01	46	36	31	28	26	24	22	21	17
	0.02	42	32	27	25	22	21	19	18	15
	0.05	35	27	23	20	18	17	15	14	11
	0.10	31	23	19	17	15	14	12	11	9
	0.20	26	18	15	13	12	11	10	9	7
0.80	0.01	37	29	25	23	21	20	18	17	14
	0.02	34	26	23	20	19	17	16	15	12
	0.05	29	22	19	17	15	14	13	12	10
	0.10	25	18	16	14	12	11	10	10	8
	0.20	21	15	12	11	10	9	8	7	6
0.85	0.01	30	24	21	19	18	16	15	15	12
	0.02	27	21	19	17	16	14	13	13	11
	0,05	23	18	15	14	13	12	11	10	8
	0.10	20	15	13	11	10	10	9	8	7
	0.20	17	12	10	9	8	8	7	6	5
0.90	0.01	25	20	17	16	15	14	13	12	11
	0.02	22	18	15	14	13	12	11	11	9
	0.05	19	15	13	12	11	10	9	9	7
	0.10	16	13	11	10	9	8	8	7	6
	0.20	14	10	9	8	7	6	6	6	4
0.95	0.01	20	16	15	13	13	12	11	11	9
	0.02	18	15	13	12	11	10	10	9	8
	0.05	15	12	11	10	9	8	8	7	6
	0.10	13	10	9	8	7	7	7	6	5
	0.20	11	8	7	6	6	5	5	5	4
				$P_1 = 0.30$						
0.35	0.01	4251	3163	2650	2328	2089	1892	1726	1578	1204
	0.02	3832	2804	2322	2022	1800	1618	1464	1329	989
	0.05	3259	2318	1882	1613	1416	1256	1122	1004	714
	0.10	2803	1937	1541	1300	1124	983	865	762	514
	0.20	2320	1541	1192	981	830	710	611	527	327
0.40	0.01	1108	827	695	612	550	499	456	418	322
	0.02	999	734	610	532	475	428	389	354	266

Продолжение

P_2	Уровень значимости	Мощность								
		0.99	0.95	0.90	0.85	0.80	0.75	0.70	0.65	0.50
	0.05	851	608	496	427	376	334	300	269	194
	0.10	733	510	408	345	300	264	233	207	142
	0.20	608	407	317	263	224	193	167	145	94
0.45	0.01	508	381	321	283	255	232	213	195	151
	0.02	459	338	282	247	221	200	182	166	126
	0.05	391	281	230	199	175	157	141	127	93
	0.10	337	236	190	161	141	124	110	98	69
	0.20	280	189	148	124	106	92	80	70	47
0.50	0.01	293	220	186	165	149	135	124	114	89
	0.02	264	196	164	144	129	117	106	97	75
	0.05	225	163	134	116	103	92	83	75	56
	0.10	194	137	111	95	83	73	66	59	42
	0.20	162	110	87	73	63	55	48	42	29
0.55	0.01	190	144	122	108	98	89	82	76	60
	0.02	172	128	107	94	85	77	71	65	50
	0.05	147	106	88	76	68	61	55	50	38
	0.10	127	90	73	63	55	49	44	39	29
	0.20	105	72	57	48	42	37	33	29	20
0.60	0.01	133	101	86	76	69	63	59	54	43
	0.02	120	90	76	67	60	55	50	46	36
	0.05	103	75	62	54	48	44	40	36	27
	0.10	89	63	52	45	39	35	32	29	21
	0.20	74	51	41	35	30	27	24	21	15
0.65	0.01	98	75	64	57	52	47	44	41	32
	0.02	88	66	56	50	45	41	38	35	27
	0.05	75	55	46	40	36	33	30	27	21
	0.10	65	47	38	33	30	27	24	22	16
	0.20	54	38	31	26	23	20	18	16	12
0.70	0.01	74	57	49	44	40	37	34	32	25
	0.02	67	51	43	38	35	32	29	27	22
	0.05	57	42	36	31	28	26	23	21	17
	0.10	49	36	30	26	23	21	19	17	13
	0.20	41	29	24	20	18	16	14	13	9
0.75	0.01	58	45	38	34	32	29	27	25	21
	0.02	52	40	34	30	28	25	23	22	17
	0.05	44	33	28	25	22	20	19	17	14
	0.10	38	28	23	20	18	17	15	14	11
	0.20	32	23	19	16	14	13	12	10	8
0.80	0.01	46	36	31	28	25	24	22	21	17
	0.02	41	32	27	24	22	21	19	18	14
	0.05	35	26	22	20	18	17	15	14	11
	0.10	30	22	19	17	15	14	12	11	9
	0.20	25	18	15	13	12	11	10	9	7
0.85	0.01	36	29	25	23	21	19	18	17	14
	0.02	33	26	22	20	18	17	16	15	12
	0.05	28	21	18	16	15	14	13	12	9
	0.10	24	18	15	14	12	11	10	10	8
	0.20	20	15	12	11	10	9	8	7	6

Продолжение

P_2	Уровень значимости	Мощность								
		0.99	0.95	0.90	0.85	0.80	0.75	0.70	0.65	0.50
0.90	0.01	29	23	20	19	17	16	15	14	12
	0.02	26	21	18	16	15	14	13	12	10
	0.05	22	17	15	13	12	11	11	10	8
	0.10	19	15	13	11	10	9	9	8	7
	0.20	16	12	10	9	8	7	7	6	5
0.95	0.01	23	19	17	15	14	13	13	12	10
	0.02	21	17	15	14	13	12	11	10	9
	0.05	18	14	12	11	10	10	9	8	7
	0.10	16	12	10	9	9	8	7	7	6
	0.20	13	10	8	7	7	6	6	5	4
		$P_1 = 0.35$								
0.40	0.01	4540	3377	2828	2484	2229	2019	1841	1683	1284
	0.02	4092	2993	2478	2157	1920	1726	1562	1417	1054
	0.05	3479	2474	2008	1721	1511	1340	1196	1070	760
	0.10	2992	2067	1644	1386	1198	1047	921	812	547
	0.20	2476	1644	1271	1046	885	756	650	560	347
0.45	0.01	1168	872	732	644	579	526	480	440	338
	0.02	1053	774	642	561	500	451	409	372	279
	0.05	897	641	522	449	395	352	315	283	204
	0.10	772	537	429	363	316	277	245	217	149
	0.20	640	429	334	276	235	202	176	152	98
0.50	0.01	530	397	334	295	265	241	221	203	157
	0.02	478	352	294	257	230	208	189	172	131
	0.05	407	293	239	207	182	163	146	132	96
	0.10	351	246	197	168	146	129	115	102	71
	0.20	292	197	154	128	110	95	83	73	48
0.55	0.01	302	227	192	169	153	139	128	118	92
	0.02	272	202	169	148	133	120	110	100	77
	0.05	232	168	138	119	106	95	85	77	57
	0.10	200	141	114	97	85	75	67	60	43
	0.20	167	113	89	75	65	56	49	43	29
0.60	0.01	194	147	124	110	100	91	84	77	61
	0.02	175	130	109	96	87	79	72	66	51
	0.05	149	109	90	78	69	62	56	51	38
	0.10	129	91	74	64	56	50	45	40	29
	0.20	108	74	59	49	43	37	33	29	20
0.65	0.01	134	102	87	77	70	64	59	55	43
	0.02	121	91	76	68	61	55	51	47	36
	0.05	104	76	63	55	49	44	40	36	28
	0.10	89	64	52	45	40	35	32	29	21
	0.20	75	52	41	35	30	27	24	21	15
0.70	0.01	98	75	64	57	52	47	44	41	32
	0.02	88	66	56	50	45	41	38	35	27
	0.05	75	55	46	40	36	33	30	27	21
	0.10	65	47	38	33	30	27	24	22	16
	0.20	54	38	31	26	23	20	18	16	12
0.75	0.01	73	56	48	43	39	36	34	31	25
	0.02	66	50	43	38	34	32	29	27	21

Продолжение

P_2	Уровень значимости	Мощность								
		0.99	0.95	0.90	0.85	0.80	0.75	0.70	0.65	0.50
	0.05	57	42	35	31	28	25	23	21	17
	0.10	49	36	29	26	23	21	19	17	13
	0.20	41	29	23	20	18	16	14	13	9
0.80	0.01	56	34	38	34	31	29	27	25	20
	0.02	51	39	33	30	27	25	23	21	17
	0.05	44	33	27	24	22	20	18	17	13
	0.10	38	28	23	20	18	16	15	14	11
	0.20	31	22	18	16	14	13	11	10	8
0.85	0.01	44	34	30	27	25	23	21	20	16
	0.02	40	31	26	24	22	20	19	17	14
	0.05	34	26	22	19	18	16	15	14	11
	0.10	29	22	18	16	15	13	12	11	9
	0.20	25	18	15	13	11	10	9	9	7
0.90	0.01	35	27	24	22	20	19	17	16	14
	0.02	31	24	21	19	18	16	15	14	12
	0.05	27	20	18	16	14	13	12	11	9
	0.10	23	17	15	13	12	11	10	9	7
	0.20	19	14	12	10	9	8	8	7	6
0.95	0.01	28	22	19	18	16	15	14	14	11
	0.02	25	20	17	16	14	13	13	12	10
	0.05	21	16	14	13	12	11	10	9	8
	0.10	18	14	12	11	10	9	8	8	6
	0.20	15	11	10	9	8	7	7	6	5
					$P_1 = 0.40$					
0.45	0.01	4732	3520	2947	2589	2322	2104	1918	1753	1337
	0.02	4265	3119	2582	2248	2001	1798	1627	1476	1097
	0.05	3626	2578	2093	1793	1573	1395	1245	1114	791
	0.10	3118	2153	1713	1444	1248	1090	959	845	568
	0.20	2581	1712	1323	1089	921	787	677	582	360
0.50	0.01	1204	898	754	664	597	542	495	453	348
	0.02	1086	797	662	578	515	464	421	383	287
	0.05	924	660	538	463	407	362	324	291	210
	0.10	796	553	442	374	325	285	252	223	153
	0.20	660	441	343	284	242	208	180	157	100
0.55	0.01	540	405	341	301	271	246	225	207	160
	0.02	488	359	299	262	234	212	192	175	133
	0.05	415	298	244	211	186	166	149	134	98
	0.10	358	250	201	171	149	131	117	104	73
	0.20	298	201	157	131	112	97	85	74	49
0.60	0.01	305	229	193	171	154	141	129	119	93
	0.02	275	204	170	149	134	121	111	101	77
	0.05	235	169	139	120	107	96	86	78	58
	0.10	202	142	115	98	86	76	68	61	43
	0.20	168	114	90	76	65	57	50	44	30
0.65	0.01	194	147	124	110	100	91	84	77	61
	0.02	175	130	109	96	87	79	72	66	51
	0.05	149	109	90	78	69	62	56	51	38
	0.10	129	91	74	64	56	50	45	40	29

Продолжение

P_2	Уровень значимости	Мощность								
		0.99	0.95	0.90	0.85	0.80	0.75	0.70	0.65	0.50
	0.20	108	74	59	49	43	37	33	29	20
0.70	0.01	133	101	86	76	69	63	59	54	43
	0.02	120	90	76	67	60	55	50	46	36
	0.05	103	75	62	54	48	44	40	36	27
	0.10	89	63	52	45	39	35	32	29	21
	0.20	74	51	41	35	30	27	24	21	15
0.75	0.01	96	73	62	56	51	47	43	40	32
	0.02	86	65	55	49	44	40	37	34	27
	0.05	74	54	45	40	36	32	29	27	21
	0.10	64	46	38	33	29	26	24	21	16
	0.20	53	37	30	26	22	20	18	16	12
0.80	0.01	71	55	47	42	38	35	33	31	25
	0.02	64	49	42	37	34	31	28	26	21
	0.05	55	41	34	30	27	25	23	21	16
	0.10	47	35	29	25	22	20	18	17	13
	0.20	40	28	23	20	17	15	14	13	9
0.85	0.01	54	42	36	33	30	28	26	24	20
	0.02	49	37	32	29	26	24	22	21	17
	0.05	42	31	26	23	21	19	18	16	13
	0.10	36	27	22	19	17	16	14	13	10
	0.20	30	22	18	15	14	12	11	10	8
0.90	0.01	42	33	28	26	24	22	20	19	16
	0.02	38	29	25	23	21	19	18	17	14
	0.05	32	24	21	18	17	15	14	13	11
	0.10	28	21	17	15	14	13	12	11	8
	0.20	23	17	14	12	11	10	9	8	6
0.95	0.01	32	26	22	20	19	18	16	16	13
	0.02	29	23	20	18	17	15	14	13	11
	0.05	25	19	16	15	14	12	12	11	9
	0.10	27	16	14	12	11	10	9	9	7
	0.20	18	13	11	10	9	8	7	7	5
					$P_1 = 0.45$					
0.50	0.01	4828	3591	3007	2641	2369	2146	1956	1788	1364
	0.02	4352	3182	2635	2293	2041	1834	1659	1505	1119
	0.05	3700	2630	2135	1829	1605	1423	1270	1136	806
	0.10	3181	2197	1747	1472	1273	1112	978	861	579
	0.20	2633	1747	1350	1110	939	802	690	593	367
0.55	0.01	1216	907	762	670	603	547	499	458	352
	0.02	1097	805	668	583	520	469	425	387	290
	0.05	933	667	543	467	411	366	328	294	212
	0.10	804	558	446	378	328	288	254	225	155
	0.20	667	446	347	287	244	210	182	158	101
0.60	0.01	540	405	341	301	271	246	225	207	160
	0.02	488	359	299	262	234	212	192	175	133
	0.05	415	298	244	211	186	166	149	134	98
	0.10	358	250	201	171	149	131	117	104	73
	0.20	298	201	157	131	112	97	85	74	49
0.65	0.01	302	227	192	169	153	139	128	118	92
	0.02	272	202	169	148	133	120	110	100	77

Продолжение

P2	Уровень значимости	Мощность								
		0.99	0.95	0.90	0.85	0.80	0.75	0.70	0.65	0.50
	0.05	232	168	138	119	106	95	85	77	57
	0.10	200	141	114	97	85	75	67	60	43
	0.20	167	113	89	75	65	56	49	43	29
0.70	0.01	190	144	122	108	98	89	82	76	60
	0.02	172	128	107	94	85	77	71	65	50
	0.05	147	106	88	76	68	61	55	50	38
	0.10	127	90	73	63	55	49	44	39	29
	0.20	105	72	57	48	42	37	33	29	20
0.75	0.01	129	98	83	74	67	62	57	53	42
	0.02	117	87	74	65	59	53	49	45	35
	0.05	99	73	60	53	47	43	39	35	27
	0.10	86	61	50	43	38	34	31	28	21
	0.20	72	50	40	34	29	26	23	21	15
0.80	0.01	92	70	60	54	49	45	41	38	31
	0.02	83	63	53	47	43	39	36	33	26
	0.05	71	52	44	38	34	31	28	26	20
	0.10	61	44	36	32	28	25	23	21	16
	0.20	51	36	29	25	22	19	17	15	11
0.85	0.01	67	52	45	40	37	34	31	29	24
	0.02	61	46	39	35	32	29	27	25	20
	0.05	52	39	33	29	26	24	22	20	16
	0.10	45	33	27	24	21	19	17	16	12
	0.20	38	27	22	19	17	15	13	12	9
0.90	0.01	51	39	34	31	28	26	24	23	19
	0.02	46	35	30	27	25	23	21	20	16
	0.05	39	29	25	22	20	18	17	16	12
	0.10	34	25	21	18	16	15	14	13	10
	0.20	28	20	17	15	13	12	11	10	7
0.95	0.01	38	30	26	24	22	20	19	18	15
	0.02	35	27	23	21	19	18	17	16	13
	0.05	30	23	19	17	16	14	13	12	10
	0.10	26	19	16	14	13	12	11	10	8
	0.20	21	16	13	11	10	9	9	8	6

Продолжение

$P_1 = 0.50$

P_2	Уровень значимости	Мощность								
		0.99	0.95	0.90	0.85	0.80	0.75	0.70	0.65	0.50
0.55	0.01	4828	3591	3007	2641	2369	2146	1956	1788	1364
	0.02	4352	3182	2635	2293	2041	1834	1659	1505	1119
	0.05	3700	2630	2135	1829	1605	1423	1270	1136	806
	0.10	3181	2197	1747	1472	1273	1112	978	861	579
	0.20	2633	1747	1350	1110	939	802	690	593	367
0.60	0.01	1204	898	754	664	597	542	495	453	348
	0.02	1086	797	662	578	515	464	421	383	287
	0.05	924	660	538	463	407	362	324	291	210
	0.10	796	553	442	374	325	285	252	223	153
	0.20	660	441	343	284	242	208	180	157	100
0.65	0.01	530	397	334	295	265	241	221	203	157
	0.02	478	352	294	257	230	208	189	172	131
	0.05	407	293	239	207	182	163	146	132	96
	0.10	351	246	197	168	146	129	115	102	71
	0.20	292	197	154	128	110	95	83	73	48
0.70	0.01	293	220	186	165	149	135	124	114	89
	0.02	264	196	164	144	129	117	106	97	75
	0.05	225	163	134	116	103	92	83	75	56
	0.10	194	137	111	95	83	73	66	59	42
	0.20	162	110	87	73	63	55	48	42	29
0.75	0.01	182	138	117	104	94	86	79	73	57
	0.02	165	123	103	91	82	74	68	62	48
	0.05	141	102	85	74	65	59	53	48	36
	0.10	121	86	70	60	53	47	42	38	28
	0.20	101	70	55	47	41	36	31	28	20
0.80	0.01	122	93	79	71	64	59	54	50	40
	0.02	110	83	70	62	56	51	47	43	34
	0.05	94	69	57	50	45	41	37	34	26
	0.10	82	58	48	41	37	33	29	27	20
	0.20	68	47	38	32	28	25	22	20	14
0.85	0.01	86	66	56	50	46	42	39	36	29
	0.02	78	59	50	44	40	37	34	31	25
	0.05	66	49	41	36	32	29	27	25	19
	0.10	57	42	34	30	26	24	22	20	15
	0.20	48	34	27	23	21	18	16	15	11
0.90	0.01	62	48	41	37	34	31	29	27	22
	0.02	56	43	37	33	30	27	25	23	19
	0.05	48	36	30	27	24	22	20	19	15
	0.10	42	30	25	22	20	18	16	15	12
	0.20	35	25	20	18	16	14	13	11	9
0.95	0.01	46	36	31	28	26	24	22	21	17
	0.02	41	32	27	25	23	21	19	18	15
	0.05	35	27	23	20	18	17	16	14	12
	0.10	31	23	19	17	15	14	13	12	9
	0.20	26	19	15	13	12	11	10	9	7

$P_1 = 0.55$

P_2	Уровень значимости	Мощность								
		0.99	0.95	0.90	0.85	0.80	0.75	0.70	0.65	0.50
0.60	0.01	4732	3520	2947	2589	2322	2104	1918	1753	1337
	0.02	4265	3119	2582	2248	2001	1798	1627	1476	1097
	0.05	3626	2578	2093	1793	1573	1395	1245	1114	791
	0.10	3118	2153	1713	1444	1248	1090	959	845	568
	0.20	2581	1712	1323	1089	921	787	677	582	360
0.65	0.01	1168	872	732	644	579	526	480	440	338
	0.02	1053	774	642	561	500	451	409	372	279
	0.05	897	641	522	449	395	352	315	283	204
	0.10	772	537	429	363	316	277	245	217	149
	0.20	640	429	334	276	235	202	176	152	98
0.70	0.01	508	381	321	283	255	232	213	195	151
	0.02	459	338	282	247	221	200	182	166	126
	0.05	391	281	230	199	175	157	141	127	93
	0.10	337	236	190	161	141	124	110	98	69
	0.20	280	189	148	124	106	92	80	70	47
0.75	0.01	278	209	177	156	141	129	118	109	85
	0.02	251	186	156	137	123	111	101	93	71
	0.05	214	155	127	110	98	88	79	72	53
	0.10	185	130	105	90	79	70	63	56	40
	0.20	154	105	83	70	60	52	46	41	28
0.80	0.01	171	129	110	98	88	81	74	69	54
	0.02	154	115	97	85	77	70	64	59	46
	0.05	132	96	79	69	62	56	50	46	35
	0.10	114	81	66	57	50	45	40	36	26
	0.20	95	65	52	44	38	34	30	27	19
0.85	0.01	113	86	73	65	60	55	50	47	37
	0.02	102	77	65	57	52	47	44	40	32
	0.05	87	64	53	47	42	38	34	32	24
	0.10	75	54	44	39	34	31	28	25	19
	0.20	63	44	35	30	26	23	21	19	14
0.90	0.01	78	60	51	46	42	39	36	33	27
	0.02	71	54	46	41	37	34	31	29	23
	0.05	60	45	38	33	30	27	25	23	18
	0.10	52	38	31	27	24	22	20	18	14
	0.20	44	31	25	22	19	17	15	14	10
0.95	0.01	55	43	37	33	31	28	26	25	20
	0.02	50	38	33	29	27	25	23	21	17
	0.05	43	32	27	24	22	20	18	17	14
	0.10	37	27	23	20	18	16	15	14	11
	0.20	31	22	18	16	14	13	12	11	8

Продолжение

$P_1 = 0.60$

P_2	Уровень значимости	Мощность								
		0.99	0.95	0.90	0.85	0.80	0.75	0 70	0.65	0.50
0.65	0.01	4540	3377	2828	2484	2229	2019	1841	1683	1284
	0.02	4092	2993	2478	2157	1920	1726	1562	1417	1054
	0.05	3479	2474	2008	1721	1511	1340	1196	1070	760
	0.10	2992	2067	1644	1386	1198	1047	921	812	547
	0.20	2476	1644	1271	1046	885	756	650	560	347
0.70	0.01	1108	827	695	612	550	499	456	418	322
	0.02	999	734	610	532	475	428	389	354	266
	0.05	851	608	496	427	376	334	300	269	194
	0.10	733	510	408	345	300	264	233	207	142
	0.20	608	407	317	263	224	193	167	145	94
0.75	0.01	476	357	301	266	240	218	200	183	142
	0.02	430	317	265	232	207	188	171	156	118
	0.05	366	264	216	187	165	147	133	120	88
	0.10	316	221	178	152	133	117	104	93	65
	0.20	263	178	140	117	100	87	76	67	44
0.80	0.01	256	193	164	145	131	120	110	101	79
	0.02	232	172	144	127	114	103	94	86	66
	0.05	198	143	118	102	91	82	74	67	50
	0.10	171	121	98	84	74	65	58	52	38
	0.20	142	97	77	65	56	49	43	38	26
0.85	0.01	155	118	100	89	81	74	68	63	50
	0.02	140	105	89	78	70	64	59	54	42
	0.05	120	88	73	63	57	51	46	42	32
	0.10	104	74	60	52	46	41	37	33	25
	0.20	87	60	48	41	35	31	28	25	18
0.90	0.01	101	77	66	59	54	49	46	42	34
	0.02	91	69	58	52	47	43	39	36	29
	0.05	78	58	48	42	38	34	31	29	22
	0.10	68	49	40	35	31	28	25	23	17
	0.20	56	40	32	27	24	21	19	17	13
0.95	0.01	68	53	45	41	37	34	32	30	24
	0.02	62	47	40	36	33	30	28	26	21
	0.05	53	39	33	29	27	24	22	21	16
	0.10	46	34	28	24	22	20	18	17	13
	0.20	38	27	22	19	17	15	14	13	10

Продолжение

$P_1 = 0.65$

P_2	Уровень значимости	Мощность								
		0.99	0.95	0.90	0.85	0.80	0.75	0.70	0.65	0.50
0.70	0.01	4251	3163	2650	2328	2089	1892	1726	1578	1204
	0.02	3833	2804	2322	2022	1800	1618	1464	1329	989
	0.05	3259	2318	1882	1613	1416	1256	1122	1004	714
	0.10	2803	1937	1541	1300	1124	983	865	762	514
	0.20	2320	1541	1192	981	830	710	611	527	327
0.75	0.01	1023	765	643	566	509	462	423	387	298
	0.02	923	679	564	493	440	397	360	328	247
	0.05	786	563	459	395	348	310	278	250	181
	0.10	678	472	378	320	278	245	217	192	133
	0.20	562	377	294	244	208	179	156	135	88
0.80	0.01	433	325	274	242	219	199	183	168	131
	0.02	391	289	242	212	190	172	156	143	109
	0.05	334	240	197	171	151	135	122	110	81
	0.10	288	202	163	139	122	107	96	85	61
	0.20	240	162	128	107	92	80	70	62	41
0.85	0.01	229	173	147	130	118	108	99	91	72
	0.02	207	154	130	114	102	93	85	78	60
	0.05	177	129	106	92	82	74	67	61	45
	0.10	153	109	88	76	67	59	53	48	35
	0.20	128	88	70	59	51	45	39	35	24
0.90	0.01	136	104	88	79	72	66	61	56	45
	0.02	123	93	78	69	62	57	52	48	38
	0.05	105	77	64	56	50	45	41	38	29
	0.10	91	65	54	46	41	37	33	30	22
	0.20	76	53	43	36	32	28	25	22	16
0.95	0.01	86	66	57	51	46	43	40	37	30
	0.02	78	59	50	45	41	37	34	32	25
	0.05	67	50	42	37	33	30	28	25	20
	0.10	58	42	35	30	27	25	22	20	16
	0.20	48	34	28	24	21	19	17	16	12
		$P_1 = 0.70$								
0.75	0.01	3867	2878	2411	2119	1902	1723	1572	1438	1098
	0.02	3486	2552	2114	1841	1639	1474	1334	1211	902
	0.05	2965	2110	1714	1470	1291	1145	1023	916	652
	0.10	2550	1764	1404	1185	1025	897	789	696	471
	0.20	2112	1404	1087	895	758	649	559	482	301
0.80	0.01	915	685	576	507	456	415	379	348	268
	0.02	826	608	506	442	395	356	324	295	222
	0.05	704	504	412	355	313	279	250	225	163
	0.10	607	423	339	288	250	220	195	174	121
	0.20	504	339	265	220	188	162	141	123	80
0.85	0.01	380	286	241	213	193	176	161	148	116
	0.02	343	254	213	187	167	152	138	126	97
	0.05	293	212	174	151	134	120	108	98	72
	0.10	253	178	144	123	108	95	85	76	54
	0.20	211	143	113	95	82	71	63	55	38

Продолжение

P2	Уровень значимости	Мощность								
		0.99	0.95	0.90	0.85	0.80	0.75	0.70	0.65	0.50
0.90	0.01	196	149	126	112	102	93	86	79	63
	0.02	178	133	112	98	89	81	74	68	53
	0.05	152	111	92	80	71	64	58	53	40
	0.10	131	94	76	66	58	52	47	42	31
	0.20	110	76	61	51	45	39	35	31	22
0.95	0.01	113	87	74	66	60	55	51	48	38
	0.02	102	77	66	58	53	48	44	41	33
	0.05	88	65	54	48	43	39	35	32	25
	0.10	76	55	45	39	35	32	29	26	20
	0.20	64	45	36	31	27	24	22	20	14
$P_1 = 0.75$										
0.80	0.01	3386	2522	2114	1858	1668	1512	1379	1262	965
	0.02	3053	2236	1853	1615	1438	1294	1172	1064	794
	0.05	2597	1850	1504	1290	1134	1007	900	806	575
	0.10	2235	1547	1233	1041	901	789	695	614	417
	0.20	1852	1232	955	788	668	572	494	426	268
0.85	0.01	783	586	494	435	392	357	326	300	232
	0.02	707	521	434	380	340	307	279	254	193
	0.05	603	433	354	305	270	241	216	195	142
	0.10	520	363	292	248	216	191	169	151	106
	0.20	432	291	228	190	163	141	123	108	71
0.90	0.01	316	238	202	179	162	147	136	125	98
	0.02	285	212	178	156	140	127	116	107	82
	0.05	244	177	146	127	113	101	91	83	62
	0.10	211	149	121	104	91	81	73	65	47
	0.20	176	120	96	81	70	61	54	48	33
0.95	0.01	157	120	102	91	83	76	70	65	52
	0.02	142	107	90	80	72	66	61	56	44
	0.05	122	90	75	65	58	53	48	44	34
	0.10	106	76	62	54	48	43	39	35	26
	0.20	88	62	50	42	37	33	29	26	19
$P_1 = 0.80$										
0.85	0.01	2810	2094	1756	1545	1388	1259	1149	1052	806
	0.02	2534	1858	1541	1343	1198	1078	977	888	664
	0.05	2157	1538	1252	1075	945	840	751	674	483
	0.10	1856	1287	1027	868	753	660	582	515	351
	0.20	1539	1027	797	659	559	480	415	360	228
0.90	0.01	627	471	397	351	316	288	264	243	189
	0.02	566	419	349	306	274	248	226	206	157
	0.05	483	348	286	247	219	196	176	159	117
	0.10	417	293	236	201	176	156	139	124	88
	0.20	347	236	185	155	133	116	102	89	60
0.95	0.01	241	183	155	138	125	115	106	98	77
	0.02	218	163	137	121	109	99	91	84	65
	0.05	187	136	113	99	88	79	72	66	50
	0.10	162	115	94	81	72	64	58	52	38
	0.20	135	94	75	64	55	49	44	39	28

Продолжение

$P_1 = 0.85$

P_2	Уровень значимости	Мощность								
		0.99	0.95	0.90	0.85	0.80	0.75	0.70	0.65	0.50
0.90	0.01	2137	1595	1340	1179	1060	963	880	806	620
	0.02	1928	1416	1176	1027	916	826	749	682	513
	0.05	1642	1174	957	823	725	646	579	520	375
	0.10	1415	984	787	667	579	509	450	399	275
	0.20	1175	787	613	508	433	373	324	281	182
0.95	0.01	447	337	285	253	228	209	192	177	139
	0.02	404	300	252	221	199	180	165	151	117
	0.05	345	251	207	179	160	143	130	118	88
	0.10	299	212	172	147	130	115	103	93	67
	0.20	250	171	136	115	99	87	77	68	47
		$P_1 = 0.90$								
0.95	0.01	1368	1025	863	762	686	624	572	525	407
	0.02	1235	911	760	665	595	538	489	446	339
	0.05	1054	758	621	536	474	423	381	344	252
	0.10	910	637	513	437	381	336	299	267	188
	0.20	758	512	402	336	288	250	219	192	128

Таблица A.5

Процентные точки критерия упорядоченности пропорций Бартоломью при сравнении $m = 3$ пропорций

c	Уровень значимости				
	.10	.05	.025	.01	.005
0.0	2.952	4.231	5.537	7.289	8.628
0.1	2.885	4.158	5.459	7.208	8.543
0.2	2.816	4.081	5.378	7.122	8.455
0.3	2.742	4.001	5.292	7.030	8.360
0.4	2.664	3.914	5.200	6.932	8.258
0.5	2.580	3.820	5.098	6.822	8.146
0.6	2.486	3.715	4.985	6.700	8.016
0.7	2.379	3.593	4.852	6.556	7.865
0.8	2.251	3.446	4.689	6.377	7.677
0.9	2.080	3.245	4.465	6.130	7.413
1.0	1.642	2.706	3.841	5.413	6.635

Приводится с сокращениями по табл. A.1 Barlow R. E., Bartholomew D. J., Bremner J. M. and Brunk H. D. (1972). "Statistical inference under order restrictions." John Wiley and Sons, New York.

Таблица A.6

Процентные точки критерия упорядоченности пропорций Бартоломью при сравнении $m = 4$ пропорций. Таблица симметрична относительно c_1 и c_2

c_2	Уровень значимости	c_1							
		0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7
0.0	.10	4.010							
	.05	5.435							
	.025	6.861							
	.01	8.746							
	.005	10.171							
0.1	.10	3.952	3.891						
	.05	5.372	5.305						
	.025	6.794	6.724						
	.01	8.676	8.601						
	.005	10.098	10.020						
0.2	.10	3.893	3.827	3.758					
	.05	5.307	5.235	5.160					
	.025	6.725	6.649	6.570					
	.01	8.602	8.522	8.437					
	.005	10.022	9.939	9.851					
0.3	.10	3.831	3.760	3.685	3.606				
	.05	5.239	5.162	5.080	4.993				
	.025	6.653	6.571	6.484	6.391				
	.01	8.525	8.438	8.346	8.246				
	.005	9.942	9.852	9.756	9.653				

Продолжение

c_2	Уровень значимости	c_1							
		0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7
0.4	.10	3.765	3.688	3.607	3.519	3.423			
	.05	5.166	5.083	4.994	4.898	4.791			
	.025	6.575	6.486	6.392	6.289	6.174			
	.01	8.442	8.348	8.247	8.137	8.014			
	.005	9.855	9.758	9.653	9.539	9.411			
0.5	.10	3.695	3.610	3.521	3.423	3.313	3.187		
	.05	5.088	4.997	4.898	4.791	4.670	4.528		
	.025	6.491	6.394	6.289	6.173	6.043	5.891		
	.01	8.352	8.246	8.136	8.013	7.873	7.709		
	.005	9.761	9.654	9.537	9.409	9.264	9.092		
0.6	.10	3.617	3.523	3.422	3.310	3.183	3.031	2.837	
	.05	5.002	4.900	4.789	4.665	4.524	4.354	4.135	
	.025	6.398	6.289	6.170	6.038	5.886	5.702	5.462	
	.01	8.251	8.135	8.008	7.867	7.703	7.504	7.244	
	.005	9.656	9.535	9.404	9.256	9.085	8.877	8.604	
0.7	.10	3.530	3.422	3.305	3.172	3.017	2.822	2.550	1.987
	.05	4.904	4.787	4.657	4.510	4.337	4.118	3.805	3.137
	.025	6.291	6.166	6.027	5.870	5.682	5.443	5.100	4.346
	.01	8.135	8.002	7.854	7.684	7.482	7.223	6.846	6.000
	.005	9.534	9.395	9.242	9.065	8.853	8.581	8.183	7.279
0.8	.10	3.427	3.296	3.151	2.981	2.770	2.473	1.642	
	.05	4.787	4.644	4.483	4.294	4.056	3.715	2.706	
	.025	6.163	6.011	5.838	5.634	5.375	4.999	3.841	
	.01	7.994	7.832	7.647	7.427	7.146	6.734	5.412	
	.005	9.385	9.217	9.025	8.795	8.500	8.064	6.635	
0.9	.10	3.291	3.110	2.897	2.621	2.166			
	.05	4.631	4.432	4.195	3.883	3.353			
	.025	5.990	5.778	5.523	5.182	4.591			
	.01	7.804	7.577	7.303	6.933	6.277			
	.005	9.183	8.948	8.661	8.273	7.576			
1.0	.10	2.952							
	.05	4.231							
	.025	5.537							
	.01	7.289							
	.005	8.628							

Приводится с сокращениями по табл. A.2 Barlow R. E., Bartholomew D. J., Bremner J. M. and Brunk H. D. (1972). "Statistical inference under order restrictions." John Wiley and Sons, New York.

Таблица А.

Процентные точки критерия упорядоченности пропорций Бартоломью при сравнении пропорций при равных объемах выборок ($m = 3, 4, \ldots, 12$

m	Уровень значимости				
	.10	.05	.025	.01	.005
3	2.580	3.820	5.098	6.822	8.146
4	3.187	4.528	5.891	7.709	9.092
5	3.636	5.049	6.471	8.356	9.784
6	3.994	5.460	6.928	8.865	10.327
7	4.289	5.800	7.304	9.284	10.774
8	4.542	6.088	7.624	9.639	11.153
9	4.761	6.339	7.901	9.946	11.480
10	4.956	6.560	8.145	10.216	11.767
11	5.130	6.758	8.363	10.458	12.025
12	5.288	6.937	8.561	10.676	12.257

Приводится с сокращениями по табл. A.3 Barlow R. E., Bartholomew D. J., Bremner J. M., and Brunk H. D. (1972), "Statistical inference under order restrictions." John Wiley and Sons, New York.

Ответы к задачам

1.2. Нам дано значение $P(B) = 0{,}001$.

а) $P(A|B) = 0{,}99$ и $P(A|\bar{B}) = 0{,}01$. Согласно (1.13) $P_{F+} = 0{,}01 \cdot (1 - 0{,}001)$ $[0{,}01 + 0{,}001\,(0{,}99 - 0{,}01)] = 0{,}9098$. Согласно (1.14) $P_{F-} = (1 - 0{,}99) \cdot 0{,}001$ $[1 - 0{,}01 - 0{,}001 \cdot (0{,}99 - 0{,}01)] = 0{,}00001$ (т. е. 1 на 100000). Ложная положительная доля слишком велика для большинства случаев.

б) Теперь $P(A|B) = 0{,}98$ и $P(A|\bar{B}) = 0{,}0001$. Для нового определения положительного результата $P_{F+} = 0{,}0001$ $(1 - 0{,}001)/[0{,}0001 + 0{,}001\,(0{,}98 - 0{,}0001)] = 0{,}0925$ и $P_{F-} = (1 - 0{,}98) \cdot 0{,}001/[1 - 0{,}0001 - 0{,}001\,(0{,}98 - 0{,}0001)] = 0{,}00002$ (т. е. 2 на 100000). Ложная положительная доля примерно в десять раз меньше, чем в а), а ложная отрицательная доля по прежнему очень мала.

в) Пропорция людей, реакция которых на первый тест положительна, равна согласно (1.11) $P(A) = 0{,}99 \cdot 0{,}001 + 0{,}01 \cdot 0{,}999 = 0{,}01098$, или 1098 индивидуумов на 100 000 испытуемых. Следовательно, 98 902 из 100 000 людей отрицательно отреагируют на первый тест и им не придется повторно проходить тестирование.

1.3.

а) (1) 40 000 больных неврозом живут одиноко. (2) 200 из них будут госпитализированы. (3) 60 000 больных неврозом живут со своими семьями. (4) 360 из них будут госпитализированы. (5) $p_1 = 200/(200 + 360) = 0{,}357$. (6) $p_1 < P(A|B)$.

б) (1) 200 600 людей, не страдающих неврозом, живут одиноко. (2) 1000 из них будут госпитализированы. (3) 800 000 людей, не страдающих

процом, живут вместе со своими семьями (4) 1800 из них будут госпи[illegible]лизированы. (5) $p_2 = 1000/(1000+1800) = 0{,}357$. (6) $p_2 > P(A|\bar{B})$.

г) p_1 равно p_2, хотя $P(A|B)$ много больше $P(A|\bar{B})$.

1.4. Нам даны значения $n = 100$ и $p = 0{,}05$.

а) Согласно (1.26) $P_L = 0{,}01$. Согласно (1.27) $P_U = 0{,}15$.

б) Значение $c_{\alpha/2}\sqrt{pq/n + 1/(2n)}$ равно 0,06, так что нижняя и верхняя доверительные границы для P, основанные на (1.29), равны соответственно 0,01 и 0,11.

в) Интервал в б) уже интервала в а), но тот факт, что нижняя граница в б) отрицательна, вызывает сомнения относительно его пригодности. Если исправить в (1.29) поправки на непрерывность, то границы интервала будут совпадать с границами, найденными в б), с точностью до двух знаков. Неверный вывод не обусловлен поправками на непрерывность.

3.4. Мы хотим сделать выбор между $P_1 = 0{,}45$ для нейтрального препарата («пустышки») и $P_2 = 0{,}65$ для активного препарата с помощью одностороннего критерия.

а) Для одностороннего критерия при уровне значимости 0,01 мы возьмем $c_{0.99} = 2{,}326$ в (3.14), для мощности 0,95 возьмем $c_{0.95} = -1{,}645$. Значение 0,55. Поэтому $n' = [2{,}326\sqrt{2 \cdot 0{,}55 \cdot 0{,}45} - (-1{,}645)\sqrt{0{,}45 \cdot 0{,}55 + 0{,}65 \cdot 0{,}35}]^2/(0{,}65 - 0{,}45)^2 = 191{,}85$. Из (3.15) $n = 201{,}73$, т. е. на каждый вид лечения приходится 202 пациента.

Можно действовать иначе: найдите в табл. А.3 значение для $P_1 = 0{,}45$, $P_2 = 0{,}65$, $\alpha = 0{,}02$ (учитывая, что мы рассматриваем односторонний критерий) и мощности 0,95.

б) Если $\alpha = 0{,}05$ (по-прежнему используется односторонний критерий) и $\beta = 0{,}80$, то для каждой группы потребуется 85 пациентов.

в) Если $n = 52$, то значение $c_{1-\beta}$ в (3.17) равно $-0{,}20$. Соответствующая мощность равна 0,58. (Табл. А.2 дает значение $P = 0{,}8415$ как соответствующее $z = 0{,}2$. Площадь под кривой нормального распределения ниже 0,20 равна $P/2 = 0{,}42$, а мощность равна $1 - P/2 = 0{,}58$.)

3.5. Мы хотим сделать выбор между $P_1 = 0{,}25$ и $P_2 = 0{,}40$.

а), б) Чтобы найти значение $n_1 = m$, воспользуемся выражениями (3.19) и (3.18) для каждого значения r, взяв $c_\alpha = 2{,}576$ и $c_{1-\beta} = -1{,}645$. Тогда значение n_2 равно rm, так что суммарный объем выборки равен: $n_1 + n_2 = m(r+1)$. Суммарная стоимость $= 10n_1 + 12n_2 = 10m + 12rm = m(10 + 12r)$. Полная таблица выглядит следующим образом:

Отношение объемов выборок (r)	n_1	n_2	Объем суммарной выборки	Полная стоимость (доллары)
0,5	530	265	795	8480
0,6	473	284	757	8138
0,7	432	302	734	7944
0,8	401	321	722	7862
0,9	377	339	716	7838
1	357	357	714	7854

Полная стоимость минимизируется при $r = 0{,}9$ (т. е. при $n_1 = 377$ и $n_2 = 0{,}9 \cdot 377 = 339$).

в) Если полная стоимость должна составить 6240 долларов и если исследователь решит взять значение $r=0{,}9$, то m будет равно: $m=6240/(10+12\cdot 0{,}9)=300$. Значение m' согласно (3.20) есть $m'=300-(0{,}9+1)/(0{,}9\cdot 0{,}15)=285{,}93$. Чтобы найти $c_{1-\beta}$ из (3.19), вычислим при $r=0{,}9$ величины $\bar{P}=0{,}32$ и $\bar{P}\bar{Q}=0{,}2176$. Следовательно,

$$c_{1-\beta}=\frac{2{,}576\sqrt{(0{,}9+1)\cdot 0{,}2176}-0{,}15\sqrt{0{,}9\cdot 285{,}93}}{\sqrt{0{,}9\cdot 0{,}25\cdot 0{,}75+0{,}40\cdot 0{,}60}}=-1{,}17.$$

Проводя интерполяцию по табл. А.2, находим, что мощность критерия приближенно равна 0,88.

5.4.

а) Диагнозы, поставленные в клиниках

	Шизофрения	Эмоциональные расстройства	Сумма
Нью-Йорк	82	24	106
Лондон	51	67	118
Сумма	133	91	224

	Диагнозы, поставленные психиатрами-исследователями		Сумма
	шизофрения	эмоциональные расстройства	
Нью-Йорк	43	53	96
Лондон	33	85	118
Сумма	76	138	214

	Диагнозы, поставленные компьютером		Сумма
	шизофрения	эмоциональные расстройства	
Нью-Йорк	67	27	94
Лондон	56	37	93
Сумма	123	64	187

г)

Кто ставил диагноз	Отношение шансов по (5.16)
Клиницисты	4,49
Психиатры-исследователи	2,09
Компьютер	1,64

отношения шансов для психиатров-исследователей и для компьютера близки, и оба существенно отличаются от отношения шансов для клиницистов.

в)

Кто ставил диагноз	o	o'
Клиницисты	4,49	4,41
Психиатры-исследователи	2,09	2,08
Компьютер	1,64	1,63

5.5. Для значений n_{ij} табл. 5.1 и значений N_{ij} из табл. 5.3 значение $\chi^2=\Sigma\Sigma\,(n_{ij}-N_{ij})^2/N_{ij}=3{,}32$ близко к 3,25, найденному по (5.51).

5.6. Начальному приближению $\omega_U^{(1)}=5{,}37$ соответствуют значения $X=527{,}75$ и $Y=401{,}48$, следовательно, $N_{11}=14{,}45$ и $W=0{,}1993$. Значение F с поправкой $+1/2$ на непрерывность равно $-0{,}73$, поэтому надо применить итеративную процедуру.

Вычисляем значения $T=0{,}9345$, $U=0{,}0049$ и (заменяя $-1/2$ на $+1/2$) $V=1{,}5438$. Значит, второе приближение к верхней границе отношения шансов равно по (5.58):

$$\omega_U^{2}=5{,}37-\frac{-0{,}73}{1{,}5428}=5{,}84.$$

Для этого значения отношения шансов $N_{11}=14{,}87$ и $W=0{,}2016$. При этом $F=0{,}01$ достаточно близко к нулю, чтобы закончить итерации. Верхняя 95%-ная доверительная граница отношения шансов для истинного отношения шансов для данных табл. 5.1 есть $\omega_U=5{,}84$.

5.7. Ожидаемые частоты равны:

	B	$\bar{B}$
A	14,87	35,13
$\bar{A}$	10,13	139,87

Для этих частот соответствующее значение фи-коэффициента (и, следовательно, верхняя 95%-ная доверительная граница этого параметра) равно 0,30. Верхняя 95%-ная доверительная граница для относительного риска есть (14,87/50)/(10,13/150) = 4,40.

5.9.

а) Для небелых испытуемых оценка риска смерти новорожденного, приносимого низким весом, равна:

$$r_A = \frac{0{,}140 \cdot 0{,}8625 - 0{,}1147 \cdot 0{,}0088}{0{,}0228 \cdot 0{,}8713} = 0{,}557,$$

что совпадает с точностью до двух знаков с оценкой приносимого риска для белых.

б) Стандартная ошибка $\ln(1-r_A)$ по (5.79) равна:

$$\text{s.e.}\,(\ln(1-r_A)) = \sqrt{\frac{0{,}1147 + 0{,}557\,(0{,}0140 + 0{,}8625)}{37840 \cdot 0{,}0088}} = 0{,}013.$$

95%-ный доверительный интервал для параметра задается значениями 0,518 и 0,593. Он почти полностью перекрывается интервалом, определенным по (5.85), но немного шире его. Следовало ожидать такого различия длин интервалов, поскольку число рождений среди белых больше, чем среди небелых.

6.3.

а)

	Курящие	Некурящие	Сумма
Опыт	26	8	34
Контроль	73	141	214
Сумма	99	149	248

χ^2 (без поправки) = 21,95, $\varphi = 0{,}30$.

б)

	Курящие	Некурящие	Сумма
Опыт	163	51	214
Контроль	12	22	34
Сумма	175	73	248

χ^2 (без поправки) = 23,60, $\varphi = 0{,}31$.

Фи-коэффициенты в а) и б) близки.

)

	Курящие	Некурящие	Сумма
Опыт	94	30	124
Контроль	42	82	124
Сумма	136	112	248

χ^2 (без поправки) $=44{,}03$, $\varphi=0{,}42$.

ри-коэффициент в в) заметно отличается от значений в а) и б). Разница ежду коэффициентами в а) и в), выраженная в процентах, составляет 0,42—0,30)/0,30·100%=40%.

6.4.

а) Искомое значение $n_{.}=807$.

б) Искомое значение $N_P=702$. В процентах N_P меньше $n_{.}$ на (807—−702)/807·100%=13%.

в) Искомое значение $N_R=398$. В процентах N_R меньше $n_{.}$ на 100%·(807 — 398)/807 = 51%, а N_R меньше N_P на 100%·(702—398)/702=43%.

7.1.

а) Значение статистики критерия (с поправкой на непрерывность) равно 3.54. Разница между долями улучшения значима во второй клинике.

б) Разность между долями улучшения равна: $d_2=0{,}75—0{,}35=0{,}40$. Ее тандартная ошибка равна: s.e. $(d_2)=0{,}06$. Статистика для проверки значиости различия $d_1=0{,}20$ и $d_2=0{,}40$ равна: $z=2{,}17$. Эти две разности различаются на уровне значимости 0,05.

в) Относительный прирост доли улучшения во второй клинике равен: $_{e(2)}=(0{,}75-{-}0{,}35)/(1—0{,}35)=0{,}62$. Значение $L_2=—0{,}97$, оценка стандартной шибки L_2 равна 0,19. Статистика для сравнения L_1 и L_2 равна:

$$z=\frac{|-0.69-(-0{,}97)|}{\sqrt{0{,}28^2+0{,}19^2}}=0{,}83,$$

ак что два относительных прироста значимо не отличаются.

7.2.

а) Для выборки пациентов, проходивших лечение в порядке АБ, $n=20$ $p_1=15/20=0{,}75$.

б) Для выборки пациентов, проходивших лечение в порядке БА, $m=15$ $p_2=5/15=0{,}33$. (Вспомните, что p_2 — пропорция тех пациентов, кто по ложительно отреагировал на лечение, проведенное первым, т. е. на Б.)

в) Значение статистики для сравнения p_1 и p_2 с поправкой на непрерывность равно 2,14. Действие лечений А и Б значимо отличается.

8.1. Значение статистики Мак-Немара для сравнения пропорций пациентов с диагнозом эмоциональное расстройство равно $(|20-{-}10|—1)/(20+10)=-2{,}70$. Различие статистически незначимо.

Значение статистики Мак-Немара для сравнения пропорций пациентов диагнозом, отличным от шизофрении и эмоционального расстройства, равно: $(|15—5|—1)/(15+5)=4{,}05$. Поскольку эта величина не достигает 5,99, азличие статистически незначимо.

8.2.

а) Статистика хи-квадрат Стюарта — Максвелла принимает значение 10,43 при двух степенях свободы. Два распределения результирующего фактора различаются значимо.

б) Значение d_1 равно: 70—50 = +20, d_3 = 10—20 = —10, $d_1 - d_3 = +30$. Новое лечение выглядит лучше традиционного в том смысле, что связано с большим числом улучшений в итоге. Статистика критерия (8.20) равна 9,57, так что новое лечение значимо превосходит традиционное.

9.3.

а)

Выборка	n	Пропорция пациентов с диагнозом эмоциональное расстройство
1	105	0,019
2	192	0,068
3	145	0,166
Всего	442	0,088

Статистика хи-квадрат принимает значение 18,18 при двух степенях свободы. Пропорции пациентов с диагнозом эмоциональное расстройство различаются с уровнем значимости выше 0,01.

б) $p_{1,2} = 0{,}051$ и $n_{1,2} = 297$. Значение хи-квадрат для проверки значимости различия между $p_{1,2}$ и p_3 равно 16,06, различие значимо с уровнем значимости выше 0,01. Значение хи-квадрат для проверки различия между p_1 и p_2 равно 2,03, поэтому различие незначимо.

в) $\bar{\chi}^2 = \chi^2 = 18{,}18$ и

$$c = \sqrt{\frac{105 \cdot 145}{297 \cdot 337}} = 0{,}39.$$

Предложенное упорядочение значимо на уровне выше 0,01.

9.5.

а) Ридит-среднее для группы Б (А — контрольная) равно 0,963. Вероятность, что случайно извлеченный из группы Б объект получит травму не менее тяжелую, чем случайно извлеченный объект из группы Б, равна 0,963.

б) Ридит-среднее для группы А (Б — контрольная) равно 0,037, что точно совпадает с дополнением вероятности в а) до 1.

в) Стандартная ошибка ридит-среднего равна по (9.40) 0,040.

г) Стандартная ошибка ридит-среднего по (9.45) равна 0,041, что лишь немного больше значения, вычисленного в а).

10.1.

а) Значение $\chi^2_{2vs3} = 0{,}02$, что подтверждает возможное равенство отношений шансов во 2-м и 3-м исследованиях.

б) Взвешенное среднее L_2' и L_3' равно: $\bar{L}_{2,3} = 0{,}862$. $\chi^2_{1vs(2.3)}$ принимает значение 9,40, указывая, что L_1' значимо отличается от $\bar{L}_{2,3}$ на уровне значимости 0,01 (соответствующее критическое значение распределения хи-квадрат равно 9,21).

в) Сумма статистик хи-квадрат в а) и б) равна 9,42 и совпадает со значением χ^2_{homog}, найденным по (10.18), с точностью до ошибок округления.

10.2.

а) Среднее значение логарифма отношения шансов для групп 2 и 3 есть $\bar{L}_{2,3}=0{,}862$ (см. задачу 10.1б), а оценка его стандартной ошибки равна:

$$\text{s.e.}=(\bar{L}_{2,3})=\frac{1}{\sqrt{w'_2+w'_3}}=0{,}160.$$

Значение χ^2_{assoc} для этих двух групп равно:

$$\chi^2_{assoc}=\left[\frac{\bar{L}_{2,3}}{\text{s.e.}(\bar{L}_{2,3})}\right]^2=29{,}03,$$

поэтому средний логарифм отношения шансов значимо отличается от нуля.

б) Приближенный 95%-ный доверительный интервал для истинного логарифма отношения шансов есть

$$\bar{L}_{2,3}\pm 1{,}96\ \text{s.e.}\ (\bar{L}_{2,3}),$$

интервал (0,548, 1,176).

Средний логарифм отношения шансов равен: exp (0,862) = 2,37. Приведенный доверительный интервал с уровнем доверия 95% есть интервал, ограниченный значениями exp (0,548) и exp (1,176), т. е. интервал (1,73,

	n	Пропорция кататоников
	112	0,286
	154	0,506
	31	0,419
	151	0,364
	124	0,298
	572	0,376

тика хи-квадрат для сравнения этих пяти пропорций равна

ическое значение распределения хи-квадрат с 4 степенями свободы уровне значимости 0,001 равно 18,47. Подходы к диагностике кататонической и параноидальной шизофрении в пяти клиниках значимо различаются ($p<0{,}001$).

11.2.

а) P_L равно 0,50, а $p_L=0{,}75\cdot 0{,}50+0{,}05\cdot 0{,}50=0{,}40$.

б) P_B равно 0,40, а $p_B = 0{,}9 \cdot 0{,}40 + 0{,}1 \cdot 0{,}60 = 0{,}42$.

в) $D = P_L - P_B$ равно 0,10, а $d = p_L - p_B = -0{,}02$.

Разности имеют противоположные знаки.

г) Значение отношения шансов, вычисленное по P_L и P_B, равно: $0{,}50 \cdot 0{,}60/(0{,}40 \cdot 0{,}50) = 1{,}50$, а отношение шансов, вычисленное по p_L и p_B, равно: $0{,}40 \cdot 0{,}58/(0{,}42 \cdot 0{,}60) = 0{,}92$. Эти два значения отношения шансов находятся по разные стороны от единицы.

12.1.

а) $p_B = 60/200 = 0{,}30$, а отношение шансов принимает значение $0{,}44 \cdot 0{,}70/(0{,}30 \cdot 0{,}56) = 1{,}83$.

б) Значение $n_{00}/n_0 = 18/18 = 1$, $n_{10}/n_1 = 2/32 = 0{,}06$. Получаем значение $P_B = 1 \cdot 0{,}30 + 0{,}06 \cdot 0{,}70 = 0{,}34$. Значение отношения шансов по долям курящих с поправками есть $0{,}51 \cdot 0{,}66/(0{,}34 \cdot 0{,}49) = 2{,}02$. Связь в б) выглядит сильнее, чем в а).

в) Значение $n_{\infty}/n_0 = 16/18 = 0{,}89$, $n_{10}/n_1 = 7/32 = 0{,}22$. Получаем значение $P_B = 0{,}89 \cdot 0{,}30 + 0{,}22 \cdot 0{,}70 = 0{,}42$, а для отношения шансов $0{,}51 \cdot 0{,}58/(0{,}42 \cdot 0{,}49) = 1{,}44$. Связь в в) выглядит слабее, чем в а).

13.3.

а)

Исследование	n	p_e	$\widehat{\varkappa}$	A	B	C	$s.e.(\widehat{\varkappa})$
1	20	0,59	0,39	0,0742	0,0784	0,0009	0,21
2	20	0,71	0,48	0,0820	0,0393	0,0123	0,25
3	30	0,54	0,35	0,0714	0,1196	0,0000	0,17

Значение числителя в (13.21) есть $0{,}39/0{,}21^2 + 0{,}48/0{,}25^2 + 0{,}35/0{,}17^2 = 28{,}63$, значение числителя есть $1/0{,}21^2 + 1/0{,}25^2 + 1/0{,}17^2 = 73{,}28$. Общее значение каппа составляет: $28{,}63/73{,}28 = 0{,}39$.

б) Значение статистики хи-квадрат в (13.22) равно: $(0{,}39 - 0{,}39)^2/0{,}21^2 + (0{,}48 - 0{,}39)^2/0{,}25^2 + (0{,}35 - 0{,}39)^2/0{,}17^2 = 0{,}18$ при 2 степенях свободы. Различие трех оценок каппа незначимо.

в) Приближенный 95%-ный доверительный интервал для общего значения каппа есть $0{,}39 \pm 1{,}96 \cdot \sqrt{1/73{,}28}$, т. е. определяется величинами 0,16 и 0,62. Общее значение каппа значимо отличается от нуля (доверительный интервал не содержит нуль), но величина каппа соответствует согласованности, лишь немного лучшей случайной согласованности (даже верхняя 95%-ная граница, 0,62, означает невысокую согласованность).

14.1.

а) Доля нарушений функции легких у занятых в сфере обслуживания больше чем у рабочих промышленных предприятий в возрасте до 50 лет (иногда эти доли совпадают), но меньше чем в возрастных группах 50 лет и старше.

б) Единственным полезным следствием стандартизации будет простота сравнения, которое надо будет проводить лишь по двум стандартизованным долям. Главные потери в анализе — невозможность описать явление перекрытия и сильная зависимость направления различия между двумя стандартизованными долями от распределения по возрасту в стандартной популяции.

в)

Стандартизованные доли

Стандарт	Промышленность	Сервис	Различие
1	3,98%	4,40%	Сервис > Промышленность
2	8,37%	6,92%	Промышленность > Сервис
3	3,87%	3,84%	Приближенное равенство

)

Стандартизованные доли

Возрастной интервал	Промышленность	Сервис	Различие
20—49 лет	2,58%	3,37%	Сервис > Промышленность
50 лет и больше	8,23%	7,14%	Промышленность > Сервис

Предметный указатель

Оглавление

Дж. Флейс

) году в серии "Библиотечка иностранных книг для экономистов
ков" издательство предполагает выпустить:

л П. Теория индексов и практика экономического анализа: Пер.
М.: Финансы и статистика, 1990 (II кв.). — 24 л.: ил. — Пер. изд.:
. — ISBN 5–279–00442–1.
ографии систематически рассмотрены различные индексные методы,
ие малоизвестные советскому читателю. В центре внимания проблема
сти той или иной формы индекса экономическим реалиям. Большая
риала оригинальна и не отражена в литературе на русском языке.
цем издании книга дополнена главой по индексным методам для
одных сопоставлений и примерами программных реализаций для ПЭВМ
ирокого круга статистиков-практиков, преподавателей и студентов

9 году в серии "Библиотечка иностранных книг для экономистов
ков" издательством выпущены:

ш П. Факторный анализ с обобщениями: Пер. с чеш. — М.: Финансы
ка, 1989 (I кв.). — 14 л.: ил. — Пер. изд.: ЧССР, 1985. — 1 р. 80 к.
ге представлены математико-статистический аппарат и приложения
го анализа, в том числе нелинейные методы и модели. Автором
ата попытка методологически обобщить ряд новейших моделей,
к факторному анализу с точки зрения возможных приложений
ике, социологии, психологии (модели Макдональда и Сваминатана,
модели целевого и конфирматорного факторного анализа).
аучных сотрудников, разрабатывающих методы прикладной
и, широкого круга пользователей многомерных статистических
преподавателей и студентов вузов.

а В. Многомерный сравнительный анализ в эконометрическом
ании: Пер. с польск. — М.: Финансы и статистика, 1989 (I кв.). —
— Пер. изд.: ПНР, 1986. — 1 р. 50 к.
та посвящена построению эконометрических моделей в условиях
дности данных. Рассмотрены методы разделения данных на
ые подмножества и обеспечения адекватности моделей. Книга
дополняет уже известную советским специалистам работу
"Сравнительный многомерный анализ в экономических
ениях (методы таксономии и факторного анализа)"
истика, 1980).
специалистов, использующих количественные методы анализа
ике, преподавателей и студентов вузов.

ы И СТАТИСТИКА ISBN 5–279–00249–6