С.Гланц

МЕДИКО-БИОЛОГИЧЕСКАЯ СТАТИСТИКА

McGraw-Hill, 1994; М.: Практика, 1998. — 459 с.

В книге описаны все основные методы, которыми пользуется современная статистика, как параметрические, так и непараметрические: анализ различий, связей, планирование исследования, анализ выживаемости. Просто и наглядно — при этом вполне строго — автор описывает принцип каждого метода, дает четкую схему применения, обязательно указывает на ограничения и возможные ошибки. Изящные иллюстрации и остроумный разбор примеров, взятых из медицинских публикаций, делают чтение легким и увлекательным. Врачам-практикам книга поможет грамотно, критически читать медицинскую литературу. Для врачей-исследователей книга станет руководством по планированию, проведению и обработке результатов исследований.

Оглавление

Предметный указатель

Посвящается Марше Гланц

ТАБЛИЦЫ КРИТИЧЕСКИХ ЗНАЧЕНИЙ

УСЛОВНЫЕ ОБОЗНАЧЕНИЯ

α уровень значимости (вероятность ошибки I рода); коэффициент сдвига в уравнении регрессии
α' уровень значимости при множественном сравнении
a выборочная оценка коэффициента сдвига
β вероятность ошибки II рода; коэффициент наклона в уравнении регрессии
b выборочная оценка коэффициента наклона
δ величина эффекта (изменение количественного признака)
d выборочная оценка величины эффекта
φ параметр нецентральности
F критерий F
H критерий Крускала—Уоллиса
k число сравнений
l интервал сравнения
m число групп
μ среднее по совокупности
N число членов совокупности
n объем выборки (численность группы)

P	вероятность справедливости нулевой гипотезы
p	доля
$\hat{p}$	выборочная оценка доли
Q	критерий Данна
q'	критерий Даннета
q	критерий Ньюмена—Кейлса; критерий Тьюки
r	коэффициент корреляции Пирсона
r_s	коэффициент ранговой корреляции Спирмена
Σ	суммирование
σ	стандартное отклонение
σ^2	дисперсия
S	вариация (сумма квадратов отклонений)
$S(t)$	выживаемость
s	выборочная оценка стандартного отклонения
s^2	выборочная оценка дисперсии
s_a	стандартная ошибка коэффициента сдвига
s_b	стандартная ошибка коэффициента наклона
$s_{\hat{p}}$	стандартная ошибка доли
$s_{y\|x}$	остаточное стандартное отклонение
$s_{\bar{X}}$	стандартная ошибка среднего
T	критерий Манна—Уитни
t	критерий Стьюдента
t_α	критическое значение t при уровне значимости α
ν	число степеней свободы
$\nu_{\text{вну}}$	внутригрупповое число степеней свободы (знаменателя)
$\nu_{\text{меж}}$	межгрупповое число степеней свободы (числителя)
W	критерий Уилкоксона
χ^2	критерий χ^2
χ_r^2	критерий Фридмана
$\hat{y}$	значение уравнения регрессии
X	значение количественного признака
$\bar{X}$	выборочное среднее
z	критерий z (величина со стандартным нормальным распределением)

Предисловие

После окончания докторантуры мне часто случалось помогать друзьям и коллегам разобраться с тем или иным статистическим вопросом. Постепенно потребность в кратких, интуитивно понятных и в то же время достаточно строгих объяснениях привела к появлению двухчасовой лекции, включавшей даже демонстрацию слайдов. Эта лекция охватывала использование статистических методов в медицине, ошибки в их применении и способы избежать этих ошибок. Лекции оказались настолько успешными, что теперь уже мне пришлось выслушать многочисленные предложения написать вводный курс по статистике.

Так возникла эта книга. Адресована она студентам-медикам, научным работникам, преподавателям и врачам-практикам. Ее с равным успехом можно использовать и для самостоятельного изучения, и в качестве учебного пособия. Например, она послужила основой курса медицинской статистики в Калифорнийском университете в Сан-Франциско. Курс объемом 81 лекционный час включал первые восемь глав книги. Кроме того, еженедельно проводился семинар. Книга также использовалась при чтении краткого курса статистики для студентов стоматологи-

ческого факультета. Этот курс охватывал материал первых трех глав. Кроме того, книга пригодилась мне при чтении интенсивного курса, который занимал полсеместра и был рассчитан на основательное усвоение всего материала. Среди многочисленных слушателей были студенты старших курсов, аспиранты и научные сотрудники.

Эта книга имеет несколько отличий от других вводных курсов статистики — именно эти отличия, похоже, и обусловили ее популярность.

Во-первых, в книге отчетливо проведена мысль, что результаты многих биологических и медицинских работ основаны на неправильном использовании статистических методов и способны только ввести в заблуждение. Большинство ошибок связано с неправомерным использованием критерия Стьюдента. Причина такой концентрации, вероятно, кроется в том, что в пору учебы будущие исследователи не успели узнать о существовании других статистических методов (в учебниках, по которым они учились, первая глава обычно посвящена критерию Стьюдента). Напротив, дисперсионный анализ если и излагается, то, как правило, в последней главе, до которой редко кто добирается. Между тем медицинские данные чаще требуют именно дисперсионного анализа, и именно он служит основой для всех параметрических критериев — поэтому свою книгу я начинаю изложением дисперсионного анализа и лишь затем, как частный случай, разбираю критерий Стьюдента.

Во-вторых, насколько можно судить по публикациям, в медицинских исследованиях крайне важно умение правильно сравнить результаты, полученные по нескольким группам. Поэтому в книге подробно описаны методы множественного сравнения.

В-третьих, я считал, что книга по *медицинской статистике* не должна быть калькой даже с хорошего и логически стройного учебника *математической статистики*. Как показывает многолетняя практика, выслушав традиционный курс математической статистики, в котором методам проверки гипотез предшествует теория оценивания, студент, увы, не обретет понимания связи статистических методов с медицинскими задачами. Поэтому я избрал иной способ подачи материала. Стержень книги образуют проверка гипотез и оценка эффективности лечения. Я глубоко

убежден, что именно такой подход дидактически и практически отвечает задачам медицинских исследований.

Большинство использованных в книге примеров заимствовано из реальных медицинских исследований. В ряде случаев мне пришлось пойти на упрощение данных, например сделать равными объемы выборок. Эти упрощения позволили сосредоточиться на существе излагаемых методов, не отвлекаясь на технические детали. При этом если в тексте рассматривается случай выборок равного объема, то в приложении вы найдете формулы на случай выборок произвольного объема.

Готовя к печати первое издание этой книги, я задумывал его как *введение,* знакомящее с идеями, понятиями и методами статистики, — введение, за чтением которого последует более углубленное изучение традиционных курсов. Мои надежды оправдались, но, кроме того, оказалось, что многие исследователи стали пользоваться книгой как практическим пособием. Это побудило меня во втором издании более широко осветить методы множественного сравнения. В третьем издании обсуждение чувствительности критериев было пополнено рассмотрением планирования и анализа экспериментов. Наконец, в четвертом издании, которое вы держите в руках, появилась новая глава, посвященная анализу выживаемости. Помимо того, методы множественного сравнения пополнились критерием Тьюки, а в раздел, посвященный регрессионному анализу, были включены метод сравнения кривых регрессии и метод Блэнда—Алтмана для сравнения двух способов измерения.

Надо сказать, что некоторые пожелания читателей не нашли отражения в новом издании. И сделано это было совершенно сознательно. Часть читателей советовала вместо неявного использования понятий теории вероятностей дать строгое изложение ее основ. Другие предлагали дополнить книгу изложением многомерных статистических методов. В частности, предлагалось изложить методы множественной регрессии. Важность этих методов для меня вполне очевидна. Однако попытка рассмотреть их в рамках данной книги существенно изменила бы ее содержание. Что до пожеланий большей формальности, то они противоречат идее понятности и наглядности, то есть той

идее, из которой выросла эта книга и которая принесла ей успех*.

К появлению книги причастны многие люди, которым я искренне признателен. Первым человеком, от которого еще в студенческую пору я услышал понятное и практически ориентированное изложение статистики, был Джулиен Хоффман. Благодаря ему я сумел прочувствовать эту науку, а мое понимание статистических методов стало глубже. Его неиссякаемому интересу и готовности к обсуждению тонкостей я обязан тем, что узнал и — важнее — ощутил статистику настолько, чтобы задуматься о написании книги. Филипп Уилкинсон и Мэрион Нестле предложили отличные примеры и высказали массу полезных замечаний по рукописи. Стараниями Мэри Джиаммоны текст стал более понятным для студентов. Она же помогла подобрать задачи для первого издания. В работе над задачами для следующих изданий участвовали Брайан Слинкер и Джим Лайтвуд. Вирджиния Эрнстер и Сьюзен Сакс не только высказали множество полезных замечаний, но и «обкатали» первоначальный вариант рукописи, использовав его в качестве основного пособия для 300 своих студентов. Мои ассистенты Брайан Слинкер, Кен Рессер, Б. С. Апплйард и другие высказали множество тонких замечаний, которые помогли сделать материал книги более доходчивым.

Мэри Хуртадо с поразительной быстротой и точностью перепечатала рукопись. Томас Саммер, Соня Бок и Майкл Матригали помогли мне в окончательном редактировании текста в системе UNIX. Дейл Джонсон подготовил иллюстрации.

Я признателен Национальному институту здравоохранения, удостоившему меня в 1977 г. гранта, который позволил не только свободно развивать мои научные идеи, но и работать над книгой, первое издание которой увидело свет в 1981 г.

---

* Вместе с Б. Слинкером мы опубликовали специальный вводный курс, целиком посвященный множественной регрессии и многомерному дисперсионному анализу (S. A. Glantz, B. K. Slinker. Primer of Applied Regression and Analysis of Variance. New York: McGraw-Hill, 1990). Написан он в том же свободном стиле, что и настоящая книга.

С тех пор многое изменилось. Важность грамотного использования статистических методов осознается все шире. И, хотя ошибки не исчезли, все больше журналов прилагают усилия к их искоренению. Во многих из них рецензирование включает отдельный этап проверки статистической правильности предлагаемых работ. Приведу подтверждение, наиболее ощутимое для меня. Я являюсь внештатным редактором Journal of the American College of Cardiology, и моя работа состоит в выявлении статистических ошибок в поступающих работах. Доля статей, содержащих ошибки, как и раньше, составляет около половины, но теперь уже половины *предлагаемых к публикации*, а не *опубликованных* работ.

Наконец, я признателен многим читателям этой книги, студентам и преподавателям статистики, которые нашли время прислать мне вопросы, комментарии и предложения, как улучшить содержание книги. Насколько возможно, я постарался выполнить их пожелания при подготовке четвертого издания.

Многие из приведенных в книге иллюстраций — прямые потомки тех слайдов, которые я когда-то показывал на своих лекциях. Кстати, будет совсем не плохо, если, читая книгу, вы вообразите, что попали на такую лекцию. Большинство слушателей проникались критическим духом. И, как мне рассказывали, после моих выступлений перед докторантами из Калифорнийского университета те доставляли немало неприятностей последующим докладчикам, указывая на ошибки в использовании статистических методов. Надеюсь, что предлагаемая книга сделает читателя более критичным и поможет улучшить медицинскую литературу, а в конечном счете и саму медицину.

Стентон А. Гланц

Глава 1

Статистика и клиническая практика

Когда-то мне казалось, что медицинские журналы приходят к нам из идеального мира. В этом мире, недоступном простым смертным, авторы публикаций в совершенстве владеют статистическими методами, а строгие редакторы ни за что не пропустят работу со статистическими ошибками. Однако очень скоро *я понял, как легко опубликовать ошибочную и просто* бессмысленную статью, как невысок барьер на пути несостоятельной работы к читателю. Авторы и редакторы медицинских журналов живут в том же мире, что и мы, и имеют о статистике примерно такое же представление, что и остальные его обитатели. В этом суровом мире существует, помимо прочего, такая неприятная вещь, как ограничение финансирования.

ОГРАНИЧЕНИЕ ФИНАНСИРОВАНИЯ И СТАТИСТИКА

Медицина вступает в новую эру. Вплоть до середины XX века лечение мало влияло на сроки, да и сам факт выздоровления. Введение в клиническую практику инсулина, пенициллина, кор-

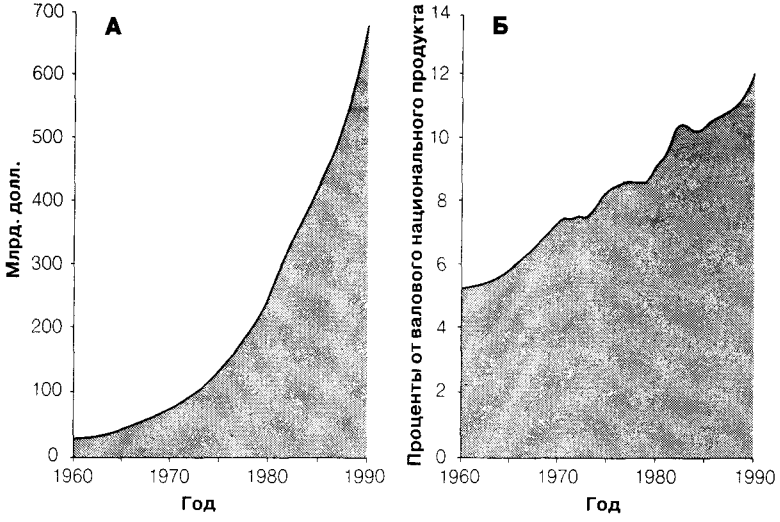


Рис. 1.1. Ежегодные расходы на здравоохранение (США, 1960—1990 гг.). **А.** Абсолютные (в миллиардах долларов). **Б.** Относительные (в процентах от валового национального продукта).

тикостероидов, витамина B_{12} радикально изменило ситуацию. Победа над ранее неизлечимыми болезнями породила веру во всесилие науки и стимулировала дальнейшие исследования. Разрабатывались все новые противоопухолевые, психотропные, гипотензивные и антиаритмические средства. Безграничный оптимизм породил почти столь же безграничное финансирование. В США расходы на медицину в 1991 г. составили 752 миллиарда долларов, или 13,2% валового национального продукта. Расходы росли как абсолютно, так и в процентах от валового национального продукта (рис. 1.1). В результате ограничение расходов на медицину сегодня превратилось в одну из первостепенных задач.

На протяжении всего этого периода, который, похоже, заканчивается, врачи и исследователи получали в свое распоряжение практически неограниченные и не обусловленные конкретными целями ресурсы. Помощь больному едва ли не выпала из числа показателей «хорошей медицины». Характерно, что даже для по-настоящему действенных методов лечения отсутствуют

достоверные оценки того, как часто и насколько эффективно они помогают*. Сложившийся подход означал не просто выбрасывание денег на ветер. Больные регулярно принимали сильнодействующие препараты или подвергались хирургическому вмешательству без серьезных оснований, но с риском серьезных осложнений.

Однако при чем тут статистика?

Когда поток не связанных с конкретными задачами средств умерит свой рост, медицинским работникам придется взглянуть на используемые ими средства с точки зрения их реальной отдачи. Потребуются строгие доказательства эффективности методов диагностики и лечения. Мало того, что придется уяснить, эффективно ли лечение, — придется выяснить также, какому проценту больных оно помогает и в какой степени. Но эти данные без помощи статистики не получишь. Естественная биологическая изменчивость, психотерапевтический эффект**, субъективность оценок — все эти факторы делают прямое суждение об эффективности лечения ненадежным. Перевести клинический опыт на язык количественных оценок — задача медицинской статистики.

Статистическому анализу может быть подвергнута не только эффективность нового метода лечения, но и эффективность работы самого врача. Так, в одном исследовании*** было показано, что больные с пиелонефритом выписываются из стационара в среднем на 2 дня раньше, если их лечение проводилось в

---

* A. L. Cockrane. Effectiveness and Efficiency: Random Reflections on Health Services. Nuffield Provincial Hospital Trust, London, 1972.

** Эффект самого факта лечения, не связанный с его физиологическим действием. Чтобы выявить психотерапевтический эффект, в клинических исследованиях применяют плацебо — неактивный препарат (например, физиологический раствор, сахарная пилюля) либо — в случае хирургического лечения — ложную операцию. В некоторых случаях, например при болях, плацебо «помогает» каждому третьему больному.

*** D. E. Knapp, D. A. Knapp, M. K. Speedie, D. M. Yager, C. I. Baker. Relationship of Inappropriate Drug Prescribing to Increased Length of Hospital Stay. *Am. J. Hosp. Pharm.*, 36:1134—1137, 1979. Эту работу мы подробно обсудим в гл. 9.

строгом соответствии с рекомендациями «Настольного справочника врача» («Physicians' desk reference»). Расходы на пребывание в стационаре составляют значительную часть всех медицинских расходов, поэтому сокращение сроков госпитализации (разумеется, не в ущерб больному) позволило бы сэкономить значительные средства. Считается, что бесконечному многообразию случаев должно соответствовать бесконечное многообразие методов лечения. Данное исследование — сильный, хотя и не бесспорный, довод в пользу большей стандартизации.

Поиск новых методов диагностики и лечения, выбор наилучшего из уже принятых — везде статистические соображения играют не последнюю роль. Чтобы принять полноправное участие в обсуждении этих вопросов, врач должен быть знаком с принципами и основными методами статистики.

До сих пор медики редко участвовали в обсуждении статистических вопросов, на первый взгляд далеких от врачебной практики и носящих сугубо технический характер. Однако по мере ужесточения требований к использованию ресурсов медикам следует научиться проверять обоснованность претензий на эффективность и с большим пониманием участвовать в распределении средств. И основой для этого служит статистика.

ДОСТОВЕРНОСТЬ И СТАТИСТИЧЕСКАЯ ЗНАЧИМОСТЬ

Рассмотрим типичный пример применения статистических методов в медицине. Создатели препарата предполагают, что он увеличивает диурез пропорционально принятой дозе. Для проверки этого предположения они назначают пяти добровольцам разные дозы препарата. По результатам наблюдений строят график зависимости диуреза от дозы (рис. 1.2А). Зависимость видна невооруженным глазом. Исследователи поздравляют друг друга с открытием, а мир — с новым диуретиком.

На самом деле данные позволяют достоверно утверждать лишь то, что зависимость диуреза от дозы наблюдалась у *этих* пяти добровольцев. То, что эта зависимость проявится у *всех* людей, которые будут принимать препарат, — не более чем предположе-

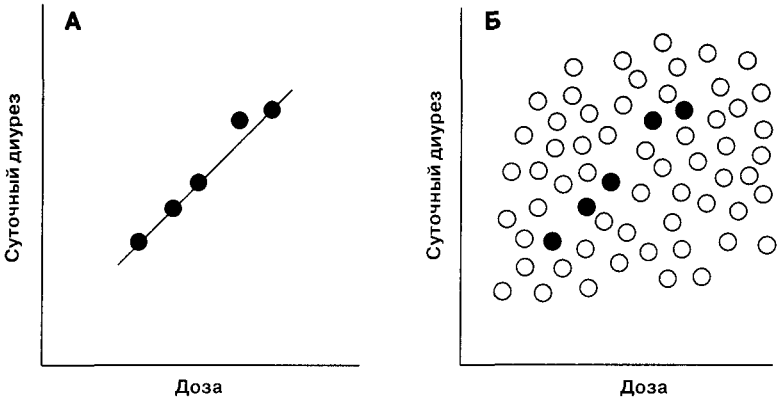


Рис. 1.2. А. У 5 добровольцев измерили суточный диурез после приема разных доз препарата (предполагаемого диуретика). Зависимость диуреза от дозы, казалось бы, налицо: чем больше доза — тем больше диурез. Можно ли считать диуретический эффект препарата доказанным? **Б.** Такую картину мы увидели бы, если бы могли исследовать связь дозы и диуреза у *всех* людей: зависимости нет и в помине. Пять человек, вошедших в первоначальное исследование, помечены черным. В данном случае мнимая зависимость порождена случайностью. С помощью статистических методов можно оценить вероятность подобной ошибки.

ние. Нельзя сказать, что оно беспочвенно — иначе зачем ставить эксперименты?

Но вот препарат поступил в продажу. Все больше людей принимают его в надежде увеличить свой диурез. И что же мы видим? Мы видим рис. 1.2Б, который свидетельствует об отсутствии какой-либо связи между дозой препарата и диурезом. Черными кружками отмечены данные первоначального исследования. Статистика располагает методами, позволяющими оценить вероятность получения столь «непредставительной», более того, сбивающей с толку выборки. Оказывается, в отсутствие связи между диурезом и дозой препарата полученная «зависимость» наблюдалась бы примерно в 5 из 1000 экспериментов. Итак, в данном случае исследователям просто не повезло. Если бы они применили даже самые совершенные статистические методы, это все равно не спасло бы их от ошибки.

Этот вымышленный, но совсем не далекий от реальности пример мы привели не для того, чтобы указать на бесполез-

ность статистики. Он говорит о другом — о вероятностном характере ее выводов. В результате применения статистического метода мы получаем не истину в последней инстанции, а всего лишь оценку вероятности того или иного предположения. Кроме того, каждый статистический метод основан на собственной математической модели, и результаты его правильны настолько, насколько эта модель соответствует действительности.

ДОВЕРЯЙ, НО ПРОВЕРЯЙ

О новых методах диагностики и лечения врачи узнают главным образом из публикаций в медицинских журналах. Познания читателей в статистике обычно скромны, поэтому выводы авторов им приходится принимать на веру. Это было бы не так страшно, если бы публикации предшествовала серьезная проверка результатов. К сожалению, проводится она далеко не всегда.

На рис. 1.3 суммированы результаты четырех исследований использования статистических методов в статьях, опубликованных в медицинских журналах с 1950 по 1976 г*. Разумеется, ис-

---

* О. Б. Росс мл. (O. B. Ross, Jr. Use of controls in medical research. *JAMA*, 145:72—75, 1951) рассмотрел 100 статей, опубликованных в Journal of the American Medical Association, American Journal of Medicine, Annals of Internal Medicine, Archives of Neurology and Psychiatry и American Journal of Medical Sciences в 1950 г. Р. Бэдгли (R. F. Badgley. An assessment of research methods reported in 103 scietific articles from two Canadian medical journals. *Can. M. A. J.*, 85:256—260, 1961) рассмотрел 103 статьи, опубликованные в журналах Canadian Medical Association Journal и Canadian Journal of Public Health в 1960 г. С. Шор и И. Картен (S. Schor, I. Karten. Statistical evaluation of medical journal manuscripts, *JAMA*, 195:1123—1128, 1966) рассмотрели 295 статей, опубликованных в журналах Annals of Internal Medicine, New England Journal of Medicine, Archives of Surgery, American Journal of Medicine, Journal of Clinical Investigation, American Archives of Neurology, Archives of Pathology и Archives of Internal Medicine в 1964 г. С. Гор, И. Джонс и Э. Риттер (S. Gore, I. G. Jones, E. C. Rytter. Misuses of statistical methods: critical assessment of articles in B. M. J. from January to March, 1976, *Br. Med. J.*, 1(6053):85—87, 1977) рассмотрели 77 статей,

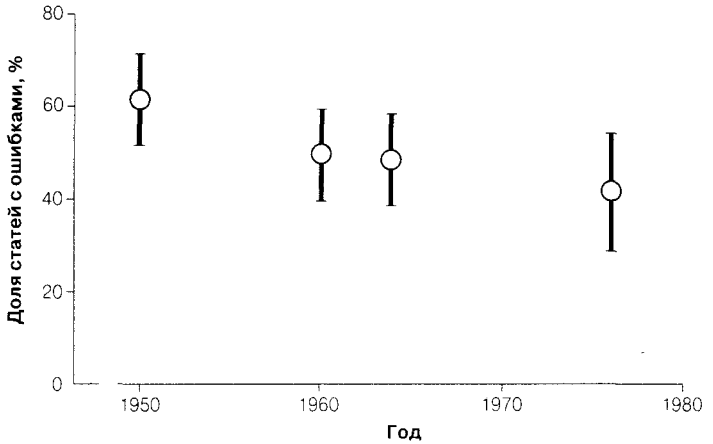


Рис. 1.3. Доля медицинских статей, содержащих статистические ошибки. Невозможно рассмотреть все статьи, публикуемые в медицинских журналах, поэтому долю определяли по некоторой случайной выборке. В результате получается оценка истинной доли статей с ошибками, на рисунке эти оценки показаны кружками. Вертикальные отрезки — это доверительный интервал, то есть пределы, в которых, скорее всего, находится истинная доля статей с ошибками.

следования могли охватить лишь часть напечатанного, поэтому выявленная в исследованиях доля статей, содержащих статистические ошибки, служит лишь приближенной оценкой истинной доли. Вертикальные черточки на рис. 1.3 указывают диапазон, называемый *доверительным интервалом*, в который с высокой вероятностью попадает истинная доля статей с ошибками. Вычисление доверительных интервалов — один из разделов статистики, с которым нам предстоит познакомиться. Как мы ви-

---

опубликованных в журнале British Medical Journal в 1976 г. Сравнительно недавнее изучение более ограниченной подборки журналов показало, что проблема статистических ошибок в медицинских публикациях не потеряла своей значимости. (См. J. Davies. A critical survey of scientific methods in two psychiatry journals, *Aust. N. Z. J. Psych.*, 21:367—373, 1987; D. F. Cruess. Review of the use of statistics in the American Journal of Tropical Medicine and Hygiene for January—December 1988. *Am. J. Trop. Med. Hyg.*, 41:619—626, 1990.)

дим, статистические ошибки встречаются примерно в половине статей. Однако дальнейшие исследования показали, что журналам, в которых взяли за правило обращать внимание не только на медицинскую, но и статистическую сторону дела, удалось существенно снизить долю ошибочных статей. Эта доля нимало не изменилась в тех журналах, которые так и не ввели статистического рецензирования.

Врачам известно множество методов диагностики и лечения, эффективность которых была «доказана» статистическими методами и которые тем не менее канули в Лету, не выдержав проверки практикой. А сколь часто приходится читать статьи, в которых статистические манипуляции с одними и теми же данными приводят к прямо противоположным выводам. Все это наводит читателя на мысль, что статистические методы либо ненадежны, либо слишком трудны для понимания, либо вообще не более чем инструмент недобросовестного исследователя. Между тем даже начального знакомства со статистикой в сочетании со здравым смыслом обычно достаточно, чтобы понять, что предлагает нам автор в качестве «доказательств». По иронии судьбы ошибки редко связаны с тонкими статистическими вопросами. Как правило, это простейшие ошибки, такие, как отсутствие контрольной группы, использование неслучайных выборок или пренебрежение статистической проверкой гипотез. По неизвестным науке причинам такие ошибки неизменно смещают результаты исследования в пользу предлагаемого автором метода.

Вред, приносимый ошибками такого рода, очевиден. Исследователь заявляет о «статистически достоверном» эффекте лечения, редактор помещает статью в журнал, врач, неспособный критически оценить публикацию, применяет неэффективный метод лечения. В конце этой цепи находится больной, который и расплачивается за все, подвергаясь ненужному риску и не получая действительно эффективного лечения. Не следует сбрасывать со счетов и ущерб от самого факта проведения бессмысленных исследований. Деньги и подопытные животные приносятся в жертву науке, больные рискуют ради сбора ошибочно интерпретируемых данных.

Сегодня грамотная проверка эффективности лечения стано-

вится первоочередной задачей. Исследования должны тщательно планироваться, а результаты правильно интерпретироваться.

ОШИБКИ ВЕЧНЫ?

Поскольку описанные ошибки совершаются в массовом порядке, ничто не побуждает исследователей корректно использовать статистические методы. Редко кому приходилось слышать критические замечания на сей счет. Наоборот, исследователи часто опасаются, что их коллеги, а особенно рецензенты, сочтут грамотно и полно изложенную статистическую процедуру высокомерной теоретизацией.

Журналы призваны быть оплотом качества научных исследований. В некоторых редакциях действительно осознали, что их рецензенты не слишком сведущи в использовании элементарной статистики, и изменили саму процедуру рецензирования. Теперь перед тем как направить рукопись на рецензию, ее тщательно проверяют на предмет правильности использования статистических методов. Результатом этого нередко становится пересмотр используемых в статье статистических методов, а иногда и самих выводов*.

Но большинство редакторов, похоже, убеждены, что каждый рецензент рассматривает статистическую сторону работы столь же тщательно, сколь и собственно медицинскую. Неясно, однако, как он может это сделать — ведь даже авторы ведущих медицинских журналов, упоминая статистическую проверку гипотез, редко затрудняют себя указанием, какой именно критерий был использован.

Коротко говоря, для грамотного чтения медицинской литературы необходимо научиться понимать и оценивать правильность применения статистических методов, используемых для анализа результатов. К счастью, основные идеи, которыми необ-

---

* Подробнее о существующей в редакциях практике работы с рукописями см. M. J. Gardner, J. Bond. An exploratory study of statistical assessment of papers published in the British Medical Journal. *JAMA,* 263:1355—1357, 1990, а также S. A. Glantz. It is all in the numbers. *J. Am. Coll. Cardiol.*, 21:835—837, 1993.

ходимо овладеть вдумчивому читателю (и, конечно, вдумчивому исследователю), довольно просты. В следующей главе мы приступим к их обсуждению.

Глава 2

Как описать данные

В этой книге мы встретимся с двумя типами задач. Первый тип задач — как сжато описать данные. Этими задачами занимается так называемая описательная статистика. Задачи второго типа связаны с оценкой статистической значимости различий и вообще с проверкой гипотез. В этой главе мы рассмотрим задачи первого типа — как наилучшим образом описать данные.

Если значения интересующего нас признака у большинства объектов близки к их среднему и с равной вероятностью отклоняются от него в большую или меньшую сторону, лучшими характеристиками совокупности будут само *среднее* значение и *стандартное отклонение*. Напротив, когда значения признака распределены несимметрично относительно среднего, совокупность лучше описать с помощью *медианы* и *процентилей*.

Возможно, сказанное давно вам известно. Тогда смело переходите к следующей главе. Тех же, для кого термины вроде *процентиля* звучат туманно, мы приглашаем приступить к изучению марсиан.

Поначалу займемся каким-нибудь *количественным* признаком, например ростом. Чтобы попусту не фантазировать, слетаем на Марс и измерим всех марсиан, благо их всего две сотни. Результаты приведены на рис. 2.1 (мы округлили рост до целого числа сантиметров). Каждому марсианину соответствует кружок, так что, например, два кружка над числом 30 означают, что имеются два марсианина ростом 30 см. Рис. 2.1 — это *распределение* марсиан по росту. Мы видим, что рост большинства марсиан — от 35 до 45 см. Коротышек (ниже 30 см) совсем немного — всего трое, и столько же великанов (выше 50 см).

Окрыленные успехом марсианского проекта, мы решаем измерить венерианцев. Легко находим деньги на путешествие и, вооружившись линейками, измеряем всех 150 обитателей Венеры. Научный отчет об экспедиции будет звучать так: «Редко встретишь венерианца ниже 10 см или выше 20 см, а чаще попадаются 15-сантиметровые, см. рис. 2.2».

Но вот остались позади нелегкие межпланетные перелеты. Настала пора скрупулезного анализа данных. Сравним рис. 2.1 и 2.2. Мы видим, что венерианцы ниже марсиан и что интервал, в

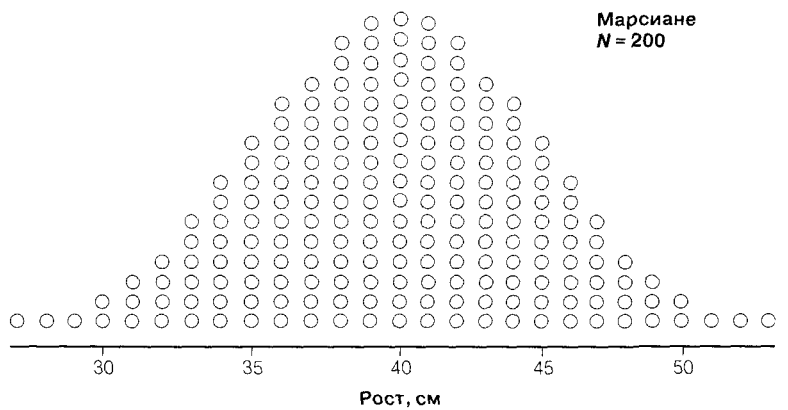


Рис. 2.1. Распределение марсиан по росту. Каждому марсианину соответствует кружок. Обратите внимание, что марсиан среднего роста (около 40 см) больше всего и что высокорослых столько же, сколько коротышек (распределение симметрично).

который умещается рост всех марсиан, шире, чем соответствующий интервал для венерианцев. Ширина интервала, в который попадают почти все марсиане (194 из 200) — 20 см (от 30 до 50 см). Рост большинства венерианцев (144 из 150) умещается в интервал от 10 до 20 см, то есть имеет ширину всего лишь 10 см. Несмотря на эти различия, между двумя совокупностями инопланетян имеется и существенное сходство. В обеих рост любого члена скорее близок к середине распределения, нежели заметно от нее удален, и одинаково вероятно может быть как выше, так и ниже середины. Распределения на рис. 2.1 и 2.2 имеют схожую форму и приближенно определяются одной и той же формулой.

Раз существует множество похожих распределений, значит, для характеристики одного из них достаточно указать, чем оно отличается от других, ему подобных, то есть всю собранную информацию мы можем свести к нескольким числам, которые называются *параметрами* распределения. Это *среднее* значение и *стандартное отклонение*.

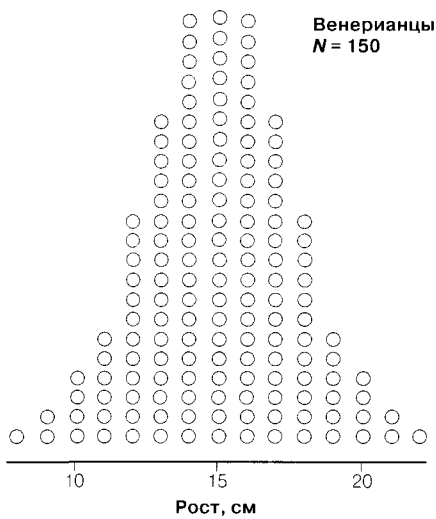


Рис. 2.2. Распределение венерианцев по росту. Венерианцы ниже марсиан, разброс значений меньше. Однако по форме распределения, напоминающей колокол, венерианцы и марсиане схожи друг с другом.

СРЕДНЕЕ

Расположив мысленно распределения марсиан и венерианцев на одной шкале роста, мы увидим, что распределение венерианцев находится ниже, чем распределение марсиан. Характеристика положения распределения на числовой оси называется средним. Среднее по совокупности обозначают греческой буквой μ (читается «мю») и вычисляют по формуле:

$$\text{Среднее по совокупности} = \frac{\text{Сумма значений признака для всех членов совокупности}}{\text{Число членов совокупности}}.$$

Эквивалентное математическое выражение имеет вид

$$\mu = \frac{\Sigma X}{N},$$

где X — значение признака, N — число членов совокупности. Как всегда, большая греческая буква Σ (читается «сигма») обозначает сумму. Подставив в формулу добытые нами данные, получим ценное дополнение к научному отчету: средний рост марсиан 40 см, а венерианцев — 15 см.

СТАНДАРТНОЕ ОТКЛОНЕНИЕ

Еще на Венере мы заметили, что тамошние жители более однородны по росту, нежели марсиане. Хотелось бы и это впечатление оформить количественно, то есть иметь показатель разброса значений относительно среднего. Ясно, что для характеристики разброса все равно, в какую сторону отклоняется значение — в большую или меньшую. Иными словами, отрицательные и положительные отклонения должны вносить равный вклад в характеристику разброса. Воспользуемся тем, что квадраты двух равных по абсолютной величине чисел равны между собой, и вычислим средний квадрат отклонения от среднего. Этот показатель носит название *дисперсии* и обозначается σ^2. Чем больше разброс значений, тем больше дисперсия. Дисперсию вычисляют по формуле:

$$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N}.$$

Как видно из формулы, дисперсия измеряется в единицах, равных квадрату единицы измерения соответствующей величины. Например, дисперсия измеряемого в сантиметрах роста сама измеряется в квадратных сантиметрах. Это довольно неудобно. Поэтому чаще используют квадратный корень из дисперсии — *стандартное отклонение* σ (маленькая греческая буква «сигма»):

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\Sigma(X-\mu)^2}{N}}.$$

Стандартное отклонение измеряется в тех же единицах, что исходные данные. Например, стандартное отклонение роста марсиан составляет 5 см, а венерианцев — 2,5 см.

НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

Таблица 2.1 сжато представляет то, что мы узнали о марсианах и венерианцах. Таблица очень информативна, из нее можно узнать об объеме совокупности, о среднем росте и о том, насколько велик разброс относительно среднего.

Вновь обратившись к рис. 2.1 и 2.2, мы обнаружим, что на обеих планетах *рост примерно 68% обитателей отличается от среднего не более чем на одно стандартное отклонение и примерно 95% — на два стандартных отклонения.* Подобные распределения встречаются очень часто. Можно сказать, что это происходит всегда, когда некая величина отклоняется от средней под действием множества слабых, независимых друг от друга факто-

Таблица 2.1. Параметры распределения марсиан и венерианцев по росту

	Объем совокупности	Среднее, см	Стандартное отклонение, см
Марсиане	200	40	5,0
Венерианцы	150	15	2,5

ров. Распределение такого рода называется *нормальным* (или гауссовым) и описывается формулой:

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}.$$

Заметим, что нормальное распределение *полностью* определяется средней μ и стандартным отклонением σ. Поэтому сведения в табл. 2.1 — это не просто удачное представление данных, но также и полное их описание.

МЕДИАНА И ПРОЦЕНТИЛИ

И снова в путь! Обогатившись теоретическими познаниями, мы отправляемся на Юпитер. Здесь мы не только измеряем всех до одного юпитериан, но также подсчитываем среднее и стандартное отклонение роста для всей их совокупности. Оказывается, средний рост юпитериан — 37,6 см, а его стандартное отклонение — 4,5 см. Можно заключить, что юпитериане очень похожи на марсиан, ведь близки оба параметра, определяющие нормальное распределение — среднее и стандартное отклонение.

Однако если взглянуть на исходные данные по юпитерианам (рис. 2.3А), то обнаружится совершенно иная картина. На самом деле типичный юпитерианин довольно приземист — около 35 см, то есть на добрых 5 см ниже марсианина. И только небольшая группа долговязых смещает значения стандартного отклонения и среднего, вводя ученых в заблуждение!

Итак, рост произвольно выбранного юпитерианина вовсе не равновероятно может оказаться выше или ниже среднего, то есть распределение юпитериан по росту *асимметрично*. В такой ситуации полагаться на среднее и стандартное отклонение нельзя. На рис. 2.3Б изображено нормальное распределение для совокупности с теми же самыми значениями среднего и стандартного отклонения, что и на рис. 2.3А. Оно ничуть не похоже на распределение юпитериан. Таким образом, доверившись среднему и стандартному отклонению, мы получим превратное представ-

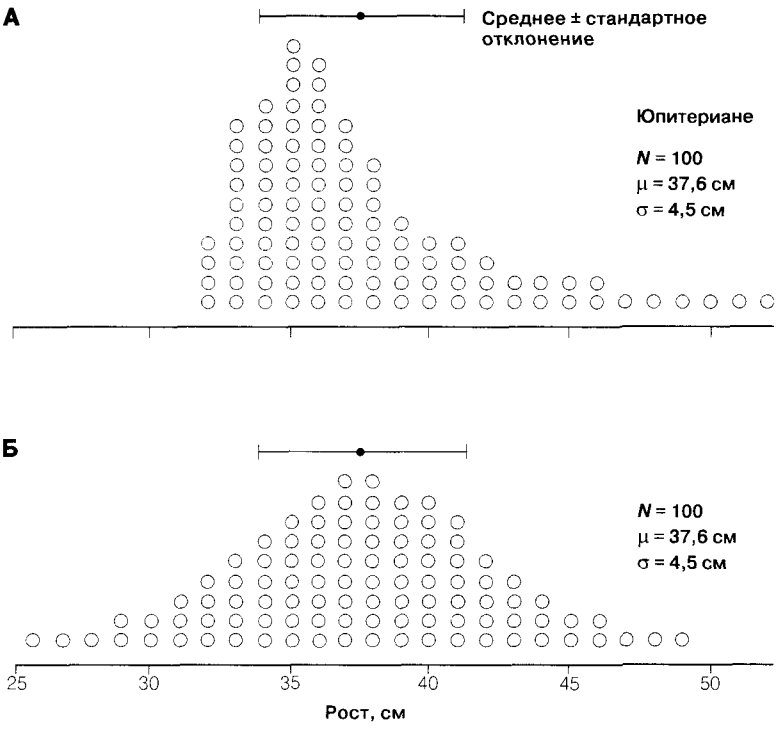


Рис. 2.3. Если распределение асимметрично, полагаться на среднее и стандартное отклонение нельзя. **А.** Распределение юпитериан по росту. **Б.** Нормальное распределение с теми же средним и стандартным отклонением: несмотря на тождественность параметров, оно ничуть не похоже на реальное распределение юпитериан.

ление о совокупности, не подчиняющейся нормальному распределению.

Для описания таких данных лучше подходит не среднее, а *медиана.* Медиана — это значение, которое делит распределение пополам: половина значений больше медианы, половина — меньше (точнее, не больше). Из рис. 2.4А видно, что ровно половина юпитериан выше 36 см. Стало быть, 36 см — это медиана роста юпитериан.

Для характеристики разброса роста юпитериан найдем значения, не выше которых оказались 25 и 75% результатов измере-

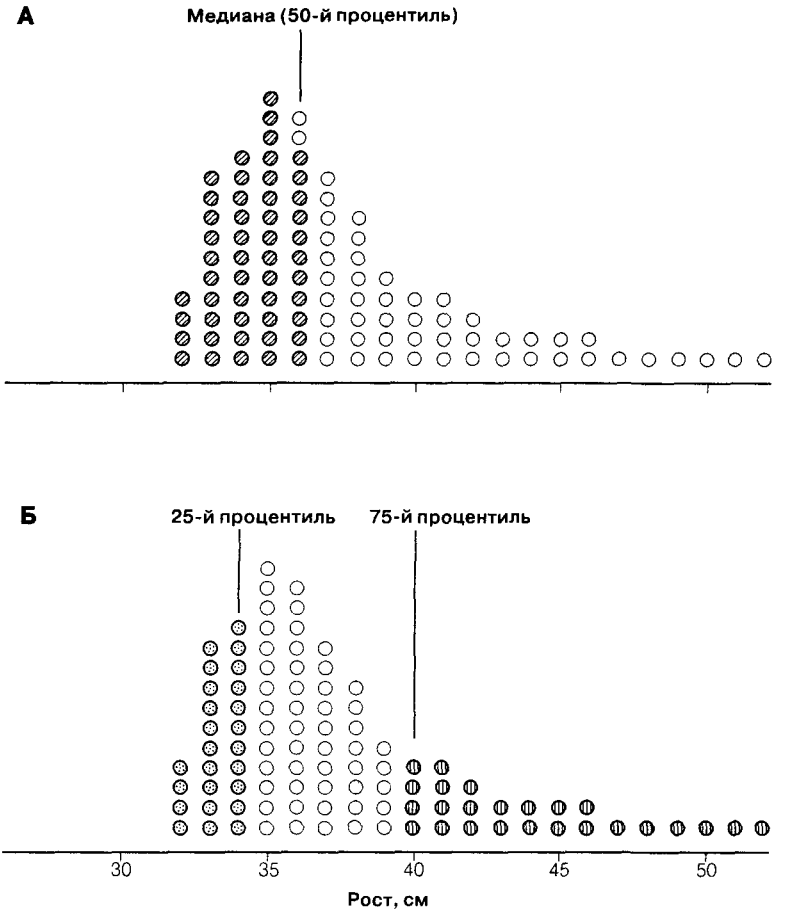


Рис. 2.4. Для описания асимметричного распределения следует использовать медиану и процентили. Медиана — это значение, которое делит распределение пополам. **А.** Медиана роста юпитериан — 36 см. **Б.** 25-й и 75-й процентили отсекают четверть самых низких и четверть самых высоких юпитериан. 25-й процентиль ближе к медиане, чем 75-й — это говорит об асимметричности распределения.

ния. Эти величины называются 25-м и 75-м *процентилями*. Если медиана делит распределение пополам, то 25-й и 75-й процентили отсекают от него по четвертушке. (Саму медиану, кстати, можно считать 50-м процентилем.) Для юпитериан, как видно из рис. 2.4Б, 25-й и 75-й процентили равны соответственно 34 см и 40 см. Конечно, медиана и процентили, в отличие от среднего и стандартного отклонения, не дают полного описания распределения. Однако между 25-м и 75-м процентилями находится половина значений — значит, мы можем судить, каков ростом средний юпитерианин. По положению медианы относительно 25-го и 75-го процентилей можно судить о том, насколько асимметрично распределение. И наконец, теперь мы примерно знаем, кто на Юпитере считается высоким (выше 75-го процентиля), а кто ростом не вышел (ниже 25-го процентиля).

Для описания распределения чаще всего применяют 25-й и 75-й процентили. Однако можно рассчитывать любые другие процентили. Например, в качестве границ нормы лабораторных показателей часто используют 5-й и 95-й процентили.

Вычисление процентилей — хороший способ разобраться в том, насколько распределение близко к нормальному. Напомним, что для нормального распределения 95% значений заключено в пределах двух стандартных отклонений от среднего и 68% — в пределах одного стандартного отклонения; медиана совпадает со средним. Соответствие между процентилями и числом стандартных отклонений от среднего таково (см. также рис. 2.5):

Процентили	Отклонения от среднего
2,5	$\mu - 2\sigma$
16	$\mu - \sigma$
50	μ
84	$\mu + \sigma$
97,5	$\mu + 2\sigma$

Если соответствие между процентилями и отклонениями от среднего не слишком отличается от приведенного, то распределение близко к нормальному и его можно описать при помощи среднего и стандартного отклонения.

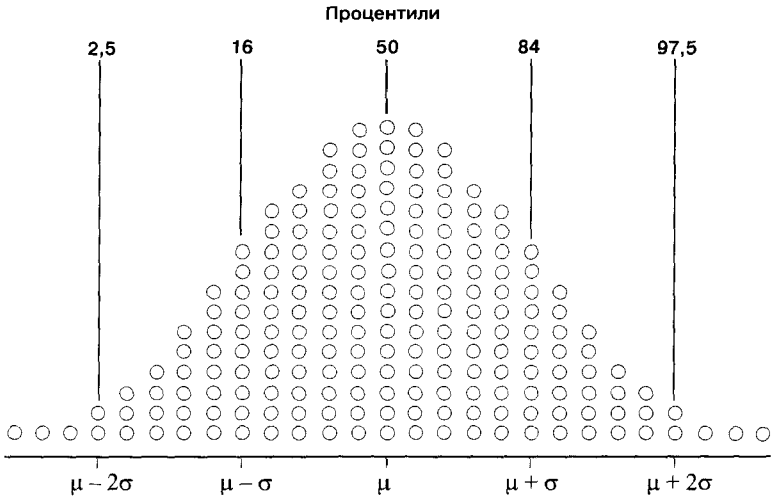


Рис. 2.5. Нормальное распределение: соответствие между числом стандартных отклонений от среднего и процентилями.

Есть еще одна, и очень важная, причина, по которой нужно знать, близко ли распределение к нормальному. Дело в том, что многие методы проверки гипотез, в частности рассматриваемые в гл. 2, 4 и 9, основаны на предположении, что распределение близко к нормальному. Только в этом случае эти методы будут надежны. (Методы, не требующие нормальности распределения, изложены в гл. 10.)

ВЫБОРОЧНЫЕ ОЦЕНКИ

До сих пор нам удавалось получить данные обо *всех* объектах совокупности, поэтому мы могли точно рассчитать значения среднего, дисперсии и стандартного отклонения. На самом деле обследовать все объекты совокупности удается редко: обычно довольствуются изучением *выборки,* полагая, что эта выборка отражает свойства совокупности. Выборку, отражающую свойства совокупности, называют представительной. Имея дело с выборкой, мы, конечно, не узнаем точных значений среднего и стан-

дартного отклонения, но можем оценить их. Оценка среднего, вычисленная по выборке, называется *выборочным средним*. Выборочное среднее обозначают $\bar{X}$ и вычисляют по формуле:

$$\bar{X} = \frac{\Sigma X}{n}.$$

где n — объем выборки.

Оценка стандартного отклонения называется *выборочным стандартным отклонением* (s) и определяется следующим образом:

$$s = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n-1}}.$$

Эта формула отличается от формулы для стандартного отклонения по совокупности. Во-первых, среднее μ заменяется его выборочной оценкой — $\bar{X}$. Во-вторых, в знаменателе из числа членов выборки вычитается единица. Строгое обоснование последнего требует основательной математической подготовки, поэтому ограничимся следующим объяснением. Разброс значений в пределах выборки никогда не бывает столь большим, как во всей совокупности, и деление не на n, а на $n-1$ компенсирует возникающее занижение оценки стандартного отклонения.

Подытожим. Если известно, что выборка скорее всего принадлежит к совокупности с нормальным распределением, лучше всего использовать выборочное среднее и выборочное стандартное отклонение. Если есть основания полагать, что распределение в совокупности отличается от нормального, следует использовать медиану, 25-й и 75-й процентили.

НАСКОЛЬКО ТОЧНЫ ВЫБОРОЧНЫЕ ОЦЕНКИ

Выборочное среднее и выборочное стандартное отклонение есть оценки среднего и стандартного отклонения для совокупности, вычисленные по случайной выборке. Понятно, что разные выборки дадут разные оценки. Для характеристики точности выборочных оценок используют *стандартную ошибку*. Стандартную ошибку можно подсчитать для любого показателя, но сейчас мы остановимся на *стандартной ошибке среднего* — она позволяет

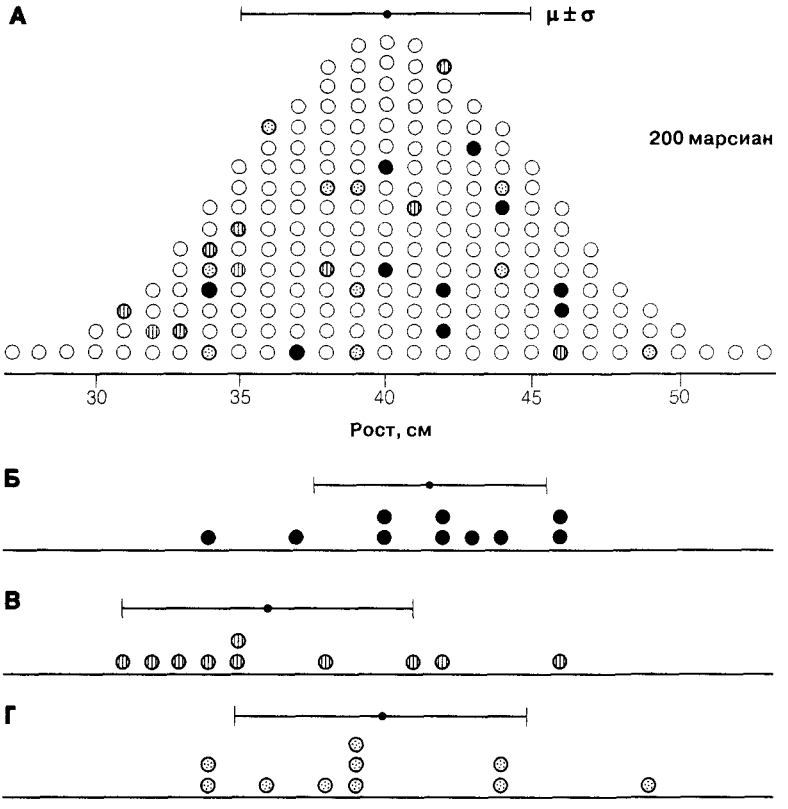


Рис. 2.6. Три случайные выборки из одной совокупности дают три разных оценки среднего и стандартного отклонения.

оценить точность, с которой выборочное среднее характеризует значение среднего по всей совокупности.

На рис. 2.6А представлено уже знакомое нам распределение марсиан по росту. Мы уже знаем рост каждого марсианина. Посмотрим, что получится, если оценивать средний рост по выборке объемом, скажем, 10 марсиан.

Из 200 обитателей Марса наугад выберем 10 и пометим их черными кружками (рис. 2.6А). На рис. 2.6Б эта выборка изображена в виде, принятом в журнальных публикациях. Точка и два

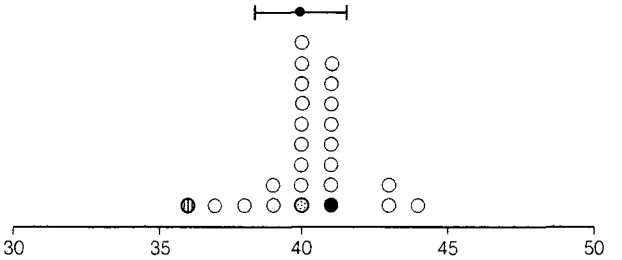


Рис. 2.7. Такое распределение мы получим, выбрав 25 раз по 10 марсиан из совокупности, представленной на рис. 2.6А, и рассчитав среднее для каждой выборки (средние для трех выборок с рис. 2.6 показаны заполненными кружками). Если построить распределение средних для всех возможных выборок, оно окажется нормальным. Среднее этого распределения будет равно среднему той совокупности, из которой извлекаются выборки. Стандартное отклонение этого распределения называется стандартной ошибкой среднего.

отрезка по бокам от нее изображают выборочное среднее (X = = 41,5 см) и выборочное стандартное отклонение (s = 3,8 см). Эти значения близки, но не равны среднему по совокупности (μ = 40 см) и стандартному отклонению (σ = 5 см).

Извлечем еще одну случайную выборку того же объема. Результат показан на рис. 2.6В. На рис. 2.6А попавшие в эту выборку марсиане изображены заштрихованными кружками. Выборочное среднее (36 см) по-прежнему близко к среднему по совокупности, хотя и отличается от него; что касается выборочного стандартного отклонения (5 см), то на этот раз оно совпало со стандартным отклонением по совокупности.

На рис. 2.6Г представлена третья выборка. Попавшие в нее марсиане на рис. 2.6А изображены кружками с точками. Среднее и стандартное отклонение для этой выборки составляют соответственно 40 и 5 см.

Теперь пора поставить добычу случайных выборок на промышленную основу. Рассмотрим *совокупность средних для каждой из возможных выборок по 10 марсиан*. Общее число таких выборок превышает 10^{16}. Три из них мы уже обследовали. Средние по этим выборкам представлены на рис. 2.7 в виде заполненных кружков. Пустые кружки — это средние еще для 22 выборок. Итак, теперь каждому выборочному среднему соответствует кру-

жок, точно так же, как до сих пор кружки соответствовали отдельному объекту.

Посмотрим на рис. 2.7. Набор из 25 выборочных средних имеет колоколообразное распределение, похожее на нормальное. Это не случайно. Можно доказать, что если переменная представляет собой сумму большого числа независимых переменных, то ее распределение стремится к нормальному, какими бы ни были распределения переменных, образующих сумму. Так как выборочное среднее определяется именно такой суммой, его распределение стремится к нормальному, причем чем больше объем выборок, тем точнее приближение. (Если выборки принадлежат совокупности с нормальным распределением, распределение выборочных средних будет нормальным независимо от объема выборок.)

Поскольку распределение на рис. 2.7 нормальное, его можно описать с помощью среднего и стандартного отклонения.

Так как среднее значение для рассматриваемых 25 точек есть среднее величин, которые сами являются средними значениями, обозначим его $\overline{X}_{\overline{X}}$. Аналогично, стандартное отклонение обозначим $s_{\overline{X}}$. По формулам для среднего и стандартного отклонения находим: $\overline{X}_{\overline{X}} = 40$ см и $s_{\overline{X}} = 1{,}6$ см.

Среднее выборочных средних $\overline{X}_{\overline{X}}$ оказалось равно среднему μ всей совокупности из 200 марсиан. Ничего неожиданного в этом нет. Действительно, если бы мы провели исследования всех возможных выборок, то каждый из 200 марсиан был бы выбран равное число раз. Итак, *среднее выборочных средних совпадет со средним по совокупности.*

Интересно, равно ли $s_{\overline{X}}$ стандартному отклонению σ совокупности из 200 марсиан? Стандартное отклонение для совокупности выборочных средних $s_{\overline{X}}$ равно 1,6 см, а стандартное отклонение самой совокупности — 5 см. Почему $s_{\overline{X}}$ меньше, чем σ? В общих чертах это можно понять, если учесть, что в случайную выборку редко будут попадать одни только коротышки и одни гиганты. Чаще их будет примерно поровну, и отклонения роста от среднего будут сглаживаться. Даже в выборке, куда попадут 10 самых высоких марсиан, средний рост составит только 50 см, тогда как рост самого высокого марсианина — 53 см.

Подобно тому как стандартное отклонение исходной выбор-

ки из 10 марсиан s служит оценкой изменчивости роста марсиан, $s_{\bar{X}}$ является оценкой изменчивости значений средних для выборок по 10 марсиан в каждой. Таким образом, величина $s_{\bar{X}}$ служит мерой точности, с которой выборочное среднее $\bar{X}$ является оценкой среднего по совокупности μ. Поэтому $s_{\bar{X}}$ носит название *стандартной ошибки среднего*.

Чем больше выборка, тем точнее оценка среднего и тем меньше его стандартная ошибка. Чем больше изменчивость исходной совокупности, тем больше изменчивость выборочных средних; поэтому стандартная ошибка среднего возрастает с увеличением стандартного отклонения совокупности.

Истинная стандартная ошибка среднего по выборкам объемом n, извлеченным из совокупности, имеющей стандартное отклонение σ, равна*:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

Собственно стандартная ошибка — это наилучшая оценка величины $\sigma_{\bar{X}}$ по одной выборке:

$$s_{\bar{X}} = \frac{s}{\sqrt{n}},$$

где s — выборочное стандартное отклонение.

Так как возможные значения выборочного среднего стремятся к нормальному распределению, истинное среднее по совокупности примерно в 95% случаев лежит в пределах 2 стандартных ошибок выборочного среднего.

Как уже говорилось, распределение выборочных средних приближенно всегда следует нормальному распределению независимо от распределения совокупности, из которой извлечены выборки. В этом и состоит суть утверждения, называемого *центральной предельной теоремой*. Эта теорема гласит следующее.

- Выборочные средние имеют приближенно нормальное распределение *независимо от* распределения исходной совокупности, из которой были извлечены выборки.

---

* Вывод этой формулы приведен в гл. 4.

- Среднее значение всех возможных выборочных средних равно среднему исходной совокупности.
- Стандартное отклонение всех возможных средних по выборкам данного объема, называемое стандартной ошибкой среднего, зависит как от стандартного отклонения совокупности, так и от объема выборки.

На рис. 2.8 показано, как связаны между собой выборочное среднее, выборочное стандартное отклонение и стандартная ошибка среднего и как они изменяются в зависимости от объема выборки*. По мере того как мы увеличиваем объем выборки, выборочное среднее $\bar{X}$ и стандартное отклонение s дают все более точные оценки среднего μ и стандартного отклонения σ по совокупности. Увеличение точности оценки среднего отражается в уменьшении стандартной ошибки среднего $\sigma_{\bar{X}}$. Набрав достаточное количество марсиан, можно сделать стандартную ошибку среднего сколь угодно малой. В отличие от стандартного отклонения стандартная ошибка среднего ничего не говорит о разбросе данных — она *лишь показывает точность* выборочной оценки среднего.

Хотя разница между стандартным отклонением и стандартной ошибкой среднего совершенно очевидна, их часто путают. Большинство исследователей приводят в публикациях значение стандартной ошибки среднего, которая заведомо меньше стандартного отклонения. Авторам кажется, что в таком виде их данные внушают больше доверия. Может быть, так оно и есть, однако беда в том, что стандартная ошибка среднего измеряет именно точность оценки среднего, но никак не разброс данных, который и интересен читателю. Мораль состоит в том, что, описывая совокупность, всегда нужно приводить значение стандартного отклонения.

---

* Рис. 2.8 получился следующим образом. Из совокупности марсиан (рис. 2.1) взяли наугад двух марсиан. По этой выборке вычислили $\bar{X}$, s и $s_{\bar{X}}$. Потом опять же наугад выбрали еще одного марсианина и, добавив его к выборке, снова рассчитали эти показатели. Добавляя каждый раз по одному случайно выбранному марсианину, объем выборки довели до 100. Если бы мы повторили эксперимент, очередность извлечения марсиан была бы иной и рисунок выглядел бы немного иначе.

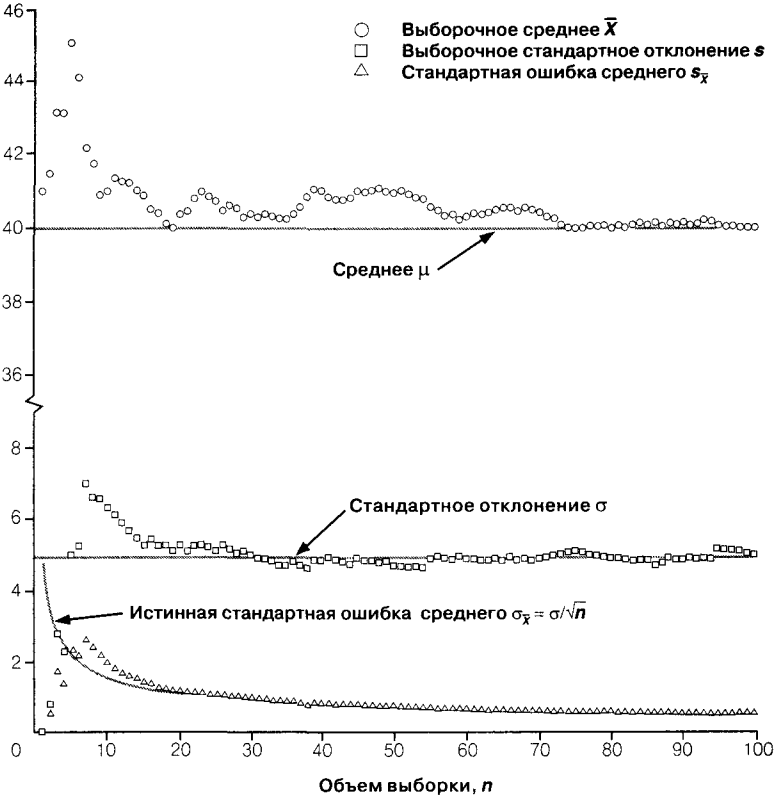


Рис. 2.8. С увеличением объема выборки возрастает точность оценки параметров распределения. Выборочное среднее $\bar{X}$ стремится к среднему в совокупности μ, выборочное стандартное отклонение s стремится к стандартному отклонению в совокупности σ, а стандартная ошибка среднего стремится к нулю.

Рассмотрим пример, позволяющий почувствовать различие между стандартным отклонением и стандартной ошибкой среднего, а также уяснить, почему не следует пренебрегать стандартным отклонением. Положим, исследователь, обследовав выборку из 20 человек, пишет в статье, что средний сердечный выброс составлял 5,0 л/мин со стандартным отклонением 1 л/мин. Мы знаем, что 95% нормально распределенной совокупности попадает в интервал среднее плюс-минус два стандартных отклоне-

ния. Тем самым, из статьи видно, что почти у всех обследованных сердечный индекс составил от 3 до 7 л/мин. Такие сведения весьма полезны, их легко использовать во врачебной практике.

Увы, приведенный пример далек от реальности. Скорее автор укажет не стандартное отклонение, а стандартную ошибку среднего. Тогда из статьи вы узнаете, что «сердечный выброс составил 5,0 ± 0,22 л/мин». И если бы мы спутали стандартную ошибку среднего со стандартным отклонением, то пребывали бы в уверенности, что 95% совокупности заключено в интервал от 4,56 до 5,44 л/мин. На самом деле в этом интервале (с вероятностью 95%) находится *среднее* значение сердечного выброса. (В гл. 7 мы поговорим о доверительных интервалах более подробно.) Впрочем, стандартное отклонение можно рассчитать самому — для этого нужно умножить стандартную ошибку среднего на квадратный корень из объема выборки (численности группы). Правда, для этого нужно знать, что же именно приводит автор — стандартное отклонение или стандартную ошибку среднего.

ВЫВОДЫ

Когда совокупность подчиняется нормальному распределению, она исчерпывающе описывается *параметрами распределения* — средним и стандартным отклонением. Когда же распределение сильно отличается от нормального, более информативны медиана и процентили.

Так как наблюдать всю совокупность удается редко, мы *оцениваем* параметры распределения по выборке, случайным образом извлеченной из совокупности. Стандартная ошибка среднего служит мерой точности, с которой выборочное среднее является оценкой среднего по совокупности.

Эти величины полезны не только для описания совокупности или выборки. Их можно также использовать для проверки статистических гипотез, в частности о различиях между группами.

Этому и будет посвящена следующая глава.

ЗАДАЧИ

2.1. Найдите среднее, стандартное отклонение, медиану, 25-й и 75-й процентили для следующей выборки: 0; 0; 0; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 2; 2; 2; 2; 3; 3; 3; 3; 4; 4; 5; 5; 5; 5; 6; 7; 9; 10; 11. Можно ли считать, что выборка извлечена из совокупности с нормальным распределением? Обоснуйте свой ответ. (Приведенные числа — клинические оценки тяжести серповидноклеточной анемии. Подробный анализ этого исследования см. в задаче 8.9. Данные заимствованы из работы: R. Hebbel et al. Erythrocyte adherence to endothelium in sickle-cell anemia: a possible determinant of disease severity. *N. Engl. J. Med.*, 302, 992—995, 1980.)

2.2. Найдите среднее, стандартное отклонение, медиану, 25-й и 75-й процентили для следующих данных: 289; 203; 359; 243; 232; 210; 251; 246; 224; 239; 220; 211. Можно ли считать, что выборка извлечена из совокупности с нормальным распределением? Обоснуйте свой ответ. (Эти числа — продолжительность (в секундах) физической нагрузки до развития приступа стенокардии у 12 человек с ишемической болезнью сердца. Данные заимствованы из работы: W. Aronow. Effect of nonnicotine cigaretts and carbon monoxide on angina. *Circulation*, 61:262—265, 1979. Более подробно эта работа описана в задаче 9.5.)

2.3. Найдите среднее, стандартное отклонение, медиану, 25-й и 75-й процентили для следующих данных: 1,2; 1,4; 1,6; 1,7; 1,7; 1,8; 2,2; 2,3; 2,4; 6,4; 19,0; 23,6. Можно ли считать, что это — выборка из совокупности с нормальным распределением? Обоснуйте свой ответ. (Приведены результаты оценки проницаемости сосудов сетчатки из работы: G. A. Fishman et al. Blood-retinal barrier function in patients with cone or cone-rod dystrophy. *Arch. Ophthalmol.*, 104:545—548, 1986.)

2.4. Опишите распределение числа очков, выпадающих при бросании игральной кости. Найдите среднее число очков.

2.5. Бросьте одновременно две игральные кости, посмотрите, сколько очков выпало на каждой из них, и рассчитайте среднее. Повторите опыт 20 раз и постройте распределение средних, найденных после каждого броска. Что это за распределение? Вычислите его среднее и стандартное отклонение. Что они характеризуют?

2.6. Р. Флетчер и С. Флетчер (R. Fletcher, S. Fletcher. Clinical research in general medical journals: a 30-year perspective. *N. Engl. J. Med.*, 301:180—183, 1979) изучили библиографические характеристики 612 случайно выбранных статей, опубликованных в журналах Journal of American Medical Association, New England Journal of Medicine и Lancet с 1946 г. Одним из показателей было число авторов статьи. Было установлено следующее:

Год	Число обследованных статей	Среднее число авторов	Стандартное отклонение
1946	151	2,0	1,4
1956	149	2,3	1,6
1966	157	2,8	1,2
1976	155	4,9	7,3

Нарисуйте график среднего числа авторов по годам. Может ли распределение статей по числу авторов быть нормальным? Почему?

Глава 3

Сравнение нескольких групп: дисперсионный анализ

Статистические методы используют для описания данных и для оценки статистической значимости результатов опыта. В предыдущей главе мы занимались описанием данных. Мы ввели понятия среднего, стандартного отклонения, медианы и процентилей. Мы узнали, как оценивать эти показатели по выборке. Мы разобрались, как определить, насколько точна выборочная оценка среднего. Перейдем теперь к методам оценки статистической значимости различий (их называют *критериями значимости,* или просто критериями*). Методов этих существует множество, но все они построены по одному принципу. Сначала мы формулируем *нулевую гипотезу*, то есть предполагаем, что исследуемые факторы не оказывают никакого влияния на исследуемую величину и полученные различия случайны. Затем мы определяем, какова вероятность получить наблюдаемые (или более сильные) различия при условии справедливости нулевой гипотезы. Если

---

* Критерием называют и сам метод, и ту величину, которая получается в результате его применения.

эта вероятность мала*, то мы отвергаем нулевую гипотезу и заключаем, что результаты эксперимента *статистически значимы*. Это, разумеется, еще не означает, что мы доказали действие именно изучаемых факторов (это вопрос прежде всего планирования эксперимента), но, во всяком случае, маловероятно, что результат обусловлен случайностью.

Дисперсионный анализ был разработан в 20-х годах нашего столетия английским математиком и генетиком Рональдом Фишером. На дисперсионном анализе основан широкий класс критериев значимости, со многими из которых мы познакомимся в этой книге. Сейчас мы постараемся понять общий принцип этого метода.

СЛУЧАЙНЫЕ ВЫБОРКИ ИЗ НОРМАЛЬНО РАСПРЕДЕЛЕННОЙ СОВОКУПНОСТИ

Однажды в небольшом городке (200 жителей) ученые исследовали влияние диеты на сердечный выброс. Случайным образом отобрали 28 человек, каждый из которых согласился участвовать в исследовании. После этого они, опять-таки случайным образом, были разделены на 4 группы по 7 человек в каждой. Члены первой (контрольной) группы продолжали питаться как обычно, члены второй группы стали есть только макароны, третьей группы — мясо, четвертой — фрукты. Через месяц у всех участников эксперимента измерили сердечный выброс. Результаты представлены на рис. 3.2.

Анализ данных мы начинаем с формулировки нулевой гипотезы. В данном случае она заключается в том, что ни одна из диет не влияет на сердечный выброс. Откроем маленький секрет — дело обстоит именно так. На рис. 3.1 показано распределение сердечного выброса для *всех* жителей городка: каждый житель представлен кружком. Члены наших экспериментальных групп изображены заштрихованными кружками. Все четыре группы

---

* Максимальную приемлемую вероятность отвергнуть верную нулевую гипотезу называют *уровнем значимости* и обозначают α. Обычно принимают $\alpha = 0{,}05$.

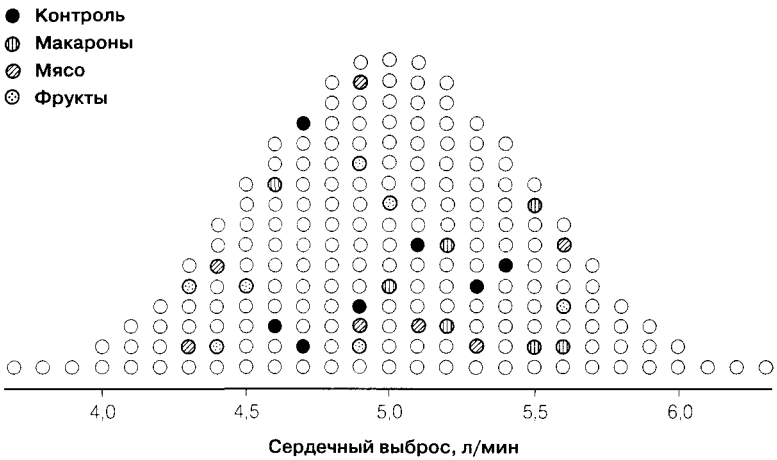


Рис. 3.1. Распределение жителей городка по величине сердечного выброса. Диета не влияет на сердечный выброс и экспериментальные группы представляют собой просто четыре случайные выборки из нормально распределенной совокупности.

представляют собой просто случайные выборки из нормально распределенной совокупности.

Однако как убедиться в этом, располагая только результатами эксперимента (рис 3.2)? Как видно из рисунка 3.2, группы все же различаются по средней величине сердечного выброса. Вопрос можно поставить так: какова вероятность получить такие различия, извлекая случайные выборки из нормально распределенной совокупности? Прежде чем ответить на этот вопрос, нам надо получить показатель, характеризующий величину различий.

Оставим на время наш эксперимент и зададимся вопросом, что заставляет нас, взглянув на несколько выборок, думать, что различия между ними не случайны.

Попробуем (исключительно в учебных целях) так изменить наши данные, чтобы читатель поверил во влияние диеты на сердечный выброс. Результат этой подтасовки представлен на рис. 3.3. Взаимное расположение точек в группах осталось прежним, но сами группы значительно раздвинуты по горизонтальной оси. Сравнив рис. 3.2 и 3.3, всякий скажет, что четыре вы-

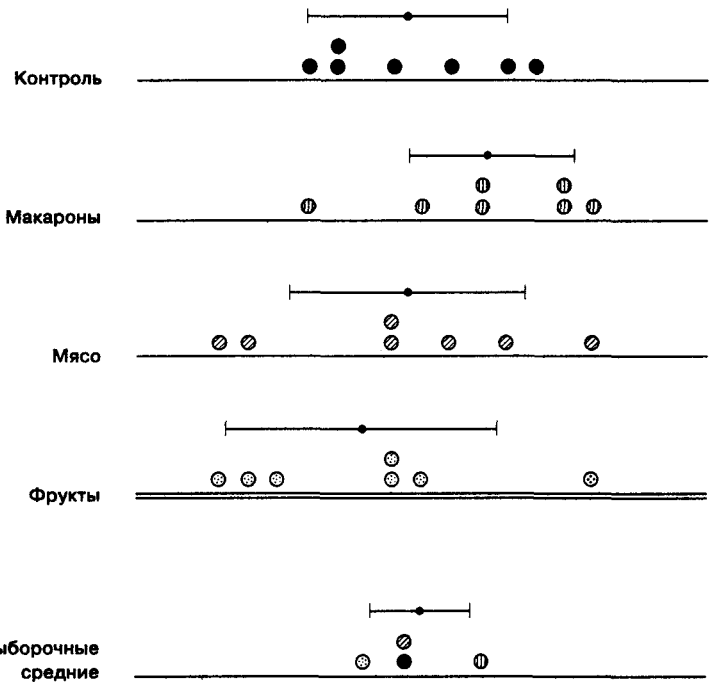


Рис. 3.2. Исследователь не может наблюдать совокупность; все, чем он располагает, — это его экспериментальные группы. На этом рисунке данные с рис. 3.1 представлены такими, какими их видит исследователь. Результаты в разных группах несколько различаются. Вызваны эти различия диетой или просто случайностью? Внизу рисунка показаны средние величины сердечного выброса в четырех группах (выборочные средние), а также среднее и стандартное отклонение этих четырех средних.

борки на рис. 3.2 «не различаются», а выборки на рис. 3.3 — «различаются». Почему? Сравним разброс значений внутри выборок с разбросом выборочных средних. Разброс выборочных средних на рис. 3.2 значительно меньше разброса значений в каждой из выборок. На рис. 3.3 картина обратная — разброс выборочных средних превышает разброс в каждой из выборок. То же самое можно сказать и о данных на рис. 3.4, хотя здесь три выборочных

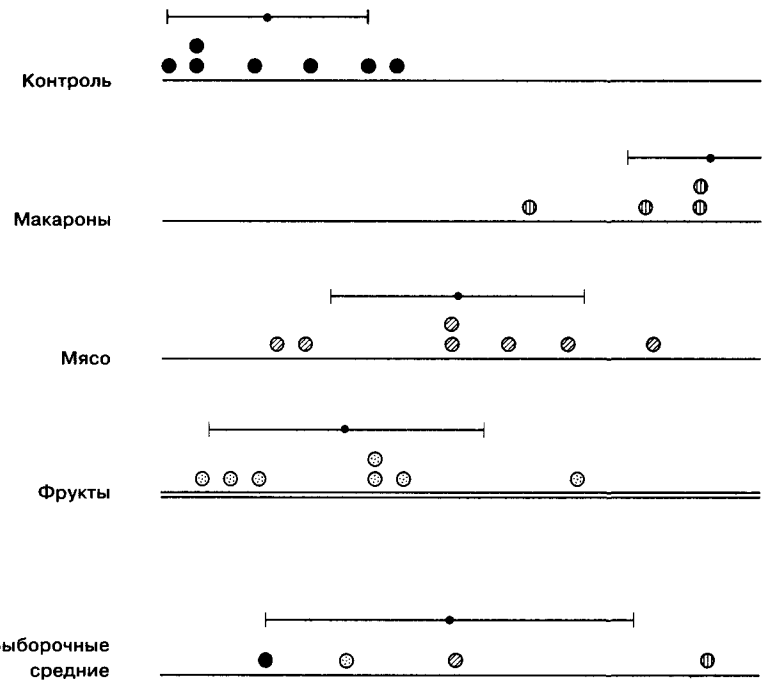


Рис. 3.3. Те же группы, что на предыдущих рисунках; теперь они раздвинуты по горизонтальной оси. Вряд ли такие различия можно отнести за счет случайности — влияние диеты налицо! Обратите внимание, что разброс выборочных средних превышает разброс внутри групп. На предыдущем рисунке картина была иной — разброс выборочных средних был меньше разброса внутри групп.

средних близки друг другу и заметно отличается от них только одна.

Итак, чтобы оценить величину различий, нужно каким-то образом сравнить разброс выборочных средних с разбросом значений внутри групп. Сейчас мы покажем, как это можно сделать с помощью дисперсии (как мы выяснили в предыдущей главе, этот показатель характеризует именно разброс), но прежде сделаем несколько замечаний.

Дисперсия правильно характеризует разброс только в том случае, если совокупность имеет нормальное распределение (вспомните

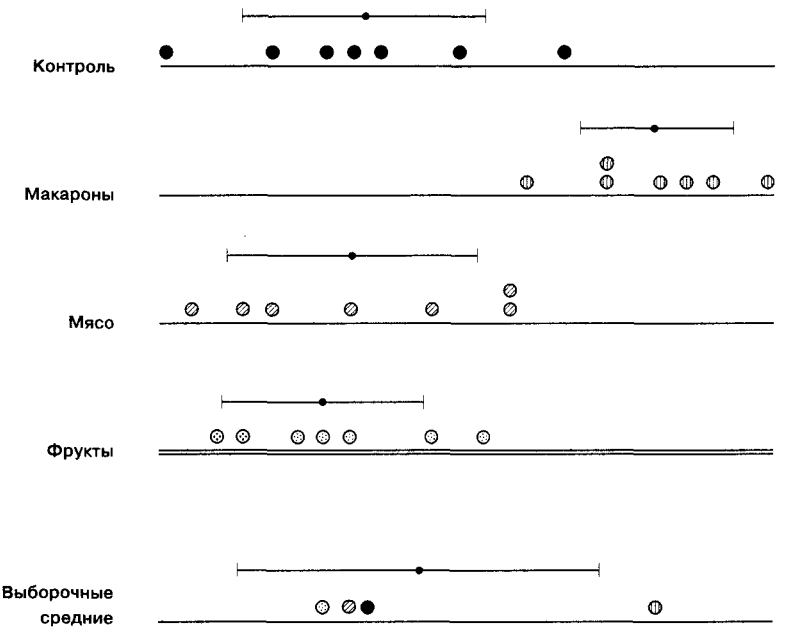


Рис. 3.4. Еще один возможный исход эксперимента с диетой. В трех группах средние примерно равны, и только в группе макаронной диеты сердечный выброс явно повысился. Такой результат, как и предыдущий, никто не отнесет на счет случайности. И снова разброс выборочных средних превышает разброс внутри групп.

обследование юпитериан, чуть было не приведшее к ошибочным заключениям). Поэтому и критерий, основанный на дисперсии, применим только для нормально распределенных совокупностей. Вообще, все критерии, основанные на оценке параметров распределения (они называются *параметрическими*), применимы только в случае, если данные подчиняются соответствующему распределению (чаще всего речь идет о нормальном распределении). Если распределение отличается от нормального, следует пользоваться так называемыми непараметрическими критериями. Эти критерии не основаны на оценке параметров распределения и вообще не требуют, чтобы данные подчинялись какому-то определенному типу распределения. Более подробно мы рассмотрим непараметри-

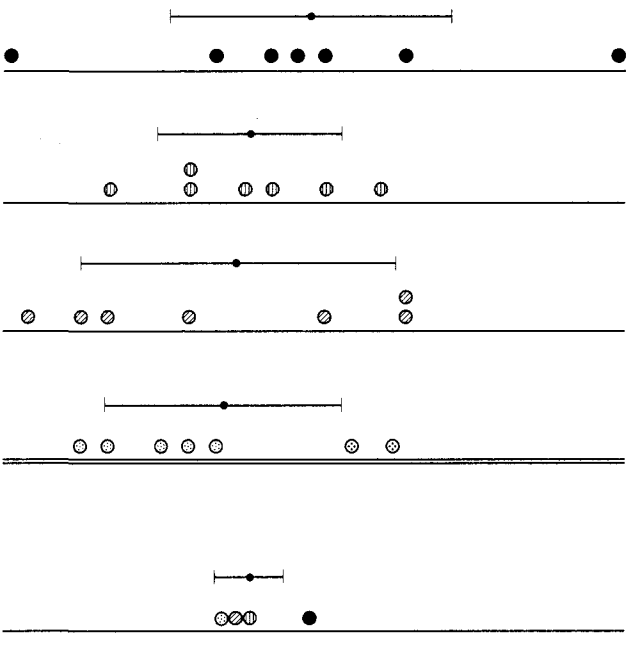


Рис. 3.5. Еще один набор из четырех случайных выборок по семь человек в каждой, извлеченных из совокупности в 200 человек (население городка, где изучали влияние диеты на сердечный выброс).

ческие критерии в гл. 5, 8 и 10. Непараметрические критерии дают более грубые оценки, чем параметрические. Параметрические методы более точны, но лишь в случае, если правильно определено распределение совокупности.

ДВЕ ОЦЕНКИ ДИСПЕРСИИ

Мы уже выяснили, что чем больше разброс средних и чем меньше разброс значений внутри групп, тем меньше вероятность того, что наши группы — это случайные выборки из одной совокупности. Осталось только оформить это суждение количественно.

Дисперсию совокупности можно оценить двумя способами. Во-первых, дисперсия, вычисленная для каждой группы, — это

оценка дисперсии совокупности. Поэтому дисперсию совокупности можно оценить на основании групповых дисперсий. Такая оценка не будет зависеть от различий групповых средних. Например, для данных на рис. 3.2 и 3.3 она будет одинаковой. Во-вторых, разброс выборочных средних тоже позволяет оценить дисперсию совокупности. Понятно, что такая оценка дисперсии зависит от различий выборочных средних.

Если экспериментальные группы — это четыре случайные выборки из одной и той же нормально распределенной совокупности (применительно к нашему эксперименту это значило бы, что диета не влияет на сердечный выброс), то обе оценки дисперсии совокупности дали бы примерно одинаковые результаты. Поэтому, если эти оценки оказываются близки, то мы не можем отвергнуть нулевую гипотезу. В противном случае мы отвергаем нулевую гипотезу, то есть заключаем: маловероятно, что мы получили бы такие различия между группами, если бы они были просто четырьмя случайными выборками из одной нормально распределенной совокупности.

Перейдем к вычислениям. Как оценить дисперсию совокупности по четырем выборочным дисперсиям? Если верна гипотеза о том, что диета не влияет на величину сердечного выброса, то любая из них дает одинаково хорошую оценку. Поэтому в качестве оценки дисперсии совокупности возьмем среднее выборочных дисперсий. Эта оценка называется внутригрупповой дисперсией; обозначим ее $s^2_{\text{вну}}$.

$$s^2_{\text{вну}} = \frac{1}{4}(s^2_{\text{кон}} + s^2_{\text{мак}} + s^2_{\text{мяс}} + s^2_{\text{фру}}),$$

где $s^2_{\text{кон}}$, $s^2_{\text{мак}}$, $s^2_{\text{мяс}}$, $s^2_{\text{фру}}$ — выборочные оценки дисперсии в группах, питавшихся как обычно (контроль), макаронами, мясом и фруктами. Дисперсия внутри каждой группы вычисляется относительно среднего для группы. Поэтому внутригрупповая дисперсия не зависит от того, насколько различаются эти средние.

Оценим теперь дисперсию совокупности по выборочным средним. Так как мы предположили, что все четыре выборки извлечены из одной совокупности, стандартное отклонение четырех выборочных средних служит оценкой ошибки среднего. На-

помним, что стандартная ошибка среднего $\sigma_{\bar{X}}$ связана со стандартным отклонением совокупности σ и объемом выборки n следующим соотношением:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

Тем самым, дисперсию совокупности σ^2 можно рассчитать следующим образом:

$$\sigma^2 = n\sigma_{\bar{X}}^2.$$

Воспользуемся этим, чтобы оценить дисперсию совокупности по разбросу значений выборочных средних. Эта оценка называется межгрупповой дисперсией; обозначим ее $s^2_{\text{меж}}$.

$$s^2_{\text{меж}} = ns_{\bar{X}}^2,$$

где $s_{\bar{X}}^2$ — оценка стандартного отклонения выборки из четырех средних.

Если верна нулевая гипотеза, то как внутригрупповая, так и межгрупповая дисперсии служат оценками одной и той же дисперсии и должны быть приближенно равны. Исходя из этого, вычислим критерий F:

$$F = \frac{\begin{array}{l}\text{Дисперсия совокупности,}\\ \text{оцененная по выборочным средним}\end{array}}{\begin{array}{l}\text{Дисперсия совокупности, оцененная}\\ \text{по выборочным дисперсиям}\end{array}},$$

или

$$F = \frac{s^2_{\text{меж}}}{s^2_{\text{вну}}}.$$

И числитель, и знаменатель этого отношения — это оценки одной и той же величины — дисперсии совокупности σ^2, поэтому значение F должно были близко к 1. Для четырех групп, представленных на рис. 3.2, значение F действительно близко к единице. Теперь наши исследователи влияния диеты на сердечный выброс могут сделать определенные выводы. Получен-

ные в эксперименте данные не противоречат нулевой гипотезе, следовательно, нет оснований считать, что диета влияет на сердечный выброс. Что касается данных, которые мы специально сконструировали, чтобы убедить читателя в таком «влиянии» (рис. 3.3), то для них $F = 68{,}0$. Для данных, изображенных на рис. 3.4, $F = 24{,}5$. Как видим, величина F хорошо согласуется с впечатлением, которое складывается при взгляде на рисунок.

Итак, если F значительно превышает 1, нулевую гипотезу следует отвергнуть. Если же значение F близко к 1, нулевую гипотезу следует принять. Осталось понять, начиная с какой именно величины F следует отвергать нулевую гипотезу.

КРИТИЧЕСКОЕ ЗНАЧЕНИЕ *F*

Если извлекать случайные выборки из нормально распределенной совокупности, значение F будет меняться от опыта к опыту. Например, на рис. 3.5 представлен еще один набор из четырех случайных выборок по семь человек в каждой, извлеченных из нашей совокупности в 200 человек. На этот раз $F = 0{,}5$. Положим, что нам удалось повторить эксперимент с жителями того же городка, скажем, 200 раз. Каждый раз мы заново набирали по четыре группы и каждый раз вычисляли F. На рис. 3.6А приведены результаты этого многократного эксперимента. Значения F округлены до одного знака после запятой и изображены кружками. Два черных кружка соответствуют данным с рис. 3.2 и 3.5. Как и следовало ожидать, большинство значений F близко к единице (попадая в интервал от 0 до 2); только в 10 из 200 опытов (то есть в 5% случаев) мы получили значение F, большее или равное 3. (На рис. 3.6Б эти 10 значений показаны черными кружками.) Значит, отвергая нулевую гипотезу при $F \geq 3$, мы будем ошибаться в 5% случаев. Если такой процент ошибок не чрезмерен, то будем считать «большими» те значения F, которые больше или равны 3. Значение критерия, начиная с которого мы отвергаем нулевую гипотезу, называется *критическим значением*.

Вероятность ошибочно отвергнуть верную нулевую гипотезу, то есть найти различия там, где их нет, обозначается P. Как правило, считают достаточным, чтобы эта вероятность не превышала

5%. (Максимальная приемлемая вероятность ошибочно отвергнуть нулевую гипотезу называется *уровнем значимости* и обозначается α.) Почему бы не повысить критическое значение F, тем самым уменьшая эту вероятность? Однако в этом случае возрастет риск ошибочно *принять* неверную нулевую гипотезу (то есть не найти различий там, где они есть). Подробнее мы поговорим об этом в гл. 6.

Итак, мы решили, приняв допустимой 5% вероятность ошибки, отвергать нулевую гипотезу при $F > 3$. Однако критическое значение F следовало бы выбрать на основе не 200, а всех 10^{42} экспериментов, которые можно провести на совокупности из 200 человек. Предположим, что нам удалось провести все эти эксперименты. По их результатам мы вычислили соответствующие значения F и нанесли их на график (рис. 3.6В). Здесь каждое значение F изображено «песчинкой». На долю темных песчинок в правой части горки приходится 5% всех значений. Картина, в общем, похожа на ту, что мы видели рис. 3.6Б. На практике совокупности гораздо больше, чем население нашего городка, а число возможных значений F несравненно больше 10^{42}. Если мысленно увеличить объем совокупности до бесконечности, то песчинки сольются и получится гладкая кривая, изображенная на рис. 3.6Г. Площади под кривой аналогичны долям от общего числа кружков или песчинок на рис. 3.6А, Б и В. Заштрихованная область на рис. 3.6Г составляет 5% всей площади под кривой. Эта область начинается от $F = 3{,}01$; это и есть критическое значение F.

В нашем примере число групп равнялось 4, в каждую группу входило 7 человек. Если бы число групп или число членов в каждой группе было другим, кривая пошла бы по-другому и критическое значение F тоже было бы другим. Вообще, критическое значение F однозначно определяется уровнем значимости (обычно 0,05 или 0,01) и еще двумя параметрами, которые называются внутригрупповым и межгрупповым числом степеней свободы и обозначаются греческой буквой ν («ню»). Оставим в стороне вопрос о происхождении этих названий и просто укажем, как их определять. Межгрупповое число степеней свободы — это число групп минус единица: $\nu_{меж} = m - 1$. Внутригрупповое число степеней свободы — это произведение числа групп на численность

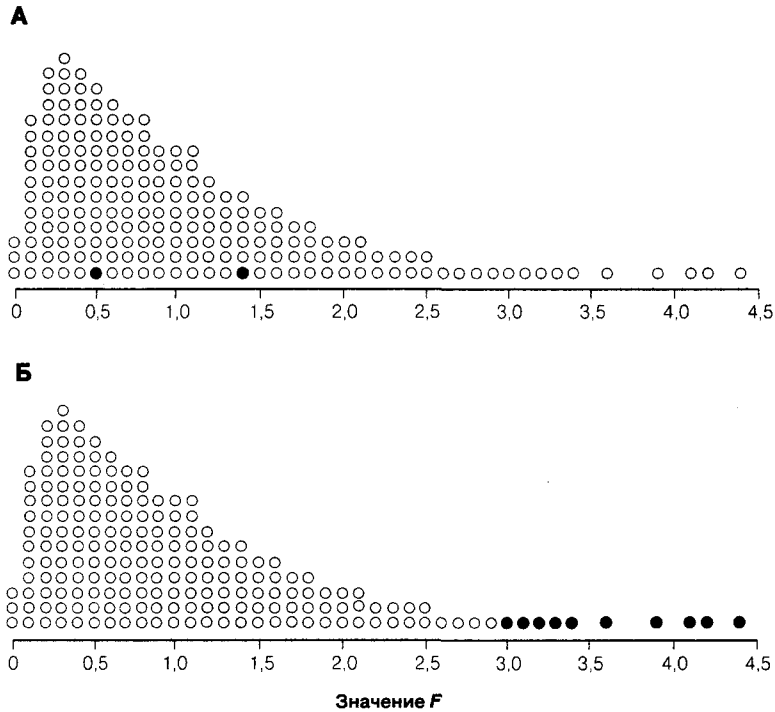


Рис. 3.6. А. Четыре случайные выборки по 7 человек в каждой извлекли из той же совокупности (население городка) 200 раз. Каждый раз рассчитывали значение *F* и наносили его на график. Результаты для выборок с рис. 3.2 и 3.5 помечены черным. **Б.** Десять наибольших значений помечены черным. Область черных кружков начинается со значения *F*, равного 3,0.

каждой из групп минус единица: $\nu_{\text{вну}} = m(n-1)$. В примере с исследованием диеты межгрупповое число степеней свободы равно $4-1=3$, а внутригрупповое $4(7-1)=24$. Вычислить критическое значение *F* довольно сложно, поэтому пользуются таблицами критических значений *F* для разных α, $\nu_{\text{меж}}$ и $\nu_{\text{вну}}$ (табл. 3.1).

Математическая модель, на которой основано вычисление критических значений *F*, предполагает следующее.

- Каждая выборка независима от остальных выборок.
- Каждая выборка случайным образом извлечена из исследуемой совокупности.

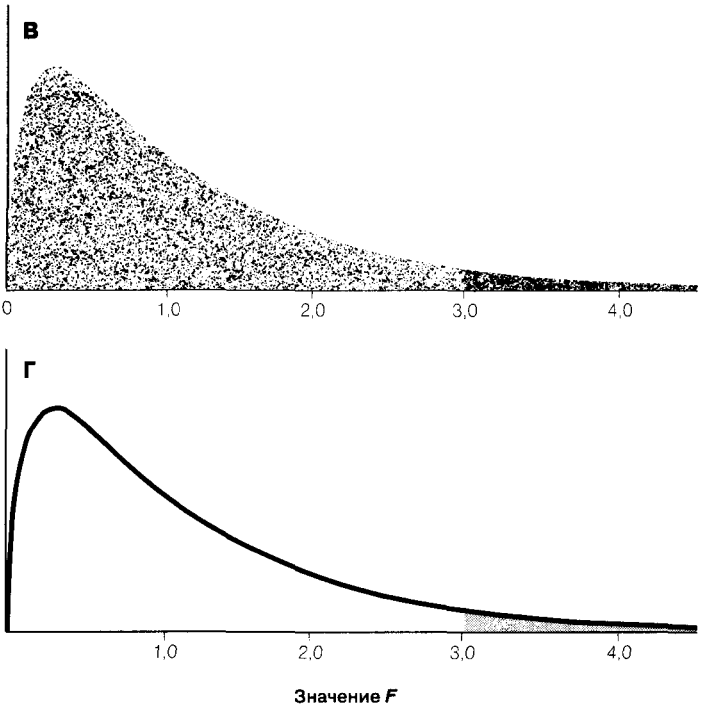


Рис. 3.6 (продолжение). В. Из той же совокупности извлекли все возможные наборы из 4 выборок по 7 человек в каждой и построили распределение *F*. Отдельные значения слились, превратившись в песчинки. 5% песчинок с самыми большими значениями *F* помечены черным. **Г.** Такое распределение *F* получится, если извлекать выборки из бесконечной совокупности. Пяти процентам самых *высоких значений F соответствует заштрихованная область* (*ее площадь* составляет 5% от общей площади под кривой). «Большие» значения *F* начинаются там, где начинается эта область, то есть с $F = 3{,}01$.

- Совокупность нормально распределена.
- Дисперсии всех выборок равны.

При существенном нарушении хотя бы одного из этих условий нельзя пользоваться ни таблицей 3.1, ни вообще дисперсионным анализом.

В рассмотренном нами эксперименте исследовалась зависимость только от одного фактора — диеты. Дисперсионный ана-

Таблица 3.1. Критические значения F для $\alpha = 0{,}05$ (обычный шрифт) и $\alpha = 0{,}01$ (жирный шрифт)

$\nu_{вну}$	$\nu_{меж}$ 1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞
1	161	200	216	225	230	234	237	239	241	242	243	244	245	246	248	249	250	251	252	253	253	254	254	254
	4052	**4999**	**5403**	**5625**	**5764**	**5859**	**5928**	**5981**	**6022**	**6056**	**6082**	**6106**	**6142**	**6169**	**6208**	**6234**	**6261**	**6286**	**6302**	**6323**	**6334**	**6352**	**6361**	**6366**
2	18,51	19,00	19,16	19,25	19,30	19,33	19,36	19,37	19,38	19,39	19,40	19,41	19,42	19,43	19,44	19,45	19,46	19,47	19,47	19,48	19,49	19,49	19,50	19,50
	98,49	**99,00**	**99,17**	**99,25**	**99,30**	**99,33**	**99,36**	**99,37**	**99,39**	**99,40**	**99,41**	**99,42**	**99,43**	**99,44**	**99,45**	**99,46**	**99,47**	**99,48**	**99,48**	**99,49**	**99,49**	**99,49**	**99,50**	**99,50**
3	10,13	9,55	9,28	9,12	9,01	8,94	8,88	8,84	8,81	8,78	8,76	8,74	8,71	8,69	8,66	8,64	8,62	8,60	8,58	8,57	8,56	8,54	8,54	8,53
	34,12	**30,82**	**29,46**	**28,71**	**28,24**	**27,91**	**27,67**	**27,49**	**27,34**	**27,23**	**27,13**	**27,05**	**26,92**	**26,83**	**26,69**	**26,60**	**26,50**	**26,41**	**26,35**	**26,27**	**26,23**	**26,18**	**26,14**	**26,12**
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,93	5,91	5,87	5,84	5,80	5,77	5,74	5,71	5,70	5,68	5,66	5,65	5,64	5,63
	21,20	**18,00**	**16,69**	**15,98**	**15,52**	**15,21**	**14,98**	**14,80**	**14,66**	**14,54**	**14,45**	**14,37**	**14,24**	**14,15**	**14,02**	**13,93**	**13,83**	**13,74**	**13,69**	**13,61**	**13,57**	**13,52**	**13,48**	**13,46**
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,78	4,74	4,70	4,68	4,64	4,60	4,56	4,53	4,50	4,46	4,44	4,42	4,40	4,38	4,37	4,36
	16,26	**13,27**	**12,06**	**11,39**	**10,97**	**10,67**	**10,45**	**10,29**	**10,15**	**10,05**	**9,96**	**9,89**	**9,77**	**9,68**	**9,55**	**9,47**	**9,38**	**9,29**	**9,24**	**9,17**	**9,13**	**9,07**	**9,04**	**9,02**
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00	3,96	3,92	3,87	3,84	3,81	3,77	3,75	3,72	3,71	3,69	3,68	3,67
	13,74	**10,92**	**9,78**	**9,15**	**8,75**	**8,47**	**8,26**	**8,10**	**7,98**	**7,87**	**7,79**	**7,72**	**7,60**	**7,52**	**7,39**	**7,31**	**7,23**	**7,14**	**7,09**	**7,02**	**6,99**	**6,94**	**6,90**	**6,88**
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,63	3,60	3,57	3,52	3,49	3,44	3,41	3,38	3,34	3,32	3,29	3,28	3,25	3,24	3,23
	12,25	**9,55**	**8,45**	**7,85**	**7,46**	**7,19**	**7,00**	**6,84**	**6,71**	**6,62**	**6,54**	**6,47**	**6,35**	**6,27**	**6,15**	**6,07**	**5,98**	**5,90**	**5,85**	**5,78**	**5,75**	**5,70**	**5,67**	**5,65**
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,34	3,31	3,28	3,23	3,20	3,15	3,12	3,08	3,05	3,03	3,00	2,98	2,96	2,94	2,93
	11,26	**8,65**	**7,59**	**7,01**	**6,63**	**6,37**	**6,19**	**6,03**	**5,91**	**5,82**	**5,74**	**5,67**	**5,56**	**5,48**	**5,36**	**5,28**	**5,20**	**5,11**	**5,06**	**5,00**	**4,96**	**4,91**	**4,88**	**4,86**
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,13	3,10	3,07	3,02	2,98	2,93	2,90	2,86	2,82	2,80	2,77	2,76	2,73	2,72	2,71
	10,56	**8,02**	**6,99**	**6,42**	**6,06**	**5,80**	**5,62**	**5,47**	**5,35**	**5,26**	**5,18**	**5,11**	**5,00**	**4,92**	**4,80**	**4,73**	**4,64**	**4,56**	**4,51**	**4,45**	**4,41**	**4,36**	**4,33**	**4,31**
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,97	2,94	2,91	2,86	2,82	2,77	2,74	2,70	2,67	2,64	2,61	2,59	2,56	2,55	2,54
	10,04	**7,56**	**6,55**	**5,99**	**5,64**	**5,39**	**5,21**	**5,06**	**4,95**	**4,85**	**4,78**	**4,71**	**4,60**	**4,52**	**4,41**	**4,33**	**4,25**	**4,17**	**4,12**	**4,05**	**4,01**	**3,96**	**3,93**	**3,91**
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,86	2,82	2,79	2,74	2,70	2,65	2,61	2,57	2,53	2,50	2,47	2,45	2,42	2,41	2,40
	9,65	**7,20**	**6,22**	**5,67**	**5,32**	**5,07**	**4,88**	**4,74**	**4,63**	**4,54**	**4,46**	**4,40**	**4,29**	**4,21**	**4,10**	**4,02**	**3,94**	**3,86**	**3,80**	**3,74**	**3,70**	**3,66**	**3,62**	**3,60**
12	4,75	3,88	3,49	3,26	3,11	3,00	2,92	2,85	2,80	2,76	2,72	2,69	2,64	2,60	2,54	2,50	2,46	2,42	2,40	2,36	2,35	2,32	2,31	2,30
	9,33	**6,93**	**5,95**	**5,41**	**5,06**	**4,82**	**4,65**	**4,50**	**4,39**	**4,30**	**4,22**	**4,16**	**4,05**	**3,98**	**3,86**	**3,78**	**3,70**	**3,61**	**3,56**	**3,49**	**3,46**	**3,41**	**3,38**	**3,36**
13	4,67	3,80	3,41	3,18	3,02	2,92	2,84	2,77	2,72	2,67	2,63	2,60	2,55	2,51	2,46	2,42	2,38	2,34	2,32	2,28	2,26	2,24	2,22	2,21
	9,07	**6,70**	**5,74**	**5,20**	**4,86**	**4,62**	**4,44**	**4,30**	**4,19**	**4,10**	**4,02**	**3,96**	**3,85**	**3,78**	**3,67**	**3,59**	**3,51**	**3,42**	**3,37**	**3,30**	**3,27**	**3,21**	**3,18**	**3,16**
14	4,60	3,74	3,34	3,11	2,96	2,85	2,77	2,70	2,65	2,60	2,56	2,53	2,48	2,44	2,39	2,35	2,31	2,27	2,24	2,21	2,19	2,16	2,14	2,13
	8,86	**6,51**	**5,56**	**5,03**	**4,69**	**4,46**	**4,28**	**4,14**	**4,03**	**3,94**	**3,86**	**3,80**	**3,70**	**3,62**	**3,51**	**3,43**	**3,34**	**3,26**	**3,21**	**3,14**	**3,11**	**3,06**	**3,02**	**3,00**
15	4,54	3,68	3,29	3,06	2,90	2,79	2,70	2,64	2,59	2,55	2,51	2,48	2,43	2,39	2,33	2,29	2,25	2,21	2,18	2,15	2,12	2,10	2,08	2,07
	8,68	**6,36**	**5,42**	**4,89**	**4,56**	**4,32**	**4,14**	**4,00**	**3,89**	**3,80**	**3,73**	**3,67**	**3,56**	**3,48**	**3,36**	**3,29**	**3,20**	**3,12**	**3,07**	**3,00**	**2,97**	**2,92**	**2,89**	**2,87**
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,45	2,42	2,37	2,33	2,28	2,24	2,20	2,16	2,13	2,09	2,07	2,04	2,02	2,01
	8,53	**6,23**	**5,29**	**4,77**	**4,44**	**4,20**	**4,03**	**3,89**	**3,78**	**3,69**	**3,61**	**3,55**	**3,45**	**3,37**	**3,25**	**3,18**	**3,10**	**3,01**	**2,96**	**2,98**	**2,86**	**2,80**	**2,77**	**2,75**
17	4,45	3,59	3,20	2,96	2,81	2,70	2,62	2,55	2,50	2,45	2,41	2,38	2,33	2,29	2,23	2,19	2,15	2,11	2,08	2,04	2,02	1,99	1,97	1,96
	8,40	**6,11**	**5,18**	**4,67**	**4,34**	**4,10**	**3,93**	**3,79**	**3,68**	**3,59**	**3,52**	**3,45**	**3,35**	**3,27**	**3,16**	**3,08**	**3,00**	**2,92**	**2,86**	**2,79**	**2,76**	**2,70**	**2,67**	**2,65**
18	4,41	3,55	3,16	2,93	2,77	3,66	2,58	2,51	2,46	2,41	2,37	2,34	2,29	2,25	2,19	2,15	2,11	2,07	2,04	2,00	1,98	1,95	1,93	1,92
	8,28	**6,01**	**5,09**	**4,58**	**4,25**	**4,01**	**3,85**	**3,71**	**3,60**	**3,51**	**3,44**	**3,37**	**3,27**	**3,19**	**3,07**	**3,00**	**2,91**	**2,83**	**2,78**	**2,71**	**2,68**	**2,62**	**2,59**	**2,57**
19	4,38	3,52	3,13	2,90	2,74	2,63	2,55	2,48	2,43	2,38	2,34	2,31	2,26	2,21	2,15	2,11	2,07	2,02	2,00	1,96	1,94	1,91	1,90	1,88
	8,18	**5,93**	**5,01**	**4,50**	**4,17**	**3,94**	**3,77**	**3,63**	**3,52**	**3,43**	**3,36**	**3,30**	**3,19**	**3,12**	**3,00**	**2,92**	**2,84**	**2,76**	**2,70**	**2,63**	**2,60**	**2,54**	**2,51**	**2,49**
20	4,35	3,49	3,10	2,87	2,71	2,60	2,52	2,45	2,40	2,35	2,31	2,28	2,23	2,18	2,12	2,08	2,04	1,99	1,96	1,92	1,90	1,87	1,85	1,84
	8,10	**5,85**	**4,94**	**4,43**	**4,10**	**3,87**	**3,71**	**3,56**	**3,45**	**3,37**	**3,30**	**3,23**	**3,13**	**3,05**	**2,94**	**2,86**	**2,77**	**2,69**	**2,63**	**2,56**	**2,53**	**2,47**	**2,44**	**2,42**
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,28	2,25	2,20	2,15	2,09	2,05	2,00	1,96	1,93	1,89	1,87	1,84	1,82	1,81
	8,02	**5,78**	**4,87**	**4,37**	**4,04**	**3,81**	**3,65**	**3,51**	**3,40**	**3,31**	**3,24**	**3,17**	**3,07**	**2,99**	**2,88**	**2,80**	**2,72**	**2,63**	**2,58**	**2,51**	**2,47**	**2,42**	**2,38**	**2,36**
22	4,30	3,44	3,05	2,82	2,66	2,55	2,47	2,40	2,35	2,30	2,26	2,23	2,18	2,13	2,07	2,03	1,98	1,93	1,91	1,87	1,84	1,81	1,80	1,78
	7,94	**5,72**	**4,82**	**4,31**	**3,99**	**3,76**	**3,59**	**3,45**	**3,35**	**3,26**	**3,18**	**3,12**	**3,02**	**2,94**	**2,83**	**2,75**	**2,67**	**2,58**	**2,53**	**2,46**	**2,42**	**2,37**	**2,33**	**2,31**
23	4,28	3,42	3,03	2,80	2,64	2,53	2,45	2,38	2,32	2,28	2,24	2,20	2,14	2,10	2,04	2,00	1,96	1,91	1,88	1,84	1,82	1,79	1,77	1,76
	7,88	**5,66**	**4,76**	**4,26**	**3,94**	**3,71**	**3,54**	**3,41**	**3,30**	**3,21**	**3,14**	**3,07**	**2,97**	**2,89**	**2,78**	**2,70**	**2,62**	**2,53**	**2,48**	**2,41**	**2,37**	**2,32**	**2,28**	**2,26**
24	4,26	3,40	3,01	2,78	2,62	2,51	2,43	2,36	2,30	2,26	2,22	2,18	2,13	2,09	2,02	1,98	1,94	1,89	1,86	1,82	1,80	1,76	1,74	1,73
	7,82	**5,61**	**4,72**	**4,22**	**3,90**	**3,67**	**3,50**	**3,36**	**3,25**	**3,17**	**3,09**	**3,03**	**2,93**	**2,85**	**2,74**	**2,66**	**2,58**	**2,49**	**2,44**	**2,36**	**2,33**	**2,27**	**2,23**	**2,21**
25	4,24	3,38	2,99	2,76	2,60	2,49	2,41	2,34	2,28	2,24	2,20	2,16	2,11	2,06	2,00	1,96	1,92	1,87	1,84	1,80	1,77	1,74	1,72	1,71
	7,77	**5,57**	**4,68**	**4,18**	**3,86**	**3,63**	**3,46**	**3,32**	**3,21**	**3,13**	**3,05**	**2,99**	**2,89**	**2,81**	**2,70**	**2,62**	**2,54**	**2,45**	**2,40**	**2,32**	**2,29**	**2,23**	**2,19**	**2,17**
26	4,22	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,18	2,15	2,10	2,05	1,99	1,95	1,90	1,85	1,82	1,78	1,76	1,72	1,70	1,69
	7,72	**5,53**	**4,64**	**4,14**	**3,82**	**3,59**	**3,42**	**3,29**	**3,17**	**3,09**	**3,02**	**2,96**	**2,86**	**2,77**	**2,66**	**2,58**	**2,50**	**2,41**	**2,36**	**2,28**	**2,25**	**2,19**	**2,15**	**2,13**
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,30	2,25	2,20	2,16	2,13	2,08	2,03	1,97	1,93	1,88	1,84	1,80	1,76	1,74	1,71	1,68	1,67
	7,68	**5,49**	**4,60**	**4,11**	**3,79**	**3,56**	**3,39**	**3,26**	**3,14**	**3,06**	**2,98**	**2,93**	**2,83**	**2,74**	**2,63**	**2,55**	**2,47**	**2,38**	**2,33**	**2,25**	**2,21**	**2,16**	**2,12**	**2,10**
28	4,20	3,34	2,95	2,71	2,56	2,44	2,36	2,29	2,24	2,19	2,15	2,12	2,06	2,02	1,96	1,91	1,87	1,81	1,78	1,75	1,72	1,69	1,67	1,65
	7,64	**5,45**	**4,57**	**4,07**	**3,76**	**3,53**	**3,36**	**3,23**	**3,11**	**3,03**	**2,95**	**2,90**	**2,80**	**2,71**	**2,60**	**2,52**	**2,44**	**2,35**	**2,30**	**2,22**	**2,18**	**2,13**	**2,09**	**2,06**
29	4,18	3,33	2,93	2,70	2,54	2,43	2,35	2,28	2,22	2,18	2,14	2,10	2,05	2,00	1,94	1,90	1,85	1,80	1,77	1,73	1,71	1,68	1,65	1,64
	7,60	**5,42**	**4,54**	**4,04**	**3,73**	**3,50**	**3,33**	**3,20**	**3,08**	**3,00**	**2,92**	**2,87**	**2,77**	**2,68**	**2,57**	**2,49**	**2,41**	**2,32**	**2,27**	**2,19**	**2,15**	**2,10**	**2,06**	**2,03**
30	4,17	3,32	2,92	2,69	2,53	2,42	2,34	2,27	2,21	2,16	2,12	2,09	2,04	1,99	1,93	1,89	1,84	1,79	1,76	1,72	1,69	1,66	1,64	1,62
	7,56	**5,39**	**4,51**	**4,02**	**3,70**	**3,47**	**3,30**	**3,17**	**3,06**	**2,98**	**2,90**	**2,84**	**2,74**	**2,66**	**2,55**	**2,47**	**2,38**	**2,29**	**2,24**	**2,16**	**2,13**	**2,07**	**2,03**	**2,01**
32	4,15	3,30	2,90	2,67	2,51	2,40	2,32	2,25	2,19	2,14	2,10	2,07	2,02	1,97	1,91	1,86	1,82	1,76	1,74	1,69	1,67	1,64	1,61	1,59
	7,50	**5,34**	**4,46**	**3,97**	**3,66**	**3,42**	**3,25**	**3,12**	**3,01**	**2,94**	**2,86**	**2,80**	**2,70**	**2,62**	**2,51**	**2,42**	**2,34**	**2,25**	**2,20**	**2,12**	**2,08**	**2,02**	**1,98**	**1,96**
34	4,13	3,28	2,88	2,65	2,49	2,38	2,30	2,23	2,17	2,12	2,08	2,05	2,00	1,95	1,89	1,84	1,80	1,74	1,71	1,67	1,64	1,61	1,59	1,57
	7,44	**5,29**	**4,42**	**3,93**	**3,61**	**3,38**	**3,21**	**3,08**	**2,97**	**2,89**	**2,82**	**2,76**	**2,66**	**2,58**	**2,47**	**2,38**	**2,30**	**2,21**	**2,15**	**2,08**	**2,04**	**1,98**	**1,94**	**1,91**

$\nu_{вну}$	$\nu_{меж}$ 1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞
36	4,11	3,26	2,86	2,63	2,48	2,36	2,28	2,21	2,15	2,10	2,06	2,03	1,98	1,93	1,87	1,82	1,78	1,72	1,69	1,65	1,62	1,59	1,56	1,55
	7,39	**5,25**	**4,38**	**3,89**	**3,58**	**3,35**	**3,18**	**3,04**	**2,94**	**2,86**	**2,78**	**2,72**	**2,62**	**2,54**	**2,43**	**2,35**	**2,26**	**2,17**	**2,12**	**2,04**	**2,00**	**1,94**	**1,90**	**1,87**
38	4,10	3,25	2,85	2,62	2,46	2,35	2,26	2,19	2,14	2,09	2,05	2,02	1,96	1,92	1,85	1,80	1,76	1,71	1,67	1,63	1,60	1,57	1,54	1,53
	7,35	**5,21**	**4,34**	**3,86**	**3,54**	**3,32**	**3,15**	**3,02**	**2,91**	**2,82**	**2,75**	**2,69**	**2,59**	**2,51**	**2,40**	**2,32**	**2,22**	**2,14**	**2,08**	**2,00**	**1,97**	**1,90**	**1,86**	**1,84**
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,07	2,04	2,00	1,95	1,90	1,84	1,79	1,74	1,69	1,66	1,61	1,59	1,55	1,53	1,51
	7,31	**5,18**	**4,31**	**3,83**	**3,51**	**3,29**	**3,12**	**2,99**	**2,88**	**2,80**	**2,73**	**2,66**	**2,56**	**2,49**	**2,37**	**2,29**	**2,20**	**2,11**	**2,05**	**1,97**	**1,94**	**1,88**	**1,84**	**1,81**
42	4,07	3,22	2,83	2,59	2,44	2,32	2,24	2,17	2,11	2,06	2,02	1,99	1,94	1,89	1,82	1,78	1,73	1,68	1,64	1,60	1,57	1,54	1,51	1,49
	7,27	**5,15**	**4,29**	**3,80**	**3,49**	**3,26**	**3,10**	**2,96**	**2,86**	**2,77**	**2,70**	**2,64**	**2,54**	**2,46**	**2,35**	**2,26**	**2,17**	**2,08**	**2,02**	**1,94**	**1,91**	**1,85**	**1,80**	**1,78**
44	4,06	3,21	2,82	2,58	2,43	2,31	2,23	2,16	2,10	2,05	2,01	1,98	1,92	1,88	1,81	1,76	1,72	1,66	1,63	1,58	1,56	1,52	1,50	1,48
	7,24	**5,12**	**4,26**	**3,78**	**3,46**	**3,24**	**3,07**	**2,94**	**2,84**	**2,75**	**2,68**	**2,62**	**2,52**	**2,44**	**2,32**	**2,24**	**2,15**	**2,06**	**2,00**	**1,92**	**1,88**	**1,82**	**1,78**	**1,75**
46	4,05	3,20	2,81	2,57	2,42	2,30	2,22	2,14	2,09	2,04	2,00	1,97	1,91	1,87	1,80	1,75	1,71	1,65	1,62	1,57	1,54	1,51	1,48	1,46
	7,21	**5,10**	**4,24**	**3,76**	**3,44**	**3,22**	**3,05**	**2,92**	**2,82**	**2,73**	**2,66**	**2,60**	**2,50**	**2,42**	**2,30**	**2,22**	**2,13**	**2,04**	**1,98**	**1,90**	**1,86**	**1,80**	**1,76**	**1,72**
48	4,04	3,19	2,80	2,56	2,41	2,30	2,21	2,14	2,08	2,03	1,99	1,96	1,90	1,86	1,79	1,74	1,70	1,64	1,61	1,56	1,53	1,50	1,47	1,45
	7,19	**5,08**	**4,22**	**3,74**	**3,42**	**3,20**	**3,04**	**2,90**	**2,80**	**2,71**	**2,64**	**2,58**	**2,48**	**2,40**	**2,28**	**2,20**	**2,11**	**2,02**	**1,96**	**1,88**	**1,84**	**1,78**	**1,73**	**1,70**
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,02	1,98	1,95	1,90	1,85	1,78	1,74	1,69	1,63	1,60	1,55	1,52	1,48	1,46	1,44
	7,17	**5,06**	**4,20**	**3,72**	**3,41**	**3,18**	**3,02**	**2,88**	**2,78**	**2,70**	**2,62**	**2,56**	**2,46**	**2,39**	**2,26**	**2,18**	**2,10**	**2,00**	**1,94**	**1,86**	**1,82**	**1,76**	**1,71**	**1,68**
60	4,00	3,15	2,76	2,52	2,37	2,25	2,17	2,10	2,04	1,99	1,95	1,92	1,86	1,81	1,75	1,70	1,65	1,59	1,56	1,50	1,48	1,44	1,41	1,39
	7,08	**4,98**	**4,13**	**3,65**	**3,34**	**3,12**	**2,95**	**2,82**	**2,72**	**2,63**	**2,56**	**2,50**	**2,40**	**2,32**	**2,20**	**2,12**	**2,03**	**1,93**	**1,87**	**1,79**	**1,74**	**1,68**	**1,63**	**1,60**
70	3,98	3,13	2,74	2,50	2,35	2,23	2,14	2,07	2,01	1,97	1,93	1,89	1,84	1,79	1,72	1,67	1,62	1,56	1,53	1,47	1,45	1,40	1,37	1,35
	7,01	**4,92**	**4,08**	**3,60**	**3,29**	**3,07**	**2,91**	**2,77**	**2,67**	**2,59**	**2,51**	**2,45**	**2,35**	**2,28**	**2,15**	**2,07**	**1,98**	**1,88**	**1,82**	**1,74**	**1,69**	**1,62**	**1,56**	**1,53**
80	3,96	3,11	2,72	2,48	2,33	2,21	2,12	2,05	1,99	1,95	1,91	1,88	1,82	1,77	1,70	1,65	1,60	1,54	1,51	1,45	1,42	1,38	1,35	1,32
	6,96	**4,88**	**4,04**	**3,56**	**3,25**	**3,04**	**2,87**	**2,74**	**2,64**	**2,55**	**2,48**	**2,41**	**2,32**	**2,24**	**2,11**	**2,03**	**1,94**	**1,84**	**1,78**	**1,70**	**1,65**	**1,57**	**1,52**	**1,49**
100	3,94	3,09	2,70	2,46	2,30	2,19	2,10	2,03	1,97	1,92	1,88	1,85	1,79	1,75	1,68	1,63	1,57	1,51	1,48	1,42	1,39	1,34	1,30	1,28
	6,90	**4,82**	**3,98**	**3,51**	**3,20**	**2,99**	**2,82**	**2,69**	**2,59**	**2,51**	**2,43**	**2,36**	**2,26**	**2,19**	**2,06**	**1,98**	**1,89**	**1,79**	**1,73**	**1,64**	**1,59**	**1,51**	**1,46**	**1,43**
120	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91	1,87	1,84	1,78	1,73	1,66	1,61	1,56	1,50	1,46	1,39	1,37	1,32	1,28	1,25
	6,85	**4,79**	**3,95**	**3,48**	**3,17**	**2,96**	**2,79**	**2,66**	**2,56**	**2,47**	**2,40**	**2,34**	**2,23**	**2,15**	**2,03**	**1,95**	**1,86**	**1,76**	**1,70**	**1,61**	**1,56**	**1,48**	**1,42**	**1,38**
∞	3,84	2,99	2,60	2,37	2,21	2,09	2,01	1,94	1,88	1,83	1,79	1,75	1,69	1,64	1,57	1,52	1,46	1,40	1,35	1,28	1,24	1,17	1,11	1,00
	6,63	**4,60**	**3,78**	**3,32**	**3,02**	**2,80**	**2,64**	**2,51**	**2,41**	**2,32**	**2,24**	**2,18**	**2,07**	**1,99**	**1,87**	**1,79**	**1,69**	**1,59**	**1,52**	**1,41**	**1,36**	**1,25**	**1,15**	**1,00**

G. W. Snedecor, W. G. Cochran. Statistical methods. Iowa State University Press, Ames, 1978.

лиз, в котором проверяется влияние одного фактора, называется *однофакторным*. При изучении влияния более чем одного фактора используют *многофакторный дисперсионный анализ* (в этой книге не рассматривается).

ТРИ ПРИМЕРА

Сейчас мы уже можем оценивать статистическую значимость реальных данных. Покажем это на трех примерах, заимствованных из медицинской литературы. Оговорюсь, что при изложении этих примеров мне пришлось несколько отклониться от первоисточников. Тому есть две причины. Во-первых, в медицинских публикациях обычно приводят не сами данные, а средние величины и прочие обобщенные показатели. Нередко дело обстоит и того хуже. Минуя все промежуточные этапы, авторы сообщают, что «$P < 0{,}05$». Поэтому «данные из литературных источников» по большей части являются плодом моих собственных догадок, какими могли бы быть исходные данные. Во-вторых, дисперсионный анализ в том виде, как мы его изложили, требует, чтобы численность всех групп была одинаковой. Поэтому мне пришлось видоизменять приводимые в работах данные так, чтобы соблюсти это требование. Впоследствии мы обобщим наши статистические методы, и их можно будет применять и при неравной численности групп.

Позволяет ли правильное лечение сократить срок госпитализации?

Стоимость пребывания в больнице — самая весомая статья расходов на здравоохранение. Сокращение госпитализации без снижения качества лечения дало бы значительный экономический эффект. Способствует ли соблюдение официальных схем лечения сокращению госпитализации? Чтобы ответить на этот вопрос, Кнапп и соавт.* изучили истории болезни лиц, поступив-

* D. E. Knapp, D. A. Knapp, M. K. Speedie, D. M. Yaeger, C. L. Baker. Relationship of inappropriate drug prescribing to increased length of hospital stay. *Am. J. Hosp. Pharm.*, 36:1334—1337, 1979.

ших в бесплатную больницу с острым пиелонефритом. Острый пиелонефрит был выбран как заболевание, имеющее четко очерченную клиническую картину и столь же четко регламентированные методы лечения.

Эта работа — пример *обсервационного* исследования. В отличие от *экспериментального* исследования, где исследователь сам формирует группы и сам оказывает то или иное воздействие, в обсервационном исследовании он может лишь наблюдать течение процесса. С другой стороны, это исследование — *ретроспективное*, поскольку имеет дело с данными, полученными в прошлом (в отличие от *проспективного*).

В обсервационном исследовании мы никогда не можем гарантировать, что группы различаются *только* тем признаком, по которому они были сформированы. Этот неустранимый недостаток исследований такого рода. Известно, например, что курильщики чаще болеют раком легких. Это считается доказательством того, что курение вызывает рак легких. Однако возможна и другая точка зрения: у людей с генетической предрасположенностью к раку легких существует и генетическая предрасположенность к курению. В обсервационном исследовании отвергнуть такое объяснение невозможно.

Ретроспективное исследование, естественно, всегда является обсервационным; разделяя недостатки последнего, оно обладает и рядом собственных. Исследователь использует информацию, собранную для других целей, — естественно, часть ее приходится реконструировать; еще часть неизбежно теряется. Меняются методы исследования, диагностические критерии и сами представления о нозологических единицах; наконец, истории болезни ведутся порой небрежно. Кроме того, имея весь материал в руках, здесь особенно трудно удержаться от непреднамеренной подтасовки.

Тем не менее ретроспективные исследования проводились и будут проводиться. Они недороги и позволяют получить большой объем информации в короткий срок. Последнее особенно важно в случае редкого заболевания: при проспективном исследовании на сбор данных уйдут годы. В примере, который мы разбираем, проспективное исследование вообще невозможно: нельзя же, в самом деле, одну группу больных лечить правильно, а другую неправильно.

Чтобы избежать ловушек обсервационного (и особенно ретроспективного) исследования, чрезвычайно важно в явном виде задать критерии, по которым больных относили к той или иной группе. Самому исследователю это поможет избежать невольного самообмана, читателю работы это даст возможность судить, насколько результаты исследования приложимы к его больным.

Кнапп и соавт. сформулировали следующие критерии включения в исследование.

1. Диагноз при выписке — острый пиелонефрит.
2. При поступлении — боли в пояснице, температура выше 37,8°С.
3. Бактериурия более 100 000 колоний/мл, определена чувствительность к антибиотикам.
4. Возраст от 18 до 44 лет (больных старше 44 лет не включали в связи с высокой вероятностью сопутствующих заболеваний, ограничивающих выбор терапии).
5. Отсутствие почечной, печеночной недостаточности, а также заболеваний, требующих хирургического лечения (эти состояния тоже ограничивают выбор терапии).
6. Больной был выписан в связи с улучшением (то есть не покинул больницу самовольно, не умер и не был переведен в другое лечебное учреждение).

Кроме того, исследователи сформулировали критерий того, что считать «правильным» лечением. Правильным считалось лечение, соответствующее рекомендациям авторитетного справочника по лекарственным средствам «Physicians' Desk Reference» («Настольный справочник врача»). По этому критерию больных разделили на две группы: леченных правильно (1-я группа) и неправильно (2-я группа). В обеих группах было по 36 больных.

Результат представлен на рис. 3.7. Средняя длительность госпитализации составила: для первой группы 4,51 сут (стандартное отклонение 1,98 сут), для второй группы 6,28 сут (стандартное отклонение 2,54 сут). Можно ли считать эти различия случайными? Прибегнем к дисперсионному анализу.

Вычислим сначала внутригрупповую дисперсию как среднюю дисперсий обеих групп:

$$s^2_{\text{вну}} = \frac{1}{2}(s_1^2 + s_2^2) = \frac{1}{2}(1{,}98^2 + 2{,}54^2) = 5{,}19.$$

Рис. 3.7. Длительность госпитализации при правильном (1-я группа) и неправильном (2-я группа) лечении. Каждый больной обозначен кружком; положение кружка соответствует сроку госпитализации. Средняя длительность госпитализации в первой группе меньше, чем во второй. Можно ли отнести это различие за счет случайности?

Теперь вычислим межгрупповую дисперсию. Среднее двух выборочных средних равно

$$\bar{X} = \frac{1}{2}(\bar{X}_1 + \bar{X}_2) = \frac{1}{2}(4,51 + 6,28) = 5,40,$$

следовательно, стандартное отклонение равно

$$s_{\bar{X}} = \sqrt{\frac{(\bar{X}_1 - \bar{X})^2 + (\bar{X}_2 - \bar{X})^2}{m-1}} =$$

$$= \sqrt{\frac{(4,51 - 5,40)^2 + (6,28 - 5,40)^2}{2-1}} = 1,25$$

и, наконец, межгрупповая дисперсия равна

$$s^2_{\text{меж}} = n s^2_{\bar{X}} = 36 \times 1,25^2 = 56,25.$$

Теперь можно вычислить F — как отношение межгрупповой к внутригрупповой дисперсии:

$$F = \frac{s^2_{\text{меж}}}{s^2_{\text{вну}}} = \frac{56{,}25}{5{,}19} = 10{,}84.$$

Рассчитаем межгрупповое и внутригрупповое число степеней свободы: $\nu_{\text{меж}} = 2 - 1 = 1$, $\nu_{\text{вну}} = 2\,(36 - 1) = 70$. Теперь по таблице 3.1 найдем критическое значение F. На пересечении столбца «1» и строки «70» находим число 7,01, набранное жирным шрифтом. То есть при уровне значимости 0,01 критическое значение F составляет 7,01. Итак, на наш вопрос, можно ли считать различия в длительности госпитализации случайными, мы можем дать ответ: вероятность этого весьма мала, меньше 1%. Леченные правильно находились в больнице меньше, чем леченные неправильно, и различия эти статистически значимы. Значит ли это, что *благодаря* правильному лечению больные выздоравливают быстрее? Увы, нет. Как это всегда бывает в обсервационном исследовании, мы не можем исключить того, что группы различались чем-то еще, кроме лечения. Может быть, врачи, которые лечат «по справочнику», просто более склонны быстрее выписывать своих больных?

Галотан и морфин при операциях на открытом сердце

Галотан — препарат, широко используемый при общей анестезии. Он обладает сильным действием, удобен в применении и очень надежен. Галотан — газ, его можно вводить через респиратор. Поступая в организм через легкие, галотан действует быстро и кратковременно, поэтому, регулируя подачу препарата, можно оперативно управлять анестезией. Однако галотан имеет существенный недостаток — он угнетает сократимость миокарда и расширяет вены, что ведет к падению АД. В связи с этим было предложено вместо галотана для общей анестезии применять морфин, который не снижает АД. Т. Конахан и соавт.* сравнили

* T. J. Conahan III, A. J. Ominsky, H. Wollman, R. A. Stroth. A prospective random comparison of halothane and morphine for open-heart anesthesia: one year experience. *Anesthesiology*, 38:528—535, 1973.

галотановую и морфиновую анестезию у больных, подвергшихся операции на открытом сердце.

В исследование включали больных, у которых не было противопоказаний ни к галотану, ни к морфину. Способ анестезии (галотан или морфин) выбирали случайным образом.

Такое исследование — со случайно отобранной контрольной группой (то есть *рандомизированное*) и наличием воздействия со стороны исследователя — называется *рандомизированным контролируемым клиническим испытанием* или просто *контролируемым испытанием*. Контролируемое испытание — это всегда *проспективное* исследование (данные получают после начала исследования), кроме того, это *экспериментальное* исследование (воздействие оказывает исследователь). Эксперимент, который в естественных науках давно стал основным методом исследования, в медицине получил распространение сравнительно недавно. Значение контролируемых испытаний трудно переоценить. Благодаря рандомизации мы уверены в том, что группы различаются только исследуемым признаком, тем самым преодолевается основной недостаток обсервационных исследований. В отличие от ретроспективного исследования, в проспективном исследовании никто до его завершения не знает, к чему оно приведет. Это уменьшает риск невольной подтасовки, о которой мы говорили выше. Быть может, по этим причинам контролируемые испытания нередко приводят к заключению о неэффективности того или иного метода лечения, когда обсервационное исследование, напротив, доказывает его эффективность*.

Но почему в таком случае не все методы лечения проходят контролируемое испытание? Немаловажную роль играет консерватизм: когда метод уже вошел в практику, трудно убедить врачей и больных, что его эффективность еще нуждается в подтверждении. Рандомизация психологически трудна: предлагая

---

* Превосходное обсуждение значения контролируемых испытаний в медицине, а также нелицеприятный анализ того, сколь малая часть общепринятых методов лечения в действительности приносит хоть какую-нибудь пользу, можно найти в работе A. K. Cochran. Effectiveness and efficiency: random reflections on health services. Nuffield Provincial Hospitals Trust, London, 1972.

по жребию лечиться тем или иным способом, врач по сути дела признается в незнании и призывает больного стать объектом эксперимента. Чтобы охватить достаточное количество больных, исследование часто приходится проводить одновременно в нескольких местах (кооперированные испытания). Конечно, это вносит приятное разнообразие в работу координаторов проекта, однако повышает его стоимость и оборачивается дополнительной нагрузкой для сотрудников сторонних медицинских учреждений. Контролируемые испытания, как и вообще проспективные исследования, иногда занимают многие годы. За это время больной может переехать в другой город, утратить интерес к эксперименту или умереть (по причинам, не относящимся к исследованию). Нередко основная трудность состоит в том, чтобы не потерять участников испытания из виду.

С выбыванием больных из исследования связан и более принципиальный недостаток контролируемых испытаний (и проспективных исследований вообще). Если в обсервационном исследовании мы не можем гарантировать сопоставимость начального состава групп, то в проспективном исследовании мы не можем гарантировать сопоставимость выбывания из исследования. Проблема состоит в том, что выбывание может быть связано с лечением. Если, например, риск побочного действия препарата связан с тяжестью заболевания, то из группы леченных будут выбывать (из-за непереносимости препарата) наиболее тяжелые больные. Тем самым состояние группы леченных будет «улучшаться». Чтобы избежать подобных иллюзий, эффективность метода лечения следует рассчитывать как долю всех больных, включенных в исследование, а не только прошедших полный курс. Даже при соблюдении этого условия результаты исследования с большим числом выбывших всегда сомнительны. Существуют и более тонкие методы анализа результатов проспективных исследований, с ними мы познакомимся позже, в гл. 11.

Удачный выбор предмета исследования позволил Конахану и соавт. избежать большинства упомянутых трудностей. Поскольку исследователей интересовали только ближайшие результаты, проблемы выбывания не возникало. Регистрировали следующие показатели: параметры гемодинамики на разных этапах операции, длительность пребывания в реанимационном отделении и

общую длительность пребывания в больнице после операции, а также послеоперационную летальность. Данные по летальности мы проанализируем после того, как познакомимся в гл. 5 с необходимыми статистическими методами. Пока же сосредоточим внимание на артериальном давлении между началом анестезии и началом операции. Именно в этот период артериальное давление наиболее адекватно отражает гипотензивное действие анестетика, поскольку в дальнейшем начинает сказываться гипотензивный эффект самой операции. Артериальное давление между началом анестезии и началом операции измеряли многократно, каждый раз вычисляя среднее артериальное давление:

$$АД_{средн} = \frac{АД_С - АД_Д}{3} + АД_Д,$$

где $АД_{средн}$ — среднее артериальное давление, $АД_Д$ — диастолическое артериальное давление, $АД_С$ — систолическое артериальное давление. Брали минимальное из полученных значений.

В исследование вошло 122 больных. У половины больных использовали галотан (1-я группа), у половины — морфин (2-я группа). Результаты представлены на рис. 3.8. Данные округлены до ближайшего четного числа. В среднем у больных, получавших галотан, минимальное $АД_{средн}$ было на 6,3 мм рт. ст. ниже, чем у больных, получавших морфин. Разброс значений довольно велик, и диапазоны значений сильно перекрываются. Стандартное отклонение в группе галотана составило 12,2 мм рт. ст., в группе морфина — 14,4 мм рт. ст.

Достаточно ли велико различие в 6,3 мм рт. ст., чтобы его нельзя было отнести за счет случайности?

Применим дисперсионный анализ. Оценкой внутригрупповой дисперсии служит среднее двух выборочных дисперсий:

$$s^2_{вну} = \frac{1}{2}(s_1^2 + s_2^2) = \frac{1}{2}(12{,}2^2 + 14{,}4^2) = 178{,}1.$$

Эта оценка дисперсии вычислена по дисперсиям отдельных выборок, поэтому она не зависит от того, различны или нет выборочные средние.

Оценим теперь дисперсию, полагая, что галотан и морфин

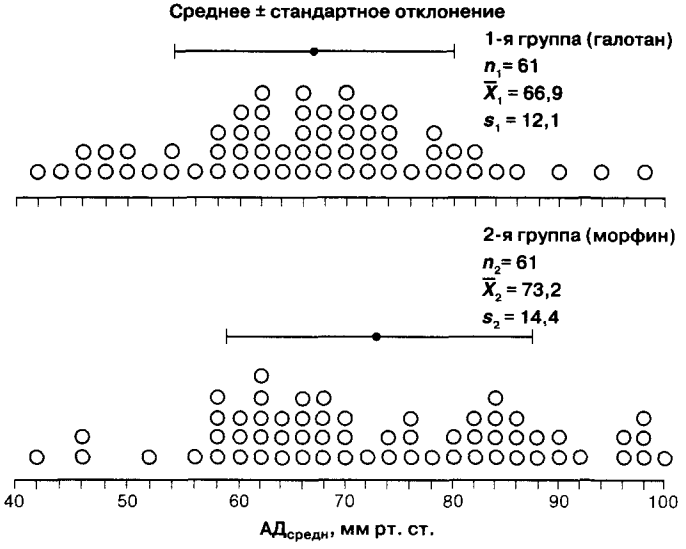


Рис. 3.8. Минимальный уровень АД$_{средн}$ между началом анестезии и началом операции при галотановой (1-я группа) и морфиновой (2-я группа) анестезии. Можно ли на основании этих данных отвергнуть нулевую гипотезу об отсутствии связи между выбором анестетика и артериальным давлением?

оказывают одинаковое действие на артериальное давление. В этом случае две группы больных, представленные на рис. 3.8, являются просто двумя случайными выборками из одной и той же совокупности. В результате стандартное отклонение выборочных средних есть оценка стандартной ошибки среднего. Среднее двух выборочных средних равно

$$\bar{X} = \frac{1}{2}(\bar{X}_1 + \bar{X}_2) = \frac{1}{2}(66{,}9 + 73{,}2) = 70.$$

Стандартное отклонение выборочных средних:

$$s_{\bar{X}} = \sqrt{\frac{(\bar{X}_1 - \bar{X})^2 + (\bar{X}_2 - \bar{X})^2}{m-1}} =$$

$$= \sqrt{\frac{(66{,}9 - 70{,}0)^2 + (73{,}2 - 70{,}0)^2}{2-1}} = 4{,}46.$$

Так как объем каждой выборки n равен 61, оценка дисперсии совокупности, полученная на основе выборочных средних, составит

$$s^2_{\text{меж}} = ns^2_{\bar{X}} = 61 \times 4{,}46^2 = 1213{,}4.$$

И наконец,

$$F = \frac{s^2_{\text{меж}}}{s^2_{\text{вну}}} = \frac{1213{,}4}{178{,}1} = 6{,}81.$$

Число степеней свободы: $\nu_{\text{меж}} = m - 1 = 2 - 1 = 1$, $\nu_{\text{вну}} = m(n-1) = 2(61-1) = 120$. В таблице 3.1 находим критическое значение F для 5% уровня значимости — 3,92. Поскольку у нас $F = 6{,}81$, то мы приходим к выводу, что различия статистически значимы. Мы можем заключить, что морфин в меньшей степени снижает артериальное давление, чем галотан. Каково клиническое значение этого результата? Мы вернемся к этому вопросу позднее.

Бег и менструации

Врачам общей практики и гинекологам очень часто приходится искать причину нерегулярности менструаций, в частности их задержки. Задержка менструации может быть признаком беременности, менопаузы; нередко она случается в начале приема пероральных контрацептивов. Задержка менструации может быть проявлением самых разных гинекологических, эндокринных и даже психических заболеваний. Среди последних особенно опасна нервная анорексия — психическое расстройство, когда женщина, убежденная в своей полноте, изнуряет себя голодом и клизмами, доходя до крайнего истощения. Без срочного и решительного врачебного вмешательства нервная анорексия может привести к смерти. Между тем есть еще одна, вполне невинная причина, которая, как полагают, может вызвать задержку менструации, — это занятия физкультурой и спортом. Чтобы проверить это предположение, Дейл и соавт.* провели обсервационное ис-

* E. Dale, D. H. Gerlach, A. L. Wilhite. Menstrual dysfunction in distance runners. *Obs. Gynecol.*, 54:47—53, 1979.

следование, целью которого было установить, есть ли связь между занятиями спортом и частотой менструаций. В исследование вошли 78 молодых женщин, разделенных на 3 группы по 26 человек в каждой. В первую — контрольную — группу вошли женщины, которые не занимались ни физкультурой, ни спортом. Вторая группа состояла из физкультурниц — они бегали трусцой и за неделю пробегали от 8 до 48 км. Женщины третьей группы — спортсменки — тренировались всерьез: за неделю они пробегали более 48 км.

На рис. 3.9 представлено распределение числа менструаций в год. В контрольной группе среднее число менструаций в год равнялось 11,5, у физкультурниц — 10,1 и у спортсменок — 9,1. Можно ли отнести эти различия на счет случайности?

Оценим дисперсию совокупности по среднему выборочных дисперсий:

$$s^2_{\text{вну}} = \frac{1}{3}(s_1^2 + s_2^2 + s_3^2) = \frac{1}{3}(1{,}3^2 + 2{,}1^2 + 2{,}4^2) = 3{,}95.$$

Чтобы оценить дисперсию по разбросу выборочных средних, нужно сначала оценить стандартную ошибку среднего, для чего вычислить стандартное отклонение среднего трех выборок. Так как среднее трех средних равно

$$\bar{X} = \frac{1}{3}(\bar{X}_1 + \bar{X}_2 + \bar{X}_3) = \frac{1}{3}(11{,}5 + 10{,}1 + 9{,}1) = 10{,}2,$$

получаем следующую оценку стандартной ошибки:

$$s_{\bar{X}} = \sqrt{\frac{(\bar{X}_1 - \bar{X})^2 + (\bar{X}_2 - \bar{X})^2 + (\bar{X}_3 - \bar{X})^2}{m-1}} =$$

$$= \sqrt{\frac{(11{,}5 - 10{,}2)^2 + (10{,}1 - 10{,}2)^2 + (9{,}1 - 10{,}2)^2}{3-1}} = 1{,}2.$$

Объем выборки n равен 26, поэтому оценка дисперсии по разбросу средних дает величину

$$s^2_{\text{меж}} = ns^2_{\bar{X}} = 26 \times 1{,}2^2 = 37{,}44.$$

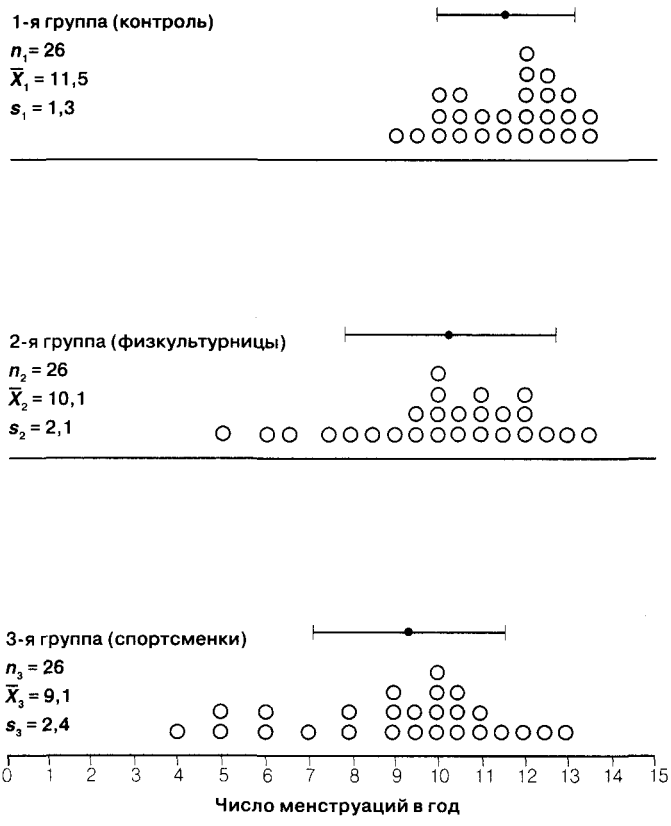


Рис. 3.9. Число менструаций в год у женщин, которые не занимались ни физкультурой, ни спортом (1-я группа), физкультурниц (2-я группа) и спортсменок (3-я группа). Среднее число менструаций различно. Можно ли отнести эти различия за счет случайности?

Наконец,

$$F = \frac{s^2_{\text{меж}}}{s^2_{\text{вну}}} = \frac{37,44}{3,95} = 9,48.$$

Число степеней свободы: $\nu_{\text{меж}} = m - 1 = 3 - 1 = 2$, $\nu_{\text{вну}} = m(n-1) = 3(26-1) = 75$. Критическое значение F при 1% уровне значимости — 4,90. Итак, различия между группами статистически зна-

чимы — вероятность случайно получить такие различия не превышает 1%. Похоже, услышав жалобы на задержку месячных, врач должен спросить: «А не занимаетесь ли вы спортом?» Однако не будем спешить — решены еще далеко не все вопросы. Можно ли утверждать, что задержки менструаций свойственны как физкультурницам, так и спортсменкам? Есть ли связь между интенсивностью нагрузок и частотой менструаций? Ответы на эти вопросы мы отложим до гл. 4.

ЗАДАЧИ

3.1. Если при родах шейка матки долго не раскрывается, то продолжительность родов увеличивается и может возникнуть необходимость кесарева сечения. Ч. О'Херлихи и Г. Мак-Дональд (C. O'Herlihy, H. MacDonald. Influence of preinduction prostaglandin E_2 vaginal gel on cervical ripening and labor. *Obstet. Gynecol.*, 54:708—710, 1979) решили выяснить, ускоряет ли гель с простагландином E_2 раскрытие шейки матки. В исследование вошло 2 группы рожениц. Роженицам первой группы вводили в шейку матки гель с простагландином E_2, роженицам второй группы вводили гель-плацебо. В обеих группах было по 21 роженице; возраст, рост и сроки беременности были примерно одинаковы. Роды в группе, получавшей гель с простагландином E_2, длились в среднем 8,5 ч (стандартное отклонение 4,7 ч), в контрольной группе — 13,9 ч (стандартное отклонение — 4,1 ч). Можно ли утверждать, что гель с простагландином E_2 сокращал продолжительность родов?

3.2. Курение считают основным фактором, предрасполагающим к хроническим обструктивным заболеваниям легких. Что касается пассивного курения, оно таким фактором обычно не считается. Дж. Уайт и Г. Фреб усомнились в безвредности пассивного курения и исследовали проходимость дыхательных путей у некурящих, пассивных и активных курильщиков (J. White, H. Froeb. Small-airways dysfunction in nonsmokers chronically exposed to tobacco smoke. *N. Engl. J. Med.*, 302:720—723, 1980). Для характеристики состояния дыхательных путей взяли один из показателей функции внешнего дыхания — максимальную объемную

скорость середины выдоха, которую измеряли во время профилактического осмотра сотрудников Калифорнийского университета в Сан-Диего. Уменьшение этого показателя — признак нарушения проходимости дыхательных путей. Данные обследования представлены в таблице.

Группа	Число обследованных	Максимальная объемная скорость середины выдоха, л/с	
		Среднее	Стандартное отклонение
Некурящие			
работающие в помещении, где не курят	200	3,17	0,74
работающие в накуренном помещении	200	2,72	0,71
Курящие			
выкуривающие небольшое число сигарет	200	2,63	0,73
выкуривающие среднее число сигарет	200	2,29	0,70
выкуривающие большое число сигарет	200	2,12	0,72

Можно ли считать максимальную объемную скорость середины выдоха одинаковой во всех группах?

3.3. Низкий уровень холестерина липопротеидов высокой плотности (ХЛПВП) — фактор риска ишемической болезни сердца. Некоторые исследования свидетельствуют, что физическая нагрузка может повысить уровень ХЛПВП. Дж. Хартунг и соавт. (G. H. Hartung et al. Relation of diet to hidh-density-lipoprotein cholesterol in middle-aged marathon runners, jogglers, and inactive men. *N. Engl. J. Med.*, 302:357—361, 1980) исследовали уровень ХЛПВП у бегунов-марафонцев, бегунов трусцой и лиц, не занимающихся спортом. Средний уровень ХЛПВП у лиц, не занимающихся спортом, составил 43,3 мг% (стандартное отклонение 14,2 мг%), у бегунов трусцой — 58,0 мг% (стандартное

отклонение 17,7 мг%) и у марафонцев — 64,8 мг% (стандартное отклонение 14,3 мг%). Будем считать, что в каждой группе было по 70 человек. Оцените статистическую значимость различий между группами.

3.4. Марихуана — наркотик, поэтому исследовать курение марихуаны на добровольцах невозможно. Исследования такого рода проводят на лабораторных животных. Г. Хубер и соавт. (G. Huber et al. Marijuana, tetrahydrocannabinol, and pulmonary arterial antibacterial defenses. *Chest*, 77:403—410, 1980) изучали влияние марихуаны на антибактериальную защиту у крыс. После ингаляционного введения бактерий крыс помещали в камеру, где специальная машина окуривала их сигаретами с марихуаной. Забив крыс, исследователи извлекали легкие и подсчитывали процент погибших бактерий, который и служил показателем состояния антибактериальной защиты. Чтобы установить, что именно влияет на антибактериальную защиту — тетрагидроканнабинолы (вещества, которые обусловливают наркотическое действие марихуаны) или просто дым, одну из групп окуривали сигаретами, из которых тетрагидроканнабинолы были удалены. В каждой группе было по 36 крыс. Являются ли различия статистически значимыми?

	Доля погибших бактерий, %	
Число сигарет	Среднее	Стандартная ошибка среднего
0 (контроль)	85,1	0,3
15	83,5	1,0
30	80,9	0,6
50	72,6	0,7
75	60	1,3
75 (тетрагидроканнабинолы удалены)	73,5	0,7
150	63,8	2,6

3.5. Стремясь отделить действие тетрагидроканнабинолов от действия дыма, Г. Хубер и соавт. изучили их действие при вну-

тривенном введении. После ингаляционного введения бактерий крысам вводили спиртовой раствор тетрагидроканнабинолов; контрольной группе вводили этиловый спирт. В обеих группах было по 36 животных. После введения тетрагидроканнабинолов доля погибших бактерий составила в среднем 51,4%, в контрольной группе — 59,4%. Стандартные ошибки среднего составили соответственно 3,2% и 3,9%. Позволяют ли эти данные утверждать, что тетрагидроканнабинолы ослабляют антибактериальную защиту?

3.6. Работа медицинской сестры сопряжена с постоянным напряжением и тяжелыми переживаниями. Груз ответственности, не уравновешенной правом принимать решения, рождает чувство усталости, раздражения и безысходности; интересная некогда работа становится ненавистным бременем. Этот синдром не совсем точно называют опустошенностью. Считается, что его развитию особенно подвержены медицинские сестры, которые работают с наиболее тяжелыми больными. Чтобы проверить это предположение, Э. Кин и соавт. (A. Keane et al. Stress in ICU and non-ICU nurses. *Nurs. Res.*, 34:231—236, 1985) провели опрос медицинских сестер с помощью специально разработанного опросника, позволяющего оценить опустошенность в баллах. Медицинских сестер разделили на три группы в зависимости от тяжести состояния больных, с которыми они работали (1-я группа — наиболее тяжелые больные, 3-я — самые легкие). Далее каждую группу разделили на две — медицинские сестры хирургических и терапевтических отделений, таким образом получилось 6 групп по 16 медицинских сестер в каждой. Являются ли различия между 6 группами статистически значимыми?

	Группа					
	1		2		3	
	Хир.	Тер.	Хир.	Тер.	Хир.	Тер.
Среднее	49,9	51,2	57,3	46,4	43,9	65,2
Стандартное отклонение	14,3	13,4	14,9	14,7	16,5	20,5
Объем выборки	16	16	16	16	16	16

3.7. Нитропруссид натрия и дофамин — препараты, которые широко используют при инфаркте миокарда. (Инфаркт мио-

карда развивается вследствие закупорки одной из коронарных артерий. Кровь перестает поступать к тому или иному участку миокарда, который в результате отмирает от недостатка кислорода.) Считается, что нитропруссид натрия облегчает работу сердца и тем самым снижает потребность миокарда в кислороде; в результате устойчивость миокарда к недостаточному кровоснабжению повышается. Дофамин препятствует падению артериального давления и увеличивает поступление крови к пораженному участку через дополнительные сосуды (так называемые коллатерали). К. Шатни и соавт. (C. Shatney et al. Effects of infusion of dopamine and nitroprusside on size of experimental myocardial infarction. *Chest*, 73:850—856, 1978) сравнили эффективность этих препаратов в опытах на собаках с инфарктом миокарда. Инфаркт миокарда вызывали перевязкой коронарной артерии, после чего вводили препарат (собакам контрольной группы вводили физиологический раствор). Через 6 часов собак забивали и взвешивали пораженный участок миокарда, результат выражали в процентах от веса левого желудочка. Препарат для каждой собаки выбирали случайным образом. Исследователь, взвешивавший миокард, не знал, какой препарат вводили собаке. Полученные данные приведены в таблице:

		Вес пораженного участка миокарда (в процентах от веса левого желудочка)	
Группа	Число животных	Среднее	Стандартная ошибка среднего
Контроль	30	15	1
Дофамин			
низкая доза	13	15	2
высокая доза	20	9	2
Нитропруссид	20	7	1

Можно ли считать различия между группами статистически значимыми? (Формулы для дисперсионного анализа при неравной численности групп найдите в прил. А.)

3.8. Считается, что выработка тромбоцитов (форменных элементов крови, играющих важную роль в ее свертывании) у но-

ворожденных регулируется иначе, чем у взрослых. Исследуя эту регуляцию, Х. Бесслер и соавт. (H. Bessler et al. Thrombopoietic activity in newborn infants. *Biol. Neonate*, 49:61—65, 1986) определили содержание тромбоцитов в крови взрослых и грудных детей разного возраста. Можно ли говорить о существовании различий в количестве тромбоцитов?

		Число тромбоцитов, $мкл^{-1}$	
Группа	Число обследованных	Среднее	Стандартное отклонение
Взрослые	15	257	159
Дети в возрасте			
4 суток	37	196	359
1 месяца	31	221	340
2 месяцев	13	280	263
4 месяцев	10	310	95

Глава 4

Сравнение двух групп: критерий Стьюдента

В предыдущей главе мы познакомились с дисперсионным анализом. Он позволяет проверить значимость различий нескольких групп. В задачах к этой главе вы видели, что нередко нужно сравнить только две группы. В этом случае можно применить критерий Стьюдента. Сейчас мы изложим его сущность и покажем, что критерий Стьюдента — это частный случай дисперсионного анализа.

Критерий Стьюдента чрезвычайно популярен, он используется более чем в половине медицинских публикаций*. Однако следует помнить, что этот критерий предназначен для сравнения именно *двух* групп, а не нескольких групп попарно. На рис. 4.1 представлено использование критерия Стьюдента в статьях из журнала Circulation. Критерий был использован в 54% статей, и чаще всего неверно. Мы покажем, что ошибочное использование критерия Стьюдента увеличивает вероятность «выявить» не-

* A. R. Feinstein. Clinical biostatistics: a survey of statistical procedures in general medical journals. *Clin. Pharmacol. Ther.*, 15:97—107, 1974.

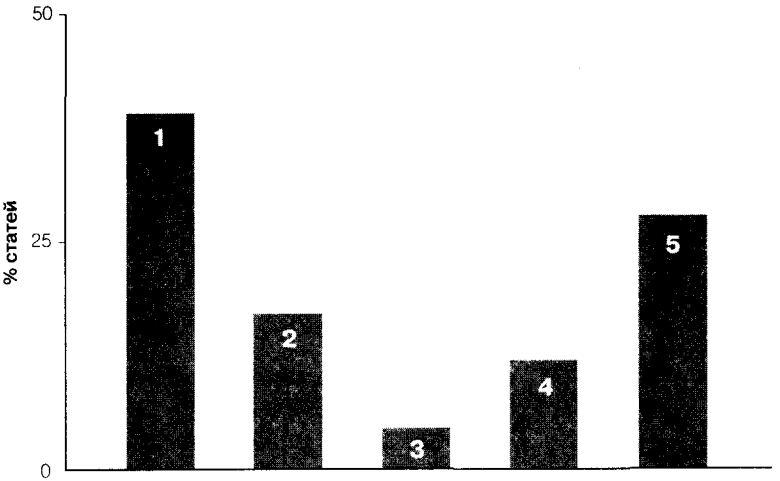


Рис. 4.1. Использование статистических методов в медицинских исследованиях. Рассмотрено 142 статьи, опубликованные в 56-м томе журнала Circulation (кроме обзоров, описаний случаев и работ по рентгенологии и патоморфологии). В 39% работ статистические методы не использовались вовсе, в 34% правильно использовали критерий Стьюдента, дисперсионный анализ или другие методы, в 27% работ критерий Стьюдента использовали неправильно — для попарного сравнения нескольких групп (S. A. Glantz. How to detect, correct, and prevent errors in the medical literature. *Circulation*, 61:1–7, 1980). **1** – не использовали статистических методов, **2** – правильно использовали критерий Стьюдента, **3** – правильно использовали дисперсионный анализ, **4** – правильно использовали другие методы, **5** – неправильно использовали критерий Стьюдента для попарного сравнения нескольких групп.

существующие различия. Например, вместо того чтобы признать несколько методов лечения равно эффективными (или неэффективными), один из них объявляют «лучшим».

ПРИНЦИП МЕТОДА

Предположим, что мы хотим испытать диуретическое действие нового препарата. Мы набираем десять добровольцев, случайным образом разделяем их на две группы — контрольную, которая получает плацебо, и экспериментальную, которая получает препарат, а затем определяем суточный диурез. Результаты пред-

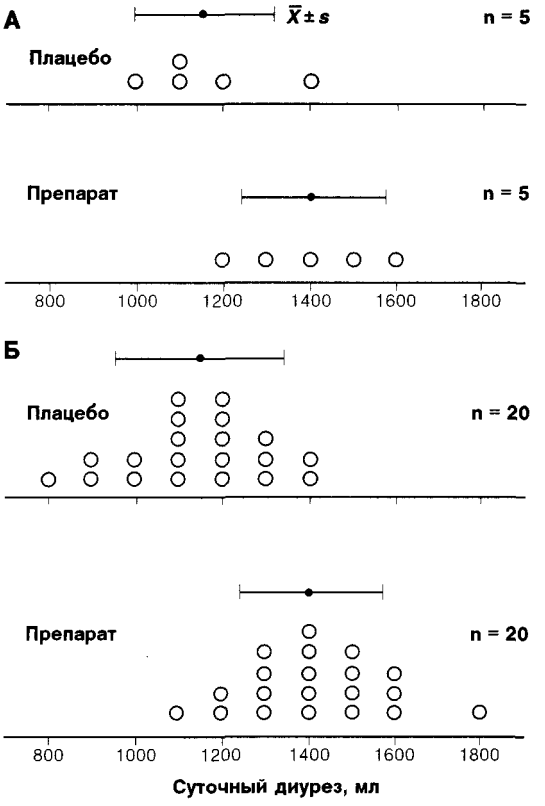


Рис. 4.2. Результаты испытаний предполагаемого диуретика. **А.** Диурез после приема плацебо и препарата. В обеих группах по 5 человек. **Б.** Теперь в обеих группах по 20 человек. Средние и стандартные отклонения остались прежними, однако доверие к результату повысилось.

ставлены на рис. 4.2А. Средний диурез в экспериментальной группе на 240 мл больше, чем в контрольной. Впрочем, подобными данными мы вряд ли кого-нибудь убедим, что препарат — диуретик. Группы слишком малы.

Повторим эксперимент, увеличив число участников. Теперь в обеих группах по 20 человек. Результаты представлены на рис. 4.2Б. Средние и стандартные отклонения примерно те же, что и в

эксперименте с меньшим числом участников. Кажется однако, что результаты второго эксперимента заслуживают большего доверия. Почему?

Вспомним, что точность выборочной оценки среднего характеризуется стандартной ошибкой среднего (см. гл. 2).

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}},$$

где n — объем выборки, а σ — стандартное отклонение совокупности, из которой извлечена выборка.

С увеличением объема выборки стандартная ошибка среднего уменьшается, следовательно уменьшается и неопределенность в оценке выборочных средних. Поэтому уменьшается и неопределенность в оценке их разности. Применительно к нашему эксперименту, мы более уверены в диуретическом действии препарата. Точнее было бы сказать, мы менее уверены в справедливости гипотезы об отсутствии диуретического действия. (Будь такая гипотеза верна, обе группы можно было бы считать двумя случайными выборками из нормально распределенной совокупности.)

Чтобы формализовать приведенные рассуждения, рассмотрим отношение:

$$t = \frac{\text{Разность выборочных средних}}{\text{Стандартная ошибка разности выборочных средних}}.$$

Для двух случайных выборок, извлеченных из одной нормально распределенной совокупности, это отношение, как правило, будет близко к нулю. Чем меньше (по абсолютной величине) t, тем больше вероятность нулевой гипотезы. Чем больше t, тем больше оснований отвергнуть нулевую гипотезу и считать, что различия статистически значимы.

Для нахождения величины t нужно знать разность выборочных средних и ее ошибку. Вычислить разность выборочных средних нетрудно — просто вычтем из одного среднего другое. Сложнее найти ошибку разности. Для этого обратимся к более общей задаче нахождения стандартного отклонения разности двух чисел, случайным образом извлеченных из одной совокупности.

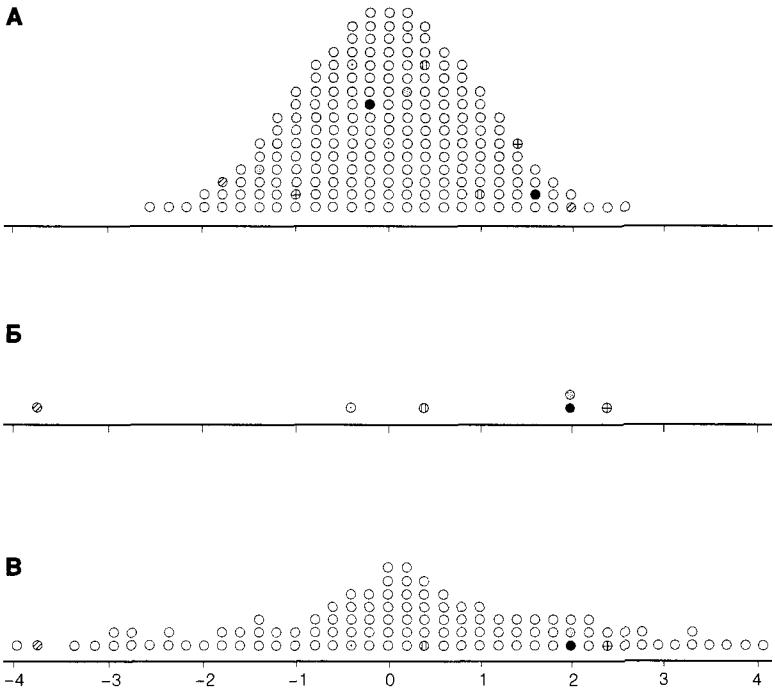


Рис. 4.3. А. Из этой совокупности мы будем наугад извлекать пары и вычислять разности. **Б.** Разности первых шести пар. **В.** Разности еще ста пар. Разброс разностей больше, чем разброс самих значений.

СТАНДАРТНОЕ ОТКЛОНЕНИЕ РАЗНОСТИ

На рис. 4.3А представлена совокупность из 200 членов. Среднее равно 0, стандартное отклонение 1. Выберем наугад два члена совокупности и вычислим разность. Выбранные члены помечены на рис. 4.3А черными кружками, полученная разность представлена таким же кружком на рис. 4.3Б. Извлечем еще пять пар (на рисунках они различаются штриховкой), вычислим разность для каждой пары, результат снова поместим на рис 4.3Б. Похоже, что разброс разностей больше разброса исходных данных. Извлечем наугад из исходной совокупности еще 100 пар, для ка-

ждой из которых вычислим разность. Теперь все разности, включая вычисленные ранее, изображены на рис. 4.3В. Стандартное отклонение для полученной совокупности разностей — примерно 1,4, то есть на 40% больше, чем в исходной совокупности.

Можно доказать, что *дисперсия разности двух случайно извлеченных значений равна сумме дисперсий совокупностей, из которых они извлечены**.

В частности, если извлекать значения из одной совокупно-

---

* Интересно, что дисперсия *суммы* двух случайно извлеченных значений тоже равна сумме дисперсий совокупностей, из которых они извлечены. Отсюда можно вывести формулу для стандартной ошибки среднего:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

Предположим, что мы случайным образом извлекли n значений из совокупности, имеющей стандартное отклонение σ. Выборочное среднее равно

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + X_3 + \ldots + X_n),$$

поэтому

$$n\bar{X} = X_1 + X_2 + X_3 + \ldots + X_n.$$

Так как дисперсия каждого из X_i равна σ^2, дисперсия величины $n\bar{X}$ составит

$$\sigma^2_{n\bar{X}} = \sigma^2 + \sigma^2 + \sigma^2 + \ldots + \sigma^2 = n\sigma^2,$$

а стандартное отклонение

$$\sigma_{n\bar{X}} = \sqrt{n}\sigma.$$

Нам нужно найти стандартное отклонение среднего $\bar{X}$, тождественно равного $n\bar{X}/n$, поэтому

$$\sigma_{\bar{X}} = \frac{\sigma_{n\bar{X}}}{n} = \sqrt{n}\frac{\sigma}{n} = \frac{\sigma}{\sqrt{n}}.$$

Мы получили формулу, которой неоднократно пользовались в предыдущих главах, — формулу для стандартной ошибки среднего. Заметим, что, выводя ее, мы не делали никаких допущений о совокупности, из которой извлечена выборка. В частности, мы не требовали, чтобы она имела нормальное распределение.

сти, то дисперсия их разности будет равна удвоенной дисперсии этой совокупности. Говоря формально, если значение X извлечено из совокупности, имеющей дисперсию σ_X^2, а значение Y из совокупности, имеющей дисперсию σ_Y^2, то распределение всех возможных значений $X - Y$ имеет дисперсию

$$\sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2.$$

Почему дисперсия разностей больше дисперсии совокупности, легко понять на нашем примере (см. рис. 4.3): в половине случаев члены пары лежат по разные стороны от среднего, поэтому их разность еще больше отклоняется от среднего, чем они сами.

Продолжим рассматривать рис. 4.3. Все пары извлекали из одной совокупности. Ее дисперсия равна 1. В таком случае дисперсия разностей будет

$$\sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2 = 1 + 1 = 2.$$

Стандартное отклонение есть квадратный корень из дисперсии. Поэтому стандартное отклонение разностей равно $\sqrt{2}$, то есть больше стандартного отклонения исходной совокупности примерно на 40%, как и получилось в нашем примере.

Чтобы оценить дисперсию разности членов двух совокупностей по выборочным данным, нужно в приведенной выше формуле заменить дисперсии их выборочными оценками:

$$s_{X-Y}^2 = s_X^2 + s_Y^2.$$

Этой формулой можно воспользоваться и для оценки стандартной ошибки разности выборочных средних. В самом деле, стандартная ошибка выборочного среднего — это стандартное отклонение совокупности средних значений всех выборок объемом n. Поэтому

$$s_{\bar{X}-\bar{Y}}^2 = s_{\bar{X}}^2 + s_{\bar{Y}}^2.$$

Тем самым, искомая стандартная ошибка разности средних

$$s_{\bar{X}-\bar{Y}} = \sqrt{s_{\bar{X}}^2 + s_{\bar{Y}}^2}.$$

Теперь мы можем вычислить отношение t.

КРИТИЧЕСКОЕ ЗНАЧЕНИЕ t

Напомним, что мы рассматриваем отношение

$$t = \frac{\text{Разность выборочных средних}}{\text{Стандартная ошибка разности выборочных средних}}.$$

Воспользовавшись результатом предыдущего раздела, имеем

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s^2_{\bar{X}_1} + s^2_{\bar{X}_2}}}.$$

Если ошибку среднего выразить через выборочное стандартное отклонение, получим другую запись этой формулы:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{n}}},$$

где n — объем выборки.

Если обе выборки извлечены из одной совокупности, то выборочные дисперсии s_1^2 и s_2^2 — это оценки одной и той же дисперсии σ^2. Поэтому их можно заменить на *объединенную оценку дисперсии*. Для выборок равного объема объединенная оценка дисперсии вычисляется как

$$s^2 = \frac{s_1^2 + s_2^2}{2}.$$

Значение t, полученное на основе объединенной оценки:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s^2}{n} + \frac{s^2}{n}}}.$$

Если объем выборок одинаков, оба способа вычисления t дадут одинаковый результат. Однако если объем выборок разный, то это не так. Вскоре мы увидим, почему важно вычислять объединенную оценку дисперсии, а пока посмотрим, какие значения

t мы будем получать, извлекая случайные пары выборок из одной и той же нормально распределенной совокупности.

Так как выборочные средние обычно близки к среднему по совокупности, значение *t* будет близко к нулю. Однако иногда мы все же будем получать большие по абсолютной величине значения *t* (вспомним опыты с *F* в предыдущей главе). Чтобы понять, какую величину *t* следует считать достаточно «большой», чтобы отвергнуть нулевую гипотезу, проведем мысленный эксперимент, подобный тому, что мы делали в предыдущей главе. Вернемся к испытаниям предполагаемого диуретика. Допустим, что в действительности препарат не оказывает диуретического действия. Тогда и контрольную группу, которая получает плацебо, и экспериментальную, которая получает препарат, можно считать случайными выборками из одной совокупности. Пусть это будет совокупность из 200 человек, представленная на рис. 4.4А. Члены контрольной и экспериментальной групп различаются штриховкой. В нижней части рисунка данные по этим двум выборкам показаны так, как их видит исследователь. Взглянув на эти данные, трудно подумать, что препарат — диуретик. Полученное по этим выборкам значение *t* равно –0,2.

Разумеется, с не меньшим успехом можно было бы извлечь любую другую пару выборок, что и сделано на рис. 4.4Б. Как и следовало ожидать, две новые выборки отличаются как друг от друга, так и от извлеченных ранее (рис. 4.4А). Интересно, что на этот раз нам «повезло» — средний диурез довольно сильно различается. Соответствующее значение *t* равно –2,1. На рис. 4.4В изображена еще одна пара выборок. Они отличаются друг от друга и от выборок с рис. 4.4А и 4.4Б. Значение *t* для них равно 0.

Разных пар выборок можно извлечь более 10^{27}. На рис. 4.5А приведено распределение значений *t*, вычисленных по 200 парам выборок. По нему уже можно судить о распределении *t*. Оно симметрично относительно нуля, поскольку любую из пары выборок можно счесть «первой». Как мы и предполагали, чаще всего значения *t* близки к нулю; значения, меньшие –2 и большие +2, встречаются редко.

На рис. 4.5Б видно, что в 10 случаях из 200 (в 5% всех случаев) *t* меньше –2,1 или больше +2,1. Иначе говоря, если обе выборки извлечены из одной совокупности, вероятность того, что значение

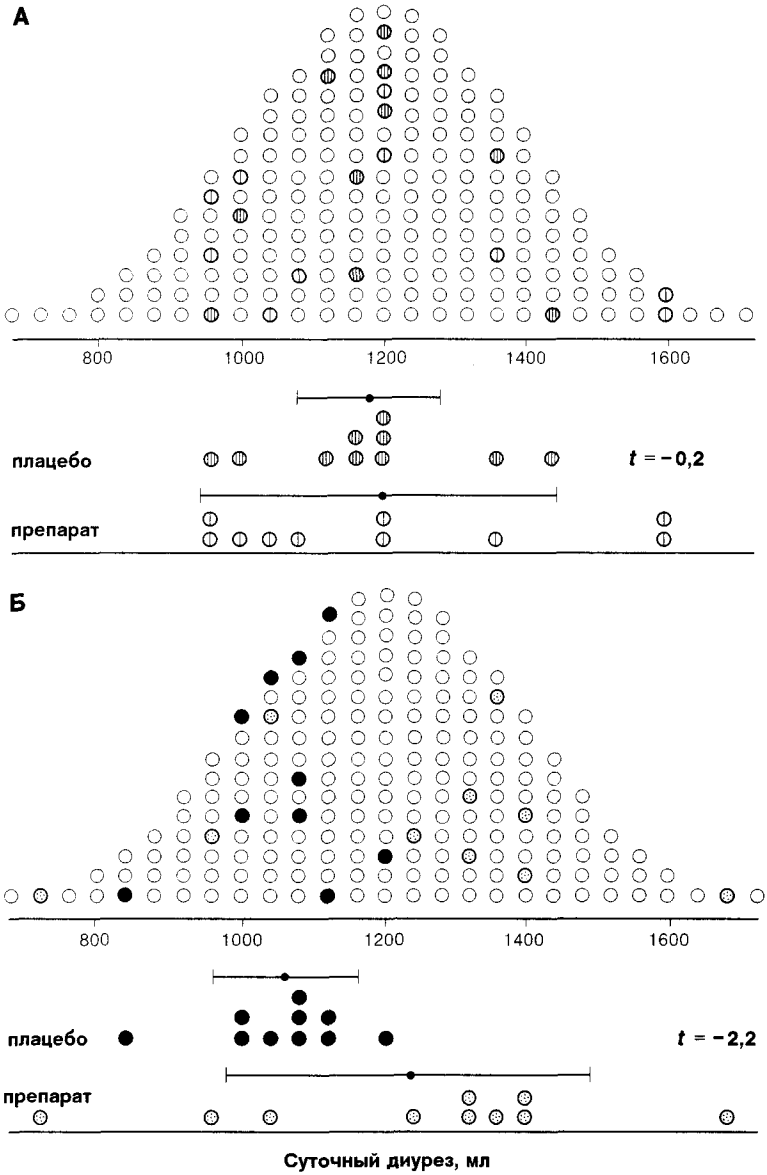

А
800
1000
1200
1400
1600
плацебо
t = −0,2
препарат
Б
800
1000
1200
1400
1600
плацебо
t = −2,2
препарат
Суточный диурез, мл

Рис. 4.4.

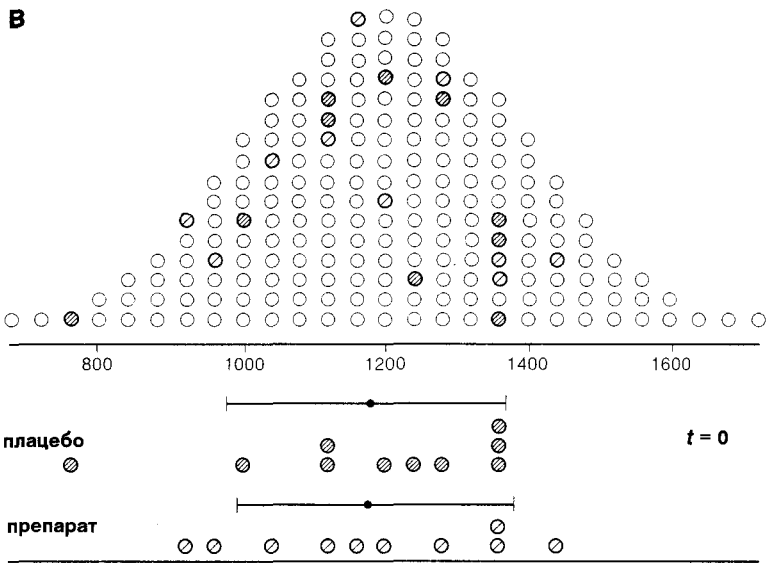


Рис. 4.4. Испытания предполагаемого диуретика. **А.** В действительности препарат не обладает диуретическим действием, поэтому обе группы — просто две случайные выборки из совокупности, показанной в верхней части рисунка. Члены совокупности, которым посчастливилось принять участие в исследовании, помечены штриховкой. В нижней части рисунка данные показаны такими, какими их видит исследователь. Вряд ли он решит, что препарат — диуретик: средний диурез в группах различается очень незначительно. **Б.** Исследователю могла бы попасться и такая пара выборок. В этом случае он наверняка счел бы препарат диуретиком. **В.** Еще две выборки из той же совокупности.

t лежит вне интервала от −2,1 до +2,1, составляет 5%. Продолжая извлекать пары выборок, мы увидим, что распределение принимает форму гладкой кривой, показанной на рис. 4.5В. Теперь 5% крайних значений соответствуют закрашенным областям графика левее −2,1 и правее +2,1. Итак, мы нашли, что если две выборки извлечены из одной и той же совокупности, то вероятность получить значение *t*, большее +2,1 или меньшее −2,1, составляет всего 5%. Следовательно, если значение *t* находится вне

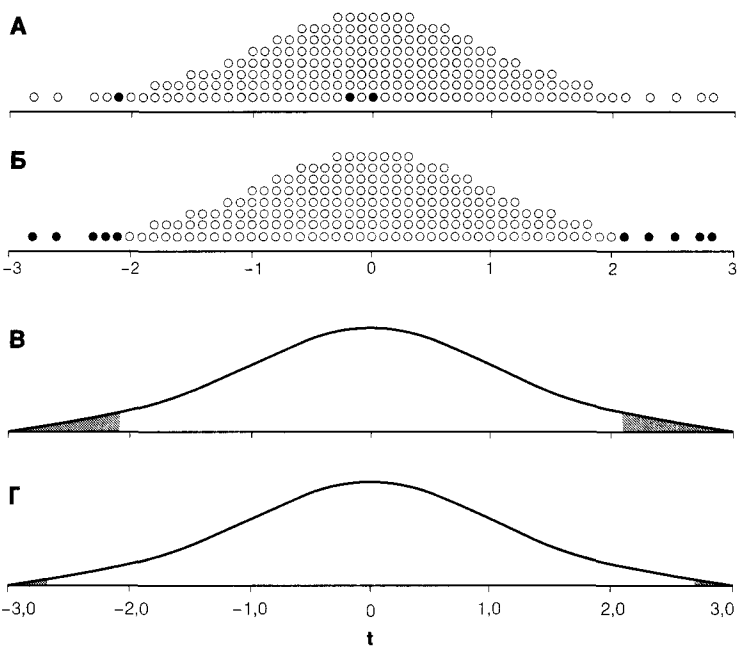


Рис. 4.5. А. Из совокупности, показанной на рис. 4.4, извлекли 200 пар случайных выборок по 10 членов в каждой, для каждой пары рассчитали значение t и нанесли его на график. Значения t для трех пар выборок с рис. 4.4 помечены черным. Большая часть значений сгруппирована вокруг нуля, однако некоторые значения по абсолютной величине превышают 1,5 и даже 2. **Б.** Число значений, по абсолютной величине превышающих 2,1, составляет 5%. **В.** Продолжая извлекать пары выборок, в конце концов мы получим гладкую кривую. 5% наибольших (по абсолютной величине) значений образуют две заштрихованные области (сумма заштрихованных площадей как раз и составляет 5% всей площади под кривой). Следовательно, «большие» значения t начинаются там, где начинается заштрихованная область, то есть с $t = \pm 2{,}1$. Вероятность получить столь высокое значение t, извлекая случайные выборки из одной совокупности, не превышает 5%. **Г.** Описанный способ выбора критического значения t предопределяет возможность ошибки: в 5% случаев мы будем находить различия там, где их нет. Чтобы снизить вероятность ошибочного заключения, мы можем выбрать более высокое критическое значение. Например, чтобы площадь заштрихованной области составляла 1% от общей площади под кривой, критическое значение должно составлять 2,878.

интервала от –2,1 до +2,1, нулевую гипотезу следует отклонить, а наблюдаемые различия признать статистически значимыми.

Обратите внимание, что таким образом мы выявляем отличия экспериментальной группы от контрольной как в меньшую, так и в большую сторону — именно поэтому мы отвергаем нулевую гипотезу как при $t < -2{,}1$, так и при $t > +2{,}1$. Этот вариант критерия Стьюдента называется *двусторонним*; именно его обычно и используют. Существует и *односторонний* вариант критерия Стьюдента. Используется он гораздо реже, и в дальнейшем, говоря о критерии Стьюдента, мы будем иметь в виду двусторонний вариант.

Вернемся к рис. 4.4Б. На нем показаны две случайные выборки из одной и той же совокупности, при этом $t = -2{,}2$. Как мы только что выяснили, нам следует отвергнуть нулевую гипотезу и признать исследуемый препарат диуретиком, что, самой собой, неверно. Хотя все расчеты были выполнены правильно, вывод ошибочен. Увы, такие случаи возможны.

Разберемся подробнее. Если значение t меньше –2,1 или больше +2,1, то при уровне значимости 0,05 мы сочтем различия статистически значимыми. Это означает, что если бы наши группы представляли собой две случайные выборки из одной и той же совокупности, то вероятность получить наблюдаемые различия (или более сильные) равна 0,05. Следовательно, ошибочный вывод о существовании различий мы будем делать в 5% случаев. Один из таких случаев и показан на рис. 4.4Б.

Чтобы застраховаться от подобных ошибок, можно принять уровень значимости не 0,05, а, скажем, 0,01. Тогда, как видно из рис. 4.5Г, мы должны отвергать нулевую гипотезу при $t < -2{,}88$ или $t > +2{,}88$. Теперь-то рис. 4.4Б нас не проведет — мы не признаем подобные различия статистически значимыми. Однако, во-первых, ошибочные выводы о существовании различий все же не исключены, просто их вероятность снизилась до 1%, и, во-вторых, вероятность *не найти различий там, где они есть,* теперь повысилась. О последней проблеме подробнее мы поговорим в гл. 6.

Критические значения t (подобно критическим значениям F, они сведены в таблицу) зависят не только от уровня значимости, но и от числа степеней свободы ν. Если объем обеих выбо-

Таблица 4.1. Критические значения *t* (двусторонний вариант)

	Уровень значимости α								
ν	0,50	0,20	0,10	0,05	0,02	0,01	0,005	0,002	0,001
1	1,000	3,078	6,314	12,706	31,821	63,657	127,321	318,309	636,619
2	0,816	1,886	2,920	4,303	6,965	9,925	14,089	22,327	31,599
3	0,765	1,638	2,353	3,182	4,541	5,841	7,453	10,215	12,924
4	0,741	1,533	2,132	2,776	3,747	4,604	5,598	7,173	8,610
5	0,727	1,476	2,015	2,571	3,365	4,032	4,773	5,893	6,869
6	0,718	1,440	1,943	2,447	3,143	3,707	4,317	5,208	5,959
7	0,711	1,415	1,895	2,365	2,998	3,499	4,029	4,785	5,408
8	0,706	1,397	1,860	2,306	2,896	3,355	3,833	4,501	5,041
9	0,703	1,383	1,833	2,262	2,821	3,250	3,690	4,297	4,781
10	0,700	1,372	1,812	2,228	2,764	3,169	3,581	4,144	4,587
11	0,697	1,363	1,796	2,201	2,718	3,106	3,497	4,025	4,437
12	0,695	1,356	1,782	2,179	2,681	3,055	3,428	3,930	4,318
13	0,694	1,350	1,771	2,160	2,650	3,012	3,372	3,852	4,221
14	0,692	1,345	1,761	2,145	2,624	2,977	3,326	3,787	4,140
15	0,691	1,341	1,753	2,131	2,602	2,947	3,286	3,733	4,073
16	0,690	1,337	1,746	2,120	2,583	2,921	3,252	3,686	4,015
17	0,689	1,333	1,740	2,110	2,567	2,898	3,222	3,646	3,965
18	0,688	1,330	1,734	2,101	2,552	2,878	3,197	3,610	3,922
19	0,688	1,328	1,729	2,093	2,539	2,861	3,174	3,579	3,883
20	0,687	1,325	1,725	2,086	2,528	2,845	3,153	3,552	3,850
21	0,686	1,323	1,721	2,080	2,518	2,831	3,135	3,527	3,819
22	0,686	1,321	1,717	2,074	2,508	2,819	3,119	3,505	3,792
23	0,685	1,319	1,714	2,069	2,500	2,807	3,104	3,485	3,768
24	0,685	1,318	1,711	2,064	2,492	2,797	3,091	3,467	3,745
25	0,684	1,316	1,708	2,060	2,485	2,787	3,078	3,450	3,725
26	0,684	1,315	1,706	2,056	2,479	2,779	3,067	3,435	3,707
27	0,684	1,314	1,703	2,052	2,473	2,771	3,057	3,421	3,690
28	0,683	1,313	1,701	2,048	2,467	2,763	3,047	3,408	3,674
29	0,683	1,311	1,699	2,045	2,462	2,756	3,038	3,396	3,659
30	0,683	1,310	1,697	2,042	2,457	2,750	3,030	3,385	3,646
31	0,682	1,309	1,696	2,040	2,453	2,744	3,022	3,375	3,633
32	0,682	1,309	1,694	2,037	2,449	2,738	3,015	3,365	3,622
33	0,682	1,308	1,692	2,035	2,445	2,733	3,008	3,356	3,611
34	0,682	1,307	1,691	2,032	2,441	2,728	3,002	3,348	3,601
35	0,682	1,306	1,690	2,030	2,438	2,724	2,996	3,340	3,591
36	0,681	1,306	1,688	2,028	2,434	2,719	2,990	3,333	3,582
37	0,681	1,305	1,687	2,026	2,431	2,715	2,985	3,326	3,574
38	0,681	1,304	1,686	2,024	2,429	2,712	2,980	3,319	3,566
39	0,681	1,304	1,685	2,023	2,426	2,708	2,976	3,313	3,558
40	0,681	1,303	1,684	2,021	2,423	2,704	2,971	3,307	3,551

Таблица 4.1. Окончание

	Уровень значимости α								
ν	0,50	0,20	0,10	0,05	0,02	0,01	0,005	0,002	0,001
42	0,680	1,302	1,682	2,018	2,418	2,698	2,963	3,296	3,538
44	0,680	1,301	1,680	2,015	2,414	2,692	2,956	3,286	3,526
46	0,680	1,300	1,679	2,013	2,410	2,687	2,949	3,277	3,515
48	0,680	1,299	1,677	2,011	2,407	2,682	2,943	3,269	3,505
50	0,679	1,299	1,676	2,009	2,403	2,678	2,937	3,261	3,496
52	0,679	1,298	1,675	2,007	2,400	2,674	2,932	3,255	3,488
54	0,679	1,297	1,674	2,005	2,397	2,670	2,927	3,248	3,480
56	0,679	1,297	1,673	2,003	2,395	2,667	2,923	3,242	3,473
58	0,679	1,296	1,672	2,002	2,392	2,663	2,918	3,237	3,466
60	0,679	1,296	1,671	2,000	2,390	2,660	2,915	3,232	3,460
62	0,678	1,295	1,670	1,999	2,388	2,657	2,911	3,227	3,454
64	0,678	1,295	1,669	1,998	2,386	2,655	2,908	3,223	3,449
66	0,678	1,295	1,668	1,997	2,384	2,652	2,904	3,218	3,444
68	0,678	1,294	1,668	1,995	2,382	2,650	2,902	3,214	3,439
70	0,678	1,294	1,667	1,994	2,381	2,648	2,899	3,211	3,435
72	0,678	1,293	1,666	1,993	2,379	2,646	2,896	3,207	3,431
74	0,678	1,293	1,666	1,993	2,378	2,644	2,894	3,204	3,427
76	0,678	1,293	1,665	1,992	2,376	2,642	2,891	3,201	3,423
78	0,678	1,292	1,665	1,991	2,375	2,640	2,889	3,198	3,420
80	0,678	1,292	1,664	1,990	2,374	2,639	2,887	3,195	3,416
90	0,677	1,291	1,662	1,987	2,368	2,632	2,878	3,183	3,402
100	0,677	1,290	1,660	1,984	2,364	2,626	2,871	3,174	3,390
120	0,677	1,289	1,658	1,980	2,358	2,617	2,860	3,160	3,373
140	0,676	1,288	1,656	1,977	2,353	2,611	2,852	3,149	3,361
160	0,676	1,287	1,654	1,975	2,350	2,607	2,846	3,142	3,352
180	0,676	1,286	1,653	1,973	2,347	2,603	2,842	3,136	3,345
200	0,676	1,286	1,653	1,972	2,345	2,601	2,839	3,131	3,340
∞	0,6745	1,2816	1,6449	1,9600	2,3263	2,5758	2,8070	3,0902	3,2905

J. H. Zar. Biostatistical analysis (2 ed.). Prentice-Hall, Englewood Cliffs, N. J., 1984.

рок — n, то число степеней свободы для критерия Стьюдента равно $2(n-1)$. Чем больше объем выборок, тем меньше критическое значение t. Это и понятно — чем больше выборка, тем менее выборочные оценки зависят от случайных отклонений и тем точнее представляют исходную совокупность.

ВЫБОРКИ ПРОИЗВОЛЬНОГО ОБЪЕМА

Критерий Стьюдента легко обобщается на случай, когда выборки содержат неодинаковое число членов. Напомним, что по определению

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s^2_{\bar{X}_1} + s^2_{\bar{X}_2}}},$$

где $s_{\bar{X}_1}$ и $s_{\bar{X}_2}$ — стандартные ошибки средних для двух выборок.

Если объем первой выборки равен n_1, а объем второй — n_2, то

$$s^2_{\bar{X}_1} = \frac{s_1^2}{n_1} \text{ и } s^2_{\bar{X}_2} = \frac{s_2^2}{n_2},$$

где s_1 и s_2 — стандартные отклонения выборок. Перепишем определение t, используя выборочные стандартные отклонения:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

Объединенная оценка дисперсии для выборок объема n_1 и n_2 равна

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

Тогда

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}.$$

Это определение t для выборок произвольного объема. Число степеней свободы $\nu = n_1 + n_2 - 2$.

Заметим, что если объемы выборок равны, то есть $n_1 = n_2 = n$, то мы получим ранее использовавшуюся формулу для t.

ПРОДОЛЖЕНИЕ ПРИМЕРОВ

Применим теперь критерий Стьюдента к тем данным, которые рассматривались при изучении дисперсионного анализа. Выводы, которые мы получим, не будут отличаться от прежних, поскольку, как говорилось, критерий Стьюдента есть частный случай дисперсионного анализа.

Позволяет ли правильное лечение сократить срок госпитализации?

Обратимся к рис. 3.7. Средняя продолжительность госпитализации 36 больных пиелонефритом, получавших правильное (соответствующее официальным рекомендациям) лечение, составила 4,51 сут, а 36 больных, получавших неправильное лечение, — 6,28 сут. Стандартные отклонения для этих групп — соответственно 1,98 сут и 2,54 сут. Так как численность групп одна и та же, объединенная оценка дисперсии $s^2 = \frac{1}{2}(1{,}98^2 + 2{,}54^2) = 5{,}18$. Подставив эту величину в выражение для t, получим

$$t = \frac{4{,}51 - 6{,}28}{\sqrt{\frac{5{,}18}{36} + \frac{5{,}18}{36}}} = -3{,}30.$$

Число степеней свободы $\nu = 2(n-1) = 2(36-1) = 70$. По таблице 4.1 находим, что для 1% уровня значимости критическое значение t составляет 2,648, то есть меньше, чем мы получили (по абсолютной величине). Следовательно, если бы наши группы представляли собой две случайные выборки из одной совокупности, то вероятность получить наблюдаемые различия была бы меньше 1%. Итак, различия в сроках госпитализации статистически значимы.

Галотан и морфин при операциях на открытом сердце

В исследовании Конахана и соавт. (рис. 3.8) минимальное $АД_{средн}$ между началом анестезии и началом операции составляло в среднем: при галотановой анестезии 66,9 мм рт. ст., при морфино-

Таблица 4.2. Показатели гемодинамики при галотановой и морфиновой анестезии

	Галотан ($n = 9$)		Морфин ($n = 16$)	
Показатель	Среднее	Стандартное отклонение	Среднее	Стандартное отклонение
Наилучший сердечный индекс	2,08	1,05	1,75	0,88
Среднее артериальное давление при наилучшем сердечном индексе, мм рт. ст.	76,8	13,8	91,4	19,6
Общее периферическое сосудистое сопротивление при наилучшем сердечном индексе, дин·с·см$^{-5}$	2210	1200	2830	1130

T. J. Conahan et al. A prospective random comparison of halothane and morphine for open-heart anesthesia: one year experience. *Anesthesiology,* 38:528–535, 1973.

вой — 73,2 мм рт. ст. Стандартные отклонения составляли соответственно 12,2 и 14,4 мм рт. ст. В каждой группе был 61 больной.

Вычислим объединенную оценку дисперсии:

$$s^2 = \frac{1}{2}\left(12{,}2^2 + 14{,}4^2\right) = 178{,}1,$$

тогда

$$t = \frac{66{,}9 - 73{,}2}{\sqrt{\frac{178{,}1}{61} + \frac{178{,}1}{61}}} = -2{,}607.$$

Число степеней свободы $\nu = 2(n-1) = 2(61-1) = 120$. По таблице 4.1 находим, что для 5% уровня значимости критическое значение t составляет 1,980, то есть меньше, чем мы получили. Заключаем, что морфин меньше снижает артериальное давление, чем галотан.

Конахан и соавт. измеряли еще один параметр гемодинамики — минутный объем сердца (объем крови, который левый желудочек перекачивает за минуту). Поскольку этот объем зависит

от размеров тела, деятельность сердца (которая и интересовала исследователей) лучше характеризуется *сердечным индексом* — отношением минутного объема сердца к площади поверхности тела. В группе галотана сердечный индекс определили у 9 больных (табл. 4.2), он составил в среднем 2,08 л/мин/м^2 (стандартное отклонение 1,05 л/мин/м^2), у 16 больных в группе морфина — 1,75 л/мин/м^2 (стандартное отклонение 0,88 л/мин/м^2). Является ли это различие статистически значимым?

Найдем объединенную оценку дисперсии

$$s^2 = \frac{(9-1)\,1{,}05^2 + (16-1)\,0{,}88^2}{9+16-2} = 0{,}89,$$

и поэтому

$$t = \frac{2{,}08 - 1{,}75}{\sqrt{\dfrac{0{,}89}{9} + \dfrac{0{,}89}{16}}} = 0{,}84.$$

Число степеней свободы $\nu = 9+16-2 = 23$. Критическое значение t при 5% уровне значимости составляет 2,069, что больше полученного нами! Итак, статистически значимых различий не найдено. Можно ли утверждать, что различий *нет*? Ответ на этот вопрос мы узнаем в гл. 6.

КРИТЕРИЙ СТЬЮДЕНТА С ТОЧКИ ЗРЕНИЯ ДИСПЕРСИОННОГО АНАЛИЗА*

Хотя критерий Стьюдента является просто вариантом дисперсионного анализа, этот факт осознается очень немногими. Покажем, что в случае двух групп справедливо равенство $F = t^2$.

Рассмотрим две выборки равного объема n со средними $\bar{X}_1$ и $\bar{X}_2$ и стандартными отклонениями s_1 и s_2.

Как вы помните, отношение F есть отношение двух оценок дисперсии. Первая, внутригрупповая оценка есть среднее выборочных дисперсий:

---

* Этот раздел посвящен сугубо математической стороне дела, и его можно пропустить без ущерба для понимания дальнейшего изложения.

$$s^2_{\text{вну}} = \frac{1}{2}(s_1^2 + s_2^2).$$

Вторая, межгрупповая оценка вычисляется по выборочным средним:

$$s_{\bar{X}} = \sqrt{\frac{(\bar{X}_1 - \bar{X})^2 + (\bar{X}_2 - \bar{X})^2}{2-1}},$$

следовательно,

$$s^2_{\bar{X}} = (\bar{X}_1 - \bar{X})^2 + (\bar{X}_2 - \bar{X})^2,$$

где $\bar{X}$ — среднее двух выборочных средних:

$$\bar{X} = \frac{1}{2}(\bar{X}_1 + \bar{X}_2).$$

Исключим $\bar{X}$ из формулы для $s^2_{\bar{X}}$:

$$s^2_{\bar{X}} = \left[\bar{X}_1 - \frac{1}{2}(\bar{X}_1 + \bar{X}_2)\right]^2 + \left[\bar{X}_2 - \frac{1}{2}(\bar{X}_1 + \bar{X}_2)\right]^2 =$$

$$= \left(\frac{1}{2}\bar{X}_1 - \frac{1}{2}\bar{X}_2\right)^2 + \left(\frac{1}{2}\bar{X}_2 - \frac{1}{2}\bar{X}_1\right)^2.$$

Если разность возводится в квадрат, все равно что из чего вычитать: $(a-b)^2 = (b-a)^2$. Поэтому

$$s^2_{\bar{X}} = \left(\frac{1}{2}\bar{X}_1 - \frac{1}{2}\bar{X}_2\right)^2 + \left(\frac{1}{2}\bar{X}_1 - \frac{1}{2}\bar{X}_2\right)^2 =$$

$$= 2\left[\frac{1}{2}(\bar{X}_1 - \bar{X}_2)\right]^2 = \frac{1}{2}(\bar{X}_1 - \bar{X}_2)^2.$$

Таким образом, межгрупповая оценка дисперсии

$$s^2_{\text{меж}} = ns^2_{\bar{X}} = \frac{n}{2}(\bar{X}_1 - \bar{X}_2)^2.$$

F есть отношение межгрупповой оценки к внутригрупповой и равно

$$F = \frac{s^2_{\text{меж}}}{s^2_{\text{вну}}} = \frac{\frac{n}{2}(\bar{X}_1 - \bar{X}_2)^2}{\frac{1}{2}(s_1^2 + s_2^2)} = \frac{(\bar{X}_1 - \bar{X}_2)^2}{\frac{s_1^2}{n} + \frac{s_2^2}{n}} =$$

$$= \left(\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{n}}} \right)^2.$$

Но величина в скобках есть не что иное, как t. Тем самым,

$$F = t^2.$$

Межгрупповое число степеней свободы в F равно числу групп минус единица, то есть $2 - 1 = 1$. Внутригрупповое число степеней свободы равно произведению числа групп на число, равное численности каждой группы минус единица, то есть $2(n-1)$. Но это как раз число степеней свободы в критерии Стьюдента.

Таким образом, можно сказать, что в случае сравнения двух групп критерий Стьюдента и дисперсионный анализ — варианты одного критерия. Конечно, если групп больше двух, дисперсионный анализ в форме критерия Стьюдента неприменим и нужно воспользоваться общим вариантом дисперсионного анализа, изложенным в гл. 3.

ОШИБКИ В ИСПОЛЬЗОВАНИИ КРИТЕРИЯ СТЬЮДЕНТА

Критерий Стьюдента предназначен для сравнения двух групп. Однако на практике он широко (и неправильно — см. рис. 4.1) используется для оценки различий большего числа групп посредством попарного их сравнения. При этом вступает в силу *эффект множественных сравнений*, который нам еще неоднократно встретится в разнообразных обличиях.

Рассмотрим пример. Исследуют влияние препаратов А и Б на уровень глюкозы плазмы. Исследование проводят на трех группах — получавших препарат А, получавших препарат Б и получавших плацебо В. С помощью критерия Стьюдента проводят

3 парных сравнения: группу А сравнивают с группой В, группу Б — с группой В и наконец А с Б. Получив достаточно высокое значение t в каком-либо из трех сравнений, сообщают, что «$P < 0{,}05$». Это означает, что вероятность ошибочного заключения о существовании различий не превышает 5%. Но это неверно: вероятность ошибки значительно превышает 5%.

Разберемся подробнее. В исследовании был принят 5% уровень значимости. Значит, вероятность ошибиться при сравнении групп А и В — 5%. Казалось бы, все правильно. Но точно так же мы ошибемся в 5% случаев при сравнении групп Б и В. И наконец, при сравнении групп А и Б ошибка возможна также в 5% случаев. Следовательно, вероятность ошибиться *хотя бы в одном* из трех сравнений составит не 5%, а значительно больше. В общем случае эта вероятность равна

$$P' = 1 - (1 - 0{,}05)^k,$$

где k — число сравнений.

При небольшом числе сравнений можно использовать приближенную формулу

$$P' = 0{,}05k,$$

то есть вероятность ошибиться хотя бы в одном из сравнений примерно равна вероятности ошибиться в одном, помноженной на число сравнений.

Итак, в нашем исследовании вероятность ошибиться хотя бы в одном из сравнений составляет примерно 15%. При сравнении четырех групп число пар и соответственно возможных попарных сравнений равно 6. Поэтому при уровне значимости в каждом из сравнений 0,05 вероятность ошибочно обнаружить различие хотя бы в одном равна уже не 0,05, а примерно $6 \times 0{,}05 = 0{,}30$. И когда исследователь, выявив таким способом «эффективный» препарат, будет говорить про 5% вероятность ошибки, на самом деле эта вероятность равна 30%.

Вернемся на минуту к нашим марсианам. Рассматривая в гл. 2 случайные выборки из населения этой планеты, мы убедились, что у разных выборок из одной совокупности могут быть заметно разные средние значения и стандартные отклонения —

взять хоть три случайные выборки на рис. 2.6. Представим себе, что это — результаты исследования влияния гормонов человека на рост марсиан. Одной группе дали тестостерон, другой — эстрадиол, а третьей — плацебо. Как известно, гормоны человека не оказывают на марсиан никакого действия, поэтому три экспериментальные группы — это просто три случайные выборки из одной совокупности, как мы это и знали с самого начала. Что хорошо известно нам, то неизвестно исследователям. На рис. 4.6 результаты исследования представлены в виде, принятом в медицинских публикациях. Столбиками изображены выборочные средние. Вертикальные черточки задают интервалы в плюс-минус одну стандартную ошибку среднего. Засучив рукава, наши исследователи приступают к попарному сравнению групп с помощью критерия Стьюдента и получают такие значения t: плацебо—тестостерон — 2,39; плацебо—эстрадиол — 0,93 и тестостерон—эстрадиол — 1,34. Так как в каждом сравнении участвуют 2 группы по 10 марсиан в каждой, число степеней свободы равно $2(10-1)=18$. По таблице 4.1 находим, что при 5% уровне значимости критическое значение t равно 2,101. Таким образом, пришлось бы заключить, что марсиане, получавшие тестостерон, стали меньше ростом, чем марсиане, получавшие плацебо, в то время как эстрадиол по влиянию на рост существенно не отличается от плацебо, а тестостерон от эстрадиола.

Задумайтесь над этим результатом. Что в нем не так?

Если тестостерон дал результаты, не отличающиеся от эстрадиола, а эстрадиол действует неотличимо от плацебо, то как тестостерон оказался отличным от плацебо? Столь странный вывод обычно не смущает исследователей, а лишь вдохновляет их на создание изощренного «Обсуждения».

Дисперсионный анализ приведенных данных дает значение $F=2{,}74$. Число степеней свободы $\nu_{меж} = m-1=3-1=2$ и $\nu_{вну} = m(n-1)=3(10-1)=27$. Критическое значение F для 5% уровня значимости равно 3,35, то есть превышает полученное нами. Итак, дисперсионный анализ говорит об отсутствии различий между группами.

В заключение приведем три правила.

- Критерий Стьюдента может быть использован для проверки гипотезы о различии средних только для двух групп.

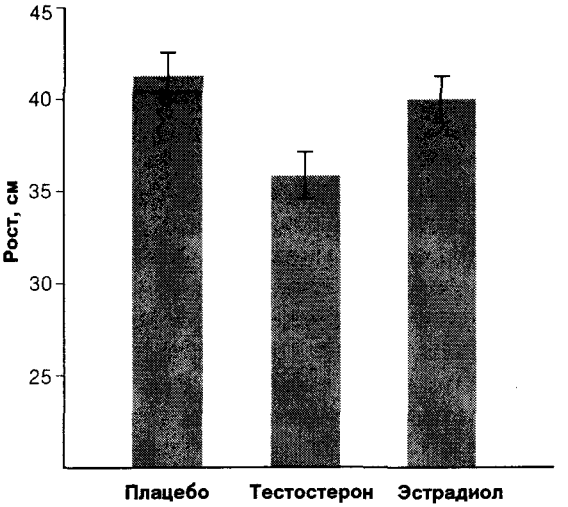


Рис. 4.6. Влияние гормонов человека на рост марсиан. Именно в таком виде результаты исследования увидели бы свет в каком-нибудь медицинском журнале. Высота столбиков соответствует средним, вертикальная черта на верхушке у каждого столбика соответствует интервалу плюс-минус одна стандартная ошибка среднего (а не стандартное отклонение).

- Если схема эксперимента предполагает большее число групп, воспользуйтесь дисперсионным анализом.
- Если критерий Стьюдента был использован для проверки различий между несколькими группами, то истинный уровень значимости можно получить, умножив уровень значимости, приводимый авторами, на число возможных сравнений.

КРИТЕРИЙ СТЬЮДЕНТА ДЛЯ МНОЖЕСТВЕННЫХ СРАВНЕНИЙ

Только что мы познакомились со злостным вредителем научных исследований — эффектом множественных сравнений. Он состоит в том, что при многократном применении критерия вероятность ошибочно найти различия там, где их нет, возрастает.

Если исследуемых групп больше двух, то следует воспользоваться дисперсионным анализом. Однако дисперсионный ана-

лиз позволяет проверить лишь гипотезу о равенстве *всех* средних. Но, если гипотеза не подтверждается, нельзя узнать, какая именно группа отличается от других.

Это позволяют сделать *методы множественного сравнения*. Все они основаны на критерии Стьюдента, но учитывают, что сравнивается более одной пары средних. Сразу поясним, когда, на наш взгляд, следует использовать эти методы. Наш подход состоит в том, чтобы в первую очередь с помощью дисперсионного анализа проверить нулевую гипотезу о равенстве всех средних, а уже затем, *если нулевая гипотеза отвергнута*, выделить среди них отличные от остальных, используя для этого методы множественного сравнения*. Простейший из методов множественного сравнения — введение *поправки Бонферрони*.

Как было показано в предыдущем разделе, при трехкратном применении критерия Стьюдента с 5% уровнем значимости вероятность обнаружить различия там, где их нет, составляет не 5%, а почти $3 \times 5 = 15\%$. Этот результат является частным случаем *неравенства Бонферрони:* если k раз применить критерий с уровнем значимости α, то вероятность хотя бы в одном случае найти различие там, где его нет, не превышает произведения k на α. Неравенство Бонферрони выглядит так:

$$\alpha' < k\alpha,$$

где α' — вероятность хотя бы один раз ошибочно выявить различия.

Можно сказать, что α', собственно, и является истинным уровнем значимости многократно примененного критерия. Из неравенства Бонферрони следует, что если мы хотим обеспечить вероятность ошибки α', то в каждом из сравнений мы должны принять уровень значимости α'/k — это и есть поправка Бонферрони. Например, при трехкратном сравнении уровень значимости должен быть $0{,}05/3 = 1{,}7\%$.

---

* Некоторые авторы считают этап дисперсионного анализа излишним и предлагают сразу применить методы множественных сравнений. Этот подход изложен в B. W. Broun, Jr., M. Hollander. Statistics: a biomedical introduction. Wiley, New York, 1977, chap. 10. Analysis of K-samples problems.

Поправка Бонферрони хорошо работает, если число сравнений невелико. Если оно превышает 8, метод становится слишком «строгим» и даже весьма большие различия приходится признавать статистически незначимыми*. Существуют не столь жесткие методы множественного сравнения, например критерий Ньюмена–Кейлса (его мы рассмотрим в следующем разделе). Все методы множественного сравнения схожи с поправкой Бонферрони в том, что, будучи модификацией критерия Стьюдента, учитывают многократность сравнений.

Один из способов смягчить строгость поправки Бонферрони состоит в том, чтобы увеличить число степеней свободы, воспользовавшись знакомой из дисперсионного анализа внутригрупповой оценкой дисперсии. Вспомним, что

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}},$$

где s^2 — объединенная оценка дисперсии совокупности.

Используя в качестве такой оценки внутригрупповую дисперсию $s^2_{\text{вну}}$ (гл. 3), получим:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s^2_{\text{вну}}}{n_1} + \frac{s^2_{\text{вну}}}{n_2}}}.$$

Если объемы выборок одинаковы, то

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{2s^2_{\text{вну}}}{n}}}.$$

Число степеней свободы $\nu = m(n-1)$. Если число групп m больше 2, то число степеней свободы при таком расчете будет

---

* Способность критерия выявлять различия называется чувствительностью, она обсуждается в гл. 6.

больше $2(n-1)$, благодаря чему критическое значение t уменьшится.

Бег и менструации. Продолжение анализа

В предыдущей главе мы выяснили, что различия в ежегодном числе менструальных циклов в группах спортсменок, физкультурниц и в контрольной группе статистически значимы. Однако осталось неясным, отличаются ли от контрольной группы и спортсменки, и физкультурницы — или только спортсменки? Отличаются ли спортсменки от физкультурниц? Способа определить межгрупповые различия у нас не было. Теперь, используя критерий Стьюдента с поправкой Бонферрони, мы можем попарно сравнить все три группы.

Внутригрупповая оценка дисперсии $s^2_{\text{вну}} = 3{,}95$. Число групп $m = 3$, численность каждой группы $n = 26$. Следовательно, число степеней свободы $\nu = m(n-1) = 3(26-1) = 75$. (Если бы мы оценивали дисперсию по двум группам, число степеней свободы было бы $2(n-1) = 2(26-1) = 50$.) Произведем попарное сравнение трех групп.

При сравнении контрольной группы и группы физкультурниц имеем:

$$t = \frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\frac{2s^2_{\text{вну}}}{n}}} = \frac{10{,}1 - 11{,}5}{\sqrt{\frac{2 \times 3{,}95}{26}}} = -2{,}54,$$

при сравнении контрольной группы и группы спортсменок:

$$t = \frac{\bar{X}_3 - \bar{X}_1}{\sqrt{\frac{2s^2_{\text{вну}}}{n}}} = \frac{9{,}1 - 11{,}5}{\sqrt{\frac{2 \times 3{,}95}{26}}} = -4{,}35,$$

и при сравнении группы физкультурниц и группы спортсменок:

$$t = \frac{\bar{X}_2 - \bar{X}_3}{\sqrt{\frac{2s^2_{\text{вну}}}{n}}} = \frac{10{,}1 - 9{,}1}{\sqrt{\frac{2 \times 3{,}95}{26}}} = 1{,}81.$$

Мы провели 3 сравнения, поэтому уровень значимости в каж-

дом должен быть 0,05/3, то есть примерно 0,017. По таблице 4.1 находим*, что при 75 степенях свободы критическое значение составляет примерно 2,45.

Таким образом, мы можем заключить, что и у спортсменок, и у физкультурниц частота менструаций ниже, чем в контрольной группе, при этом у спортсменок и физкультурниц она не отличается.

КРИТЕРИЙ НЬЮМЕНА–КЕЙЛСА**

При большом числе сравнений поправка Бонферрони делает критерий Стьюдента излишне жестким. Более изощренный *критерий Ньюмена–Кейлса* дает более точную оценку вероятности α'; чувствительность его выше, чем критерия Стьюдента с поправкой Бонферрони.

Сначала нужно с помощью дисперсионного анализа проверить нулевую гипотезу о равенстве всех средних. Если она отвергается, все средние упорядочивают по возрастанию и сравнивают попарно, каждый раз вычисляя значение критерия Ньюмена–Кейлса:

$$q = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{s^2_{\text{ВНУ}}}{2}\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}},$$

---

* Собственно говоря, значения для $\alpha = 0{,}017$ в таблице нет. В таких случаях можно либо использовать ближайшее меньшее значение α (в нашем примере это 0,01), либо приблизительно рассчитать нужное критическое значение по соседним. Если нужное нам значение $\alpha_н$ находится между α_1 и α_2, которым соответствуют критические значения t_1 и t_2, то

$$t_н = t_1 + (t_2 - t_1)\frac{(\alpha_н - \alpha_1)}{(\alpha_2 - \alpha_1)},$$

где $t_н$ – критическое значение для уровня значимости $\alpha_н$.

** Этот раздел важен для тех, кто использует нашу книгу как руководство по анализу данных. Его можно опустить без ущерба для понимания остального материала.

где $\overline{X}_A$ и $\overline{X}_B$ — сравниваемые средние, $s^2_{\text{вну}}$ — внутригрупповая дисперсия, а n_A и n_B — численность групп.

Вычисленное значение q сравнивается с критическим значением (табл. 4.3). Критическое значение зависит от α' (вероятность ошибочно обнаружить различия хотя бы в одной из всех сравниваемых пар, то есть истинный уровень значимости), числа степеней свободы $\nu = N - m$ (где N — сумма численностей всех групп, m — число групп) и величины l, которая называется интервалом сравнения. Интервал сравнения определятся так. Если сравниваются средние, стоящие соответственно на j-м и i-м месте в упорядоченном ряду, то интервал сравнения $l = j - i + 1$. Например, при сравнении 4-го и 1-го членов этого ряда $l = 4 - 1 + 1 = 4$, при сравнении 2-го и 1-го $l = 2 - 1 + 1 = 2$.

Результат применения критерия Ньюмена–Кейлса зависит от очередности сравнений, поэтому их следует проводить в определенном порядке. Этот порядок задается двумя правилами.

1. Если мы расположили средние от меньшего к большему (от 1 до m), то сначала нужно сравнить наибольшее с наименьшим, то есть m-е с 1-м, затем m-е со 2-м, 3-м и так далее, вплоть до $m-1$-го. Затем предпоследнее ($m-1$-е) тем же порядком сравниваем с 1-м, 2-м и так далее до $m-2$-го. Продолжаем эти «стягивающие сравнения», пока не переберем все пары. Например, в случае 4 групп порядок сравнений такой: 4–1, 4–2, 4–3, 3–1, 3–2, 2–1.

2. Перебирать все пары, впрочем, приходится не всегда. Если какие-либо средние не различаются, то все средние, лежащие между ними, тоже не различаются. Например, если не выявлено различий между 3-м и 1-м средним, не нужно сравнивать ни 3-е со 2-м, ни 2-е с 1-м.

Бег и менструации. Продолжение анализа

Воспользуемся критерием Ньюмена—Кейлса для анализа связи частоты менструаций с занятиями физкультурой и спортом. Среднегодовое число менструаций в контрольной группе составило 11,5, у физкультурниц — 10,1 и у спортсменок — 9,1. Упорядочим эти средние по возрастанию: 9,1; 10,1; 11,5 (спортсменки, физкультурницы, контроль) и обозначим их $\overline{X}_1$, $\overline{X}_2$, $\overline{X}_3$ соответственно. Оценка внутригрупповой дисперсии $s^2_{\text{вну}} = 3{,}95$, число степе-

Таблица 4.3А. Критические значения q для $\alpha' = 0{,}05$

	Интервал сравнения l								
ν	2	3	4	5	6	7	8	9	10
1	17,97	26,98	32,82	37,08	40,41	43,12	45,40	47,36	49,07
2	6,085	8,331	9,798	10,88	11,74	12,44	13,03	13,54	13,99
3	4,501	5,910	6,825	7,502	8,037	8,478	8,853	9,177	9,462
4	3,927	5,040	5,757	6,287	6,707	7,053	7,347	7,602	7,826
5	3,635	4,602	5,218	5,673	6,033	6,330	6,582	6,802	6,995
6	3,461	4,339	4,896	5,305	5,628	5,895	6,122	6,319	6,493
7	3,344	4,165	4,681	5,060	5,359	5,606	5,815	5,998	6,158
8	3,261	4,041	4,529	4,886	5,167	5,399	5,597	5,767	5,918
9	3,199	3,949	4,415	4,756	5,024	5,244	5,432	5,595	5,739
10	3,151	3,877	4,327	4,654	4,912	5,124	5,305	5,461	5,599
11	3,113	3,820	4,256	4,574	4,823	5,028	5,202	5,353	5,487
12	3,082	3,773	4,199	4,508	4,751	4,950	5,119	5,265	5,395
13	3,055	3,735	4,151	4,453	4,690	4,885	5,049	5,192	5,318
14	3,033	3,702	4,111	4,407	4,639	4,829	4,990	5,131	5,254
15	3,014	3,674	4,076	4,367	4,595	4,782	4,940	5,077	5,198
16	2,998	3,649	4,046	4,333	4,557	4,741	4,897	5,031	5,150
17	2,984	3,628	4,020	4,303	4,524	4,705	4,858	4,991	5,108
18	2,971	3,609	3,997	4,277	4,495	4,673	4,824	4,956	5,071
19	2,960	3,593	3,977	4,253	4,469	4,645	4,794	4,924	5,038
20	2,950	3,578	3,958	4,232	4,445	4,620	4,768	4,896	5,008
24	2,919	3,532	3,901	4,166	4,373	4,541	4,684	4,807	4,915
30	2,888	3,486	3,845	4,102	4,302	4,464	4,602	4,720	4,824
40	2,858	3,442	3,791	4,039	4,232	4,389	4,521	4,635	4,735
60	2,829	3,399	3,737	3,977	4,163	4,314	4,441	4,550	4,646
120	2,800	3,356	3,685	3,917	4,096	4,241	4,363	4,468	4,560
∞	2,772	3,314	3,633	3,858	4,030	4,170	4,286	4,387	4,474

ней свободы $\nu = 75$, численность каждой группы 26 человек. Теперь мы можем воспользоваться критерием Ньюмена—Кейлса.

Сравним $\bar{X}_3$ и $\bar{X}_1$. Имеем:

$$q = \frac{\bar{X}_3 - \bar{X}_1}{\sqrt{\frac{s^2_{\text{вну}}}{2}\left(\frac{1}{n_3} + \frac{1}{n_1}\right)}} = \frac{11{,}5 - 9{,}1}{\sqrt{\frac{3{,}95}{2}\left(\frac{1}{26} + \frac{1}{26}\right)}} = 6{,}157.$$

Интервал сравнения в данном случае: $l = 3 - 1 + 1 = 3$. По таблице 4.3А находим, что для уровня значимости $\alpha' = 0{,}05$, числа степеней свободы $\nu = 75$ и интервала сравнения $l = 3$ критическое

Таблица 4.3Б. Критические значения q для $\alpha' = 0{,}01$

	Интервал сравнения l								
ν	2	3	4	5	6	7	8	9	10
1	90,03	135,0	164,3	185,6	202,2	215,8	227,2	237,0	245,6
2	14,04	19,02	22,29	24,72	26,63	28,20	29,53	30,68	31,69
3	8,261	10,62	12,17	13,33	14,24	15,00	15,64	16,20	16,69
4	6,512	8,120	9,173	9,958	10,58	11,10	11,55	11,93	12,27
5	5,702	6,976	7,804	8,421	8,913	9,321	9,669	9,972	10,24
6	5,243	6,331	7,033	7,556	7,973	8,318	8,613	8,869	9,097
7	4,949	5,919	6,543	7,005	7,373	7,679	7,939	8,166	8,368
8	4,746	5,635	6,204	6,625	6,960	7,237	7,474	7,681	7,863
9	4,596	5,428	5,957	6,348	6,658	6,915	7,134	7,325	7,495
10	4,482	5,270	5,769	6,136	6,428	6,669	6,875	7,055	7,213
11	4,392	5,146	5,621	5,970	6,247	6,476	6,672	6,842	6,992
12	4,320	5,046	5,502	5,836	6,101	6,321	6,507	6,670	6,814
13	4,260	4,964	5,404	5,727	5,981	6,192	6,372	6,528	6,667
14	4,210	4,895	5,322	5,634	5,881	6,085	6,258	6,409	6,543
15	4,168	4,836	5,252	5,556	5,796	5,994	6,162	6,309	6,439
16	4,131	4,786	5,192	5,489	5,722	5,915	6,079	6,222	6,349
17	4,099	4,742	5,140	5,430	5,659	5,847	6,007	6,147	6,270
18	4,071	4,703	5,094	5,379	5,603	5,788	5,944	6,081	6,201
19	4,046	4,670	5,054	5,334	5,554	5,735	5,889	6,022	6,141
20	4,024	4,639	5,018	5,294	5,510	5,688	5,839	5,970	6,087
24	3,956	4,546	4,907	5,168	5,374	5,542	5,685	5,809	5,919
30	3,889	4,455	4,799	5,048	5,242	5,401	5,536	5,653	5,756
40	3,825	4,367	4,696	4,931	5,114	5,265	5,392	5,502	5,559
60	3,762	4,282	4,595	4,818	4,991	5,133	5,253	5,356	5,447
120	3,702	4,200	4,497	4,709	4,872	5,005	5,118	5,214	5,299
∞	3,643	4,120	4,403	4,603	4,757	4,882	4,987	5,078	5,157

H. L. Harter. Order statistics and their use in testing and estimation. Vol. 1: Tests based on range and studentized range of samples from a normal population. U.S. Government Printing Office, Washington, D.C., 1970.

значение q равно 3,385, то есть меньше, чем получилось у нас. Следовательно, различие статистически значимо.

Теперь сравним $\overline{X}_3$ и $\overline{X}_2$.

$$q = \frac{\overline{X}_3 - \overline{X}_2}{\sqrt{\frac{s^2_{\text{вну}}}{2}\left(\frac{1}{n_3} + \frac{1}{n_2}\right)}} = \frac{11{,}5 - 10{,}1}{\sqrt{\frac{3{,}95}{2}\left(\frac{1}{26} + \frac{1}{26}\right)}} = 3{,}592.$$

Величины α' и ν те же, что и раньше, но теперь $l = 3-2+1 = 2$. По таблице 4.3А находим критическое значение $q = 2{,}822$. Полученное нами значение снова превосходит критическое. Различие статистически значимо.

Для $\bar{X}_2$ и $\bar{X}_1$ имеем:

$$q = \frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\frac{s^2_{\text{вну}}}{2}\left(\frac{1}{n_2}+\frac{1}{n_1}\right)}} = \frac{10{,}1-9{,}1}{\sqrt{\frac{3{,}95}{2}\left(\frac{1}{26}+\frac{1}{26}\right)}} = 2{,}566.$$

Величины α', ν и $l = 2-1+1 = 2$ те же, что и в предыдущем сравнении, соответственно то же и критическое значение. Оно больше вычисленного, следовательно, различие статистически не значимо.

В данном случае вывод не отличается от полученного при применении критерия Стьюдента с поправкой Бонферрони.

КРИТЕРИЙ ТЬЮКИ

Критерий Тьюки совпадает с критерием Ньюмена–Кейлса во всем, кроме способа определения критического значения. В критерии Ньюмена–Кейлса критическое значение q зависит от интервала сравнения l. В критерии Тьюки при всех сравнениях вместо l берут число групп m, таким образом, критическое значение q все время одно и то же. Критерий Ньюмена–Кейлса был разработан как усовершенствование критерия Тьюки.

Применяя критерий Тьюки к только что рассмотренной задаче о влиянии бега на частоту менструаций, нужно было бы приравнять l к числу групп $m = 3$. Соответствующее критическое значение равно 3,385 и *неизменно* при всех сравнениях. В нашем примере при двух последних сравнениях критические значения по Тьюки будут больше, чем по Ньюмену–Кейлсу. Однако *в данном случае* результат применения обоих критериев один и тот же. Разумеется, так будет не всегда. Поскольку в критерии Тьюки при всех сравнениях используется максимальное критическое значение q, различия будут выявляться реже, чем при использовании критерия Ньюмена–Кейлса.

Критерий Тьюки слишком жесток и отвергает существование различий чаще, чем нужно, а критерий Ньюмена–Кейлса, напротив, слишком мягок. В общем, выбор критерия определяется скорее психологическим фактором: чего больше боится исследователь: найти отличия там, где их нет, или пропустить их там, где они есть. Автор предпочитает критерий Ньюмена–Кейлса.

МНОЖЕСТВЕННЫЕ СРАВНЕНИЯ С КОНТРОЛЬНОЙ ГРУППОЙ*

Иногда задача заключается в том, чтобы сравнить несколько групп с единственной — контрольной. Конечно, можно было бы использовать любой из описанных методов множественного сравнения (критерий Стьюдента с поправкой Бонферрони, Ньюмена—Кейлса или Тьюки): попарно сравнить все группы, а затем отобрать те сравнения, в которых участвовала контрольная группа. Однако в любом случае (особенно при применении поправки Бонферрони) из-за большого числа лишних сравнений критическое значение окажется неоправданно высоким. Иными словами, мы слишком часто будем пропускать реально существующие различия. Преодолеть эту трудность позволяют специальные методы сравнения, из которых мы разберем два. Это еще одна модификация критерия Стьюдента с поправкой Бонферрони и критерий Даннета. Как и другие методы множественного сравнения, их следует применять только после того, как с помощью дисперсионного анализа отвергнута нулевая гипотеза о равенстве всех средних.

Поправка Бонферрони

Применить поправку Бонферрони к сравнению нескольких групп с одной контрольной очень просто. Ход вычислений такой же, что и при применении поправки Бонферрони в общем случае. Надо только учесть, что число сравнений k составляет теперь

---

* Этот материал важен для тех, кто использует нашу книгу как руководство для анализа данных. Во вводном курсе этот раздел можно опустить.

Таблица 4.4А. Критические значения q' для $\alpha' = 0{,}05$

	Интервал сравнения l													
ν	2	3	4	5	6	7	8	9	10	11	12	13	16	21
5	2,57	3,03	3,29	3,48	3,62	3,73	3,82	3,90	3,97	4,03	4,09	4,14	4,26	4,42
6	2,45	2,86	3,10	3,26	3,39	3,49	3,57	3,64	3,71	3,76	3,81	3,86	3,97	4,11
7	2,36	2,75	2,97	3,12	3,24	3,33	3,41	3,47	3,53	3,58	3,63	3,67	3,78	3,91
8	2,31	2,67	2,88	3,02	3,13	3,22	3,29	3,35	3,41	3,46	3,50	3,54	3,64	3,76
9	2,26	2,61	2,81	2,95	3,05	3,14	3,20	3,26	3,32	3,36	3,40	3,44	3,53	3,65
10	2,23	2,57	2,76	2,89	2,99	3,07	3,14	3,19	3,24	3,29	3,33	3,36	3,45	3,57
11	2,20	2,53	2,72	2,84	2,94	3,02	3,08	3,14	3,19	3,23	3,27	3,30	3,39	3,50
12	2,18	2,50	2,68	2,81	2,90	2,98	3,04	3,09	3,14	3,18	3,22	3,25	3,34	3,45
13	2,16	2,48	2,65	2,78	2,87	2,94	3,00	3,06	3,10	3,14	3,18	3,21	3,29	3,40
14	2,14	2,46	2,63	2,75	2,84	2,91	2,97	3,02	3,07	3,11	3,14	3,18	3,26	3,36
15	2,13	2,44	2,61	2,73	2,82	2,89	2,95	3,00	3,04	3,08	3,12	3,15	3,23	3,33
16	2,12	2,42	2,59	2,71	2,80	2,87	2,92	2,97	3,02	3,06	3,09	3,12	3,20	3,30
17	2,11	2,41	2,58	2,69	2,78	2,85	2,90	2,95	3,00	3,03	3,07	3,10	3,18	3,27
18	2,10	2,40	2,56	2,68	2,76	2,83	2,89	2,94	2,98	3,01	3,05	3,08	3,16	3,25
19	2,09	2,39	2,55	2,66	2,75	2,81	2,87	2,92	2,96	3,00	3,03	3,06	3,14	3,23
20	2,09	2,38	2,54	2,65	2,73	2,80	2,86	2,90	2,95	2,98	3,02	3,05	3,12	3,22
24	2,06	2,35	2,51	2,61	2,70	2,76	2,81	2,86	2,90	2,94	2,97	3,00	3,07	3,16
30	2,04	2,32	2,47	2,58	2,66	2,72	2,77	2,82	2,86	2,89	2,92	2,95	3,02	3,11
40	2,02	2,29	2,44	2,54	2,62	2,68	2,73	2,77	2,81	2,85	2,87	2,90	2,97	3,06
60	2,00	2,27	2,41	2,51	2,58	2,64	2,69	2,73	2,77	2,80	2,83	2,86	2,92	3,00
120	1,98	2,24	2,38	2,47	2,55	2,60	2,65	2,69	2,73	2,76	2,79	2,81	2,87	2,95
∞	1,96	2,21	2,35	2,44	2,51	2,57	2,61	2,65	2,69	2,72	2,74	2,77	2,83	2,91

Таблица 4.4Б. Критические значения q' для $\alpha' = 0{,}01$

	Интервал сравнения l													
ν	2	3	4	5	6	7	8	9	10	11	12	13	16	21
5	4,03	4,63	4,98	5,22	5,41	5,56	5,69	5,80	5,89	5,98	6,05	6,12	6,30	6,52
6	3,71	4,21	4,51	4,71	4,87	5,00	5,10	5,20	5,28	5,35	5,41	5,47	5,62	5,81
7	3,50	3,95	4,21	4,39	4,53	4,64	4,74	4,82	4,89	4,95	5,01	5,06	5,19	5,36
8	3,36	3,77	4,00	4,17	4,29	4,40	4,48	4,56	4,62	4,68	4,73	4,78	4,90	5,05
9	3,25	3,63	3,85	4,01	4,12	4,22	4,30	4,37	4,43	4,48	4,53	4,57	4,68	4,82
10	3,17	3,53	3,74	3,88	3,99	4,08	4,16	4,22	4,28	4,33	4,37	4,42	4,52	4,65
11	3,11	3,45	3,65	3,79	3,89	3,98	4,05	4,11	4,16	4,21	4,25	4,29	4,30	4,52
12	3,05	3,39	3,58	3,71	3,81	3,89	3,96	4,02	4,07	4,12	4,16	4,19	4,29	4,41
13	3,01	3,33	3,52	3,65	3,74	3,82	3,89	3,94	3,99	4,04	4,08	4,11	4,20	4,32
14	2,98	3,29	3,47	3,59	3,69	3,76	3,83	3,88	3,93	3,97	4,01	4,05	4,13	4,24
15	2,95	3,25	3,43	3,55	3,64	3,71	3,78	3,83	3,88	3,92	3,95	3,99	4,07	4,18
16	2,92	3,22	3,39	3,51	3,60	3,67	3,73	3,78	3,83	3,87	3,91	3,94	4,02	4,13
17	2,90	3,19	3,36	3,47	3,56	3,63	3,69	3,74	3,79	3,83	3,86	3,90	3,98	4,08
18	2,88	3,17	3,33	3,44	3,53	3,60	3,66	3,71	3,75	3,79	3,83	3,86	3,94	4,04
19	2,86	3,15	3,31	3,42	3,50	3,57	3,63	3,68	3,72	3,76	3,79	3,83	3,90	4,00
20	2,85	3,13	3,29	3,40	3,48	3,55	3,60	3,65	3,69	3,73	3,77	3,80	3,87	3,97
24	2,80	3,07	3,22	3,32	3,40	3,47	3,52	3,57	3,61	3,64	3,68	3,70	3,78	3,87
30	2,75	3,01	3,15	3,25	3,33	3,39	3,44	3,49	3,52	3,56	3,59	3,62	3,69	3,78
40	2,70	2,95	3,09	3,19	3,26	3,32	3,37	3,41	3,44	3,48	3,51	3,53	3,60	3,68
60	2,66	2,90	3,03	3,12	3,19	3,25	3,29	3,33	3,37	3,40	3,42	3,45	3,51	3,59
120	2,62	2,85	2,97	3,06	3,12	3,18	3,22	3,26	3,29	3,32	3,35	3,37	3,43	3,51
∞	2,58	2,79	2,92	3,00	3,06	3,11	3,15	3,19	3,22	3,25	3,27	3,29	3,35	2,42

C. W. Dunnett. New tables for multiple comparisons with a control. *Biometrics*, 20:482–491, 1964.

$m-1$, и соответственно рассчитать уровень значимости в каждом из сравнений: $\alpha = \alpha'/k$. Применим этот метод к исследованию частоты менструаций. Сравним спортсменок и физкультурниц с контрольной группой. Число сравнений $k = 2$ (а не 3, как при всех возможных сравнениях). Чтобы полная вероятность ошибочно обнаружить различия не превышала 0,05, при каждом сравнении уровень значимости должен быть 0,05/2 = 0,025 (вместо 0,05/3 = 0,017). Число степеней свободы — 75; критическое значение $t = 2{,}31$ (при всех возможных сравнениях оно бы составило 2,45). Величину t для сравнения физкультурниц и спортсменок с контролем мы уже рассчитывали — 2,54 и 4,35 соответственно. Таким образом, и спортсменки, и физкультурницы статистически значимо отличаются от контрольной группы. В данном случае вывод получился тот же, что и при применении поправки Бонферрони в общем случае. Ясно, однако, что за счет снижения критического уровня t чувствительность метода повышается. Обратите внимание, что в данном случае мы *не делаем никакого заключения* о различии спортсменок и физкультурниц.

Критерий Даннета

Критерий Даннета — это вариант критерия Ньюмена–Кейлса для сравнения нескольких групп с одной контрольной. Он вычисляется как

$$q' = \frac{\bar{X}_{\text{кон}} - \bar{X}_A}{\sqrt{s^2_{\text{вну}}\left(\frac{1}{n_{\text{кон}}} + \frac{1}{n_A}\right)}}.$$

Число сравнений равно числу групп, не считая контрольной, и существенно меньше числа сравнений в исходном критерии Ньюмена–Кейлса. Соответственно, меньше и критические значения (табл. 4.4). Как и в критерии Ньюмена–Кейлса, сначала средние значения для всех групп упорядочиваются, только теперь — по абсолютной величине их отличия от контрольной группы. Затем контрольную группу сравнивают с остальными, начиная с наиболее отличной от контрольной. Если различия с очередной группой не найдены, вычисления прекращают. Параметр l постоянен и равен

числу групп, включая контрольную. Число степеней свободы вычисляют как в критерии Ньюмена–Кейлса: $\nu = N - m$.

Применим критерий Даннета к анализу влияния бега на менструации. Сначала сравним с контрольной наиболее от нее отличную группу спортсменок:

$$q' = \frac{\bar{X}_{\text{кон}} - \bar{X}_1}{\sqrt{s^2_{\text{вну}}\left(\frac{1}{n_{\text{кон}}} + \frac{1}{n_1}\right)}} = \frac{11{,}5 - 9{,}1}{\sqrt{3{,}95\left(\frac{1}{26} + \frac{1}{26}\right)}} = 4{,}35.$$

Общее число средних равно трем, поэтому $l = 3$. Число степеней свободы равно 75. По таблице 4.4 находим критическое значение для уровня значимости 0,05. Оно равно 2,28. Вычисленное значение больше критического. Тем самым, различие между спортсменками и контрольной группой статистически значимо, и сравнения можно продолжать.

Теперь сравним с контрольной группу физкультурниц:

$$q' = \frac{\bar{X}_{\text{кон}} - \bar{X}_2}{\sqrt{s^2_{\text{вну}}\left(\frac{1}{n_{\text{кон}}} + \frac{1}{n_2}\right)}} = \frac{11{,}5 - 10{,}1}{\sqrt{3{,}95\left(\frac{1}{26} + \frac{1}{26}\right)}} = 2{,}54$$

Критическое значение q' по-прежнему равно 2,28. Вычисленное значение больше. Различие между физкультурницами и контрольной группой статистически значимо.

Критерий Даннета, как вариант критерия Ньюмена–Кейлса, более чувствителен, чем критерий Стьюдента с поправкой Бонферрони, особенно при большом числе групп. Если бы групп было больше, мы убедились бы, что критерий Ньюмена–Кейлса обнаруживает те различия, которые упускает критерий Стьюдента с поправкой Бонферрони, завышающей критические значения t.

ЧТО ОЗНАЧАЕТ *P*

Поговорим еще раз о вероятности справедливости нулевой гипотезы P. Понимание смысла P требует понимания логики проверки статистической гипотезы. Например, исследователь хочет

узнать, влияет ли некий препарат на температуру тела. Очевидная схема эксперимента: взять две группы, одной дать препарат, другой плацебо, измерить температуру и вычислить для обеих групп среднюю температуру и стандартное отклонение. Средние температуры вряд ли совпадут, даже если препарат не обладает никаким действием. Поэтому естественен вопрос: сколь вероятно, что наблюдаемое различие случайно?

Для ответа на этот вопрос прежде всего нужно выразить различия одним числом — *критерием значимости*. Со многими из них мы уже встречались — это критерии F, t, q и q'. Значение критерия тем больше, чем больше различия. Если препарат не оказывает действия, то величина критерия будет мала, если оказывает — велика. Но что значит «мала» и что значит «велика»?

Чтобы разграничить «большие» и «малые» значения критерия, строится предположение, что препарат *не оказывает* влияния на температуру. Это так называемая *нулевая гипотеза*. Если нулевая гипотеза верна, то обе группы можно считать просто случайными выборками из одной и той же совокупности. Далее эксперимент мысленно проводится на всех возможных выборках, и для каждой пары вычисляется значение критерия. Чаще всего оно будет небольшим, но какая-то часть выборок даст весьма высокие значения. При этом мы сможем указать такое число (*критическое значение*), выше которого значение критерия оказывается, скажем, в 5% случаев.

Теперь вернемся к препарату и вычислим значение критерия. Если оно превышает критическое значение, то мы можем утверждать следующее: *если бы нулевая гипотеза была справедлива, то вероятность получить наблюдаемые различия была бы меньше 5%.* В принятой системе обозначений это записывается как $P < 0{,}05$. Отсюда мы заключаем, что гипотеза об отсутствии влияния препарата на температуру вряд ли справедлива, то есть различия статистически значимы (при 5% уровне значимости). Разумеется, этот вывод по сути своей носит вероятностный характер. Не исключено, что мы ошибочно признаем неэффективный препарат эффективным, то есть найдем различия там, где их нет. Однако мы можем утверждать, что вероятность подобной ошибки не превышает 5%.

Дадим определение P.

P есть вероятность того, что значение критерия окажется не меньше критического значения при условии справедливости нулевой гипотезы об отсутствии различий между группами.

Определение можно сформулировать и по-другому.

P есть вероятность ошибочно отвергнуть нулевую гипотезу об отсутствии различий.

Упрощая, можно сказать, что *P* — это *вероятность справедливости нулевой гипотезы*. Часто говорят также, что *P* — это вероятность ошибки. В общем, и это верно, однако несколько неточно. Дело в том, что существует два рода ошибок. Ошибка I рода — это ошибочное заключение о существовании различий, которых в действительности нет. Вероятность именно этой ошибки оценивает *P*. Возможна и противоположная ошибка — принять неверную нулевую гипотезу, то есть не найти действительно существующее различие. Это так называемая ошибка II рода. О вероятности этой ошибки *P* ничего не говорит, мы обсудим ее в гл. 6.

ЗАДАЧИ

4.1. Конахан и соавт. определили среднее артериальное давление и общее периферическое сосудистое сопротивление при операциях на открытом сердце с галотановой (9 больных) и морфиновой (16 больных) анестезией. Результаты приведены в табл. 4.2. Можно ли утверждать, что в группах галотановой и морфиновой анестезии эти гемодинамические показатели различаются статистически значимо?

4.2. Кокаин чрезвычайно вреден для сердца, он может вызвать инфаркт миокарда даже у молодых людей без атеросклероза. Кокаин сужает коронарные сосуды, что приводит к уменьшению притока крови к миокарду, кроме того, он ухудшает насосную функцию сердца. Нифедипин (препарат из группы антагонистов кальция) обладает способностью расширять сосуды, его применяют при ишемической болезни сердца. Ш. Хейл и соавт. (S. L. Hale, K. J. Alker, S. H. Rezkalla et al. Nifedipine protects the heart from the acute deleterious effects of cocaine if administered before but not after cocaine. *Circulation*, 83:1437—1443, 1991) предположили, что нифедипин можно использовать и при поражении сердца,

вызванном кокаином. Собакам вводили кокаин, а затем нифедипин либо физиологический раствор. Показателем насосной функции сердца служило среднее артериальное давление. Были получены следующие данные.

Среднее артериальное давление после приема кокаина, мм рт. ст.

Плацебо	Нифедипин
156	73
171	81
133	103
102	88
129	130
150	106
120	106
110	111
112	122
130	108
105	99

Влияет ли нифедипин на среднее артериальное давление после приема кокаина?

4.3. Ш. Хейл и соавт. измеряли также диаметр коронарных артерий после приема нифедипина и плацебо. Позволяют ли приводимые ниже данные утверждать, что нифедипин влияет на диаметр коронарных артерий?

Диаметр коронарной артерии, мм

Плацебо	Нифедипин
2,5	2,5
2,2	1,7
2,6	1,5
2,0	2,5
2,1	1,4
1,8	1,9
2,4	2,3
2,3	2,0
2,7	2,6
2,7	2,3
1,9	2,2

4.4. Решите задачи 3.1 и 3.5, используя критерий Стьюдента.

4.5. В задаче 3.2 приведены данные, собранные Уайтом и Фребом, о проходимости дыхательных путей у некурящих, работающих в помещении, где не курят, у пассивных курильщиков, и у курильщиков, выкуривающих различное число сигарет. Дисперсионный анализ обнаружил, что приведенные данные не согласуются с гипотезой о том, что проходимость дыхательных путей во всех группах одинакова. Выделите группы с одинаковой функцией легких. Что означает полученный результат с точки зрения первоначально поставленного вопроса: влияет ли пассивное курение на функцию легких?

4.6. Используя данные задачи 3.2, оцените статистическую значимость различий некурящих, работающих в помещении, где не курят, со всеми остальными группами. Воспользуйтесь критерием Даннета.

4.7. Решив задачу 3.3, мы пришли к заключению, что уровень холестерина липопротеидов высокой плотности (ХЛПВП) у бегунов-марафонцев, бегунов трусцой и лиц, не занимающихся спортом, неодинаков. Пользуясь критерием Стьюдента с поправкой Бонферрони, сравните эти группы попарно.

4.8. Используя данные задачи 3.3 и рассматривая группу не занимающихся спортом как контрольную, сравните ее с остальными двумя группами. Используйте поправку Бонферрони.

4.9. Пользуясь данными задачи 3.4, найдите группы с близкими показателями антибактериальной защиты.

4.10. По данным задачи 3.7 опишите различия групп. Используйте поправку Бонферрони.

4.11. Решите снова задачу 4.10, пользуясь критерием Ньюмена—Кейлса. Сравните результат с решением задачи 4.10 и объясните различия, если они есть.

4.12. В задаче 3.6 мы установили, что существуют различия в степени опустошенности у медицинских сестер, работающих с больными разной тяжести. В чем заключаются эти различия?

Глава 5

Анализ качественных признаков

Статистические процедуры, с которыми мы познакомились в предыдущих главах, предназначены для анализа *количественных* признаков. Примером таких признаков служат артериальное давление, диурез или продолжительность госпитализации. Единицей их измерения могут быть миллиметры ртутного столба, литры или дни. Над значениями количественных признаков можно производить арифметические действия. Можно, например, сказать, что диурез увеличился вдвое. Кроме того, их можно *упорядочить*, то есть расположить в порядке возрастания или убывания.

Однако очень многие признаки невозможно измерить числом. Например, можно быть либо мужчиной, либо женщиной, либо мертвым, либо живым. Можно быть врачом, юристом, рабочим, и так далее. Здесь мы имеем дело с *качественными признаками*. Эти признаки не связаны между собой никакими арифметическими соотношениями, упорядочить их также нельзя. Единственный способ описания качественных признаков состоит в том, чтобы подсчитать *число* объектов, имеющих одно и

то же значение. Кроме того, можно подсчитать, какая *доля* от общего числа объектов приходится на то или иное значение.

Существует еще один вид признаков. Это *порядковые* признаки. Их можно упорядочить, но производить над ними арифметические действия нельзя. Пример порядкового признака — состояние больного: тяжелое, средней тяжести, удовлетворительное. С такими признаками мы познакомимся в гл. 8 и 10, а сейчас продолжим обсуждение работы Т. Конахана и соавт. по сравнению галотановой и морфиновой анестезии, начатое в гл. 3.

Мы уже знаем, что галотан и морфин по-разному влияли на артериальное давление и что это различие статистически значимо. Однако для клинициста важнее знать, наблюдалось ли различие в операционной летальности? Из 61 больного, оперированного под галотановой анестезией, умерли 8, то есть 13,1%. При использовании морфина умерли 10 из 67, то есть 14,9%. (В гл. 4 мы для простоты считали размеры обеих групп одинаковыми, теперь используются реальные данные.) Летальность при использовании галотана оказалась примерно на 2% ниже, чем при использовании морфина. Можно ли считать, что морфин опаснее галотана, или такой результат мог быть результатом случайности?

Чтобы ответить на этот вопрос, нам сначала нужно найти способ оценить точность, с которой доли, вычисленные по *выборкам*, соответствуют долям во всей *совокупности*. Однако прежде нам нужно понять, каким должно быть описание самой совокупности. Здесь нам пригодятся уже несколько подзабытые марсиане.

НОВОСТИ С МАРСА

В гл. 2 мы побывали на Марсе, где измерили всех его обитателей. Хотя ранее мы не говорили об этом, но больше всего нас поразило различие в пигментации марсиан: 50 марсиан были розового, а остальные 150 — зеленого цвета (рис. 5.1).

Как описать совокупность марсиан по этому признаку? Ясно, что нужно указать *долю*, которую составляют марсиане каждого цвета во всей совокупности марсиан. В нашем случае доля розовых марсиан $p_{роз} = 50/200 = 0{,}25$ и зеленых $p_{зел} = 150/200 = 0{,}75$.

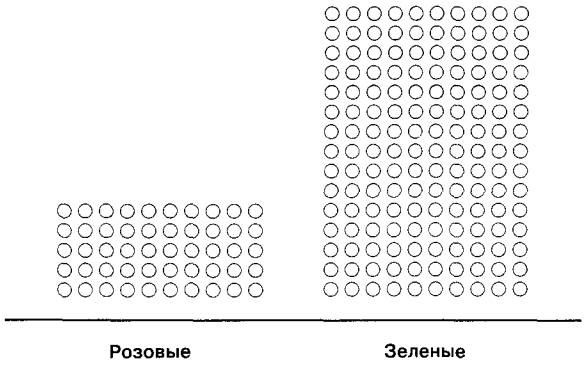


Рис. 5.1. Из 200 марсиан 150 имеют зеленую окраску, остальные 50 розовые. Если наугад извлечь марсианина, то вероятность, что он окажется розовым, составляет 50/200 = 0,25, то есть 25%.

Поскольку марсиане бывают только розовые и зеленые, справедливо тождество $p_{роз} + p_{зел} = 1$. Или, что то же самое, $p_{зел} = 1 - p_{роз}$. То есть, зная $p_{роз}$, мы легко определим и $p_{зел}$. Таким образом, для характеристики совокупности, которая состоит из двух классов, достаточно указать численность одного из них: если доля одного класса во всей совокупности равна p, то доля другого равна $1 - p$. Заметим, что $p_{роз}$ есть еще и *вероятность* того, что случайно выбранный марсианин окажется розовым.

Покажем, что доля p в некотором смысле аналогична среднему μ по совокупности. Введем числовой признак X, который принимает только два значения: 1 для розового и 0 для зеленого. Среднее значение признака X равно

$$\mu = \frac{\Sigma X}{N} = \frac{1+1+\ldots+1+0+0+\ldots+0}{200} =$$
$$= \frac{50\times 1+150\times 0}{200} = \frac{50}{200} = 0{,}25.$$

Как видим, полученное значение совпадает с долей розовых марсиан.

Повторим это рассуждение для общего случая. Пусть имеется совокупность из N членов. При этом M членов обладают каким-то качественным признаком, которого нет у остальных

$N - M$ членов. Введем числовой признак X: у членов совокупности, обладающих качественным признаком, он будет равен 1, а у членов, не обладающих этим признаком, он будет равен 0. Тогда среднее значение X равно

$$\mu = \frac{\Sigma X}{N} = \frac{M \times 1 + (N - M) \times 0}{N} = \frac{M}{N} = p,$$

то есть доле членов совокупности, обладающих качественным признаком.

Используя такой подход, легко рассчитать и показатель разброса — стандартное отклонение. Не совсем ясно, однако, что понимать под разбросом, если значений признака всего два — 0 и 1. На рис. 5.2 мы изобразили три совокупности по 200 членов в каждой. В первой из них (5.2А) все члены принадлежат к одному классу. Разброс равен нулю. На рис. 5.2Б разброс уже имеется, но он невелик. На рис. 5.2В совокупность делится на два равные класса. В этом случае разброс максимален.

Итак, найдем стандартное отклонение. По определению оно равно

$$\sigma = \sqrt{\frac{\Sigma (X - \mu)^2}{N}},$$

где для M членов совокупности значение $X = 1$, а для остальных $N - M$ членов $X = 0$. Величина $\mu = p$. Таким образом,

$$\sigma = \sqrt{\frac{(1-p)^2 + \ldots + (1-p)^2 + (0-p)^2 + \ldots + (0-p)^2}{N}} =$$

$$= \sqrt{\frac{M(1-p)^2 + (N-M)p^2}{N}} = \sqrt{\frac{M}{N}(1-p)^2 + \left(1 - \frac{M}{N}\right)p^2}.$$

Но так как $M/N = p$, то

$$\sigma = \sqrt{p(1-p)^2 + (1-p)p^2} = \sqrt{[p(1-p) + p^2](1-p)},$$

или, после преобразования,

$$\sigma = \sqrt{p(1-p)}.$$

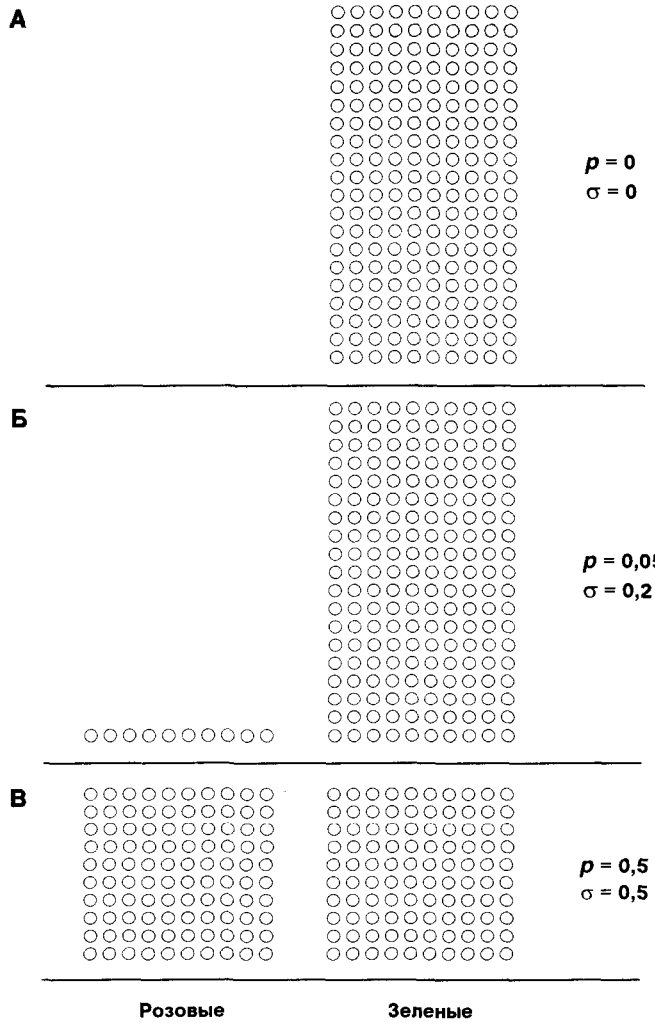


Рис. 5.2. *Что такое разброс данных, если значений признака всего два?* Возможно, это станет яснее, если вспомнить, что разброс — это отсутствие единства. Рассмотрим три совокупности из 200 марсиан. **А.** Все марсиане зеленые. Царит полное единство, разброс отсутствует, $\sigma = 0$. **Б.** Среди стройных рядов зеленых марсиан появилось 10 розовых. Единство немного нарушено, появился некоторый разброс, $\sigma = 0{,}2$. **В.** От единства марсиан не осталось и следа: они разделились поровну на зеленых и розовых. Разброс максимален, $\sigma = 0{,}5$.

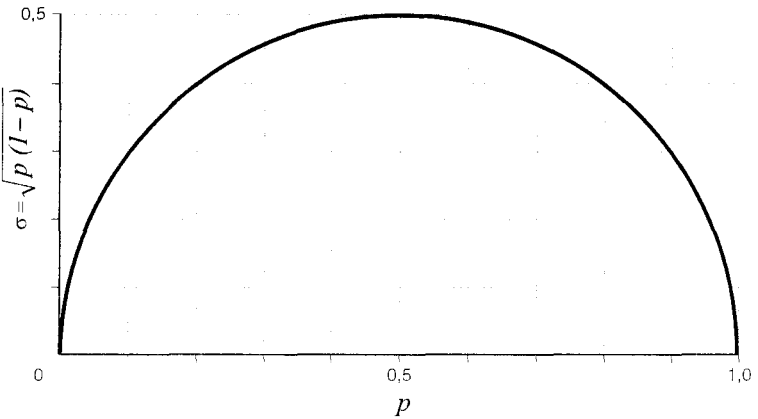


Рис. 5.3. Стандартное отклонение доли σ полностью определяется самой этой долей p. Когда доля равна 0 или 1, разброс отсутствует и σ = 0. Когда p = 0,5, разброс максимален, σ = 0,5.

Найденное стандартное отклонение σ полностью определяется величиной p. Этим оно принципиально отличается от стандартного отклонения для нормального распределения, которое не зависит от μ. На рис. 5.3 показана зависимость σ от p. Она вполне согласуется с теми впечатлениями, которые возникают при рассмотрении рис. 5.2: стандартное отклонение достигает максимума при $p = 0.5$ и равно 0, когда p равно 0 или 1.

Зная стандартное отклонение σ, можно найти стандартную ошибку для выборочной оценки p. Посмотрим, как это делается.

ТОЧНОСТЬ ОЦЕНКИ ДОЛЕЙ

Если бы в наших руках были данные по всем членам совокупности, то не было бы никаких проблем, связанных с точностью оценок. Однако нам всегда приходится довольствоваться ограниченной выборкой. Поэтому возникает вопрос, насколько точно доли в выборке соответствуют долям в совокупности. Проделаем мысленный эксперимент наподобие того, который мы провели в гл. 2, когда рассматривали, насколько хорошей оценкой среднего по совокупности является выборочное среднее.

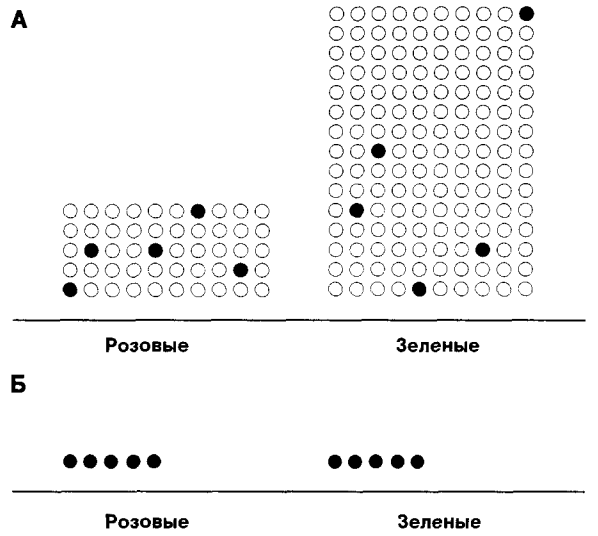


Рис. 5.4. А. Из совокупности марсиан, среди которых 150 зеленых и 50 розовых, извлекли случайную выборку из 10 особей. В выборку попало 5 зеленых и 5 розовых марсиан, на рисунке они помечены черным. **Б.** В таком виде данные предстанут перед исследователем, который не может наблюдать всю совокупность и вынужден судить о ней по выборке. Оценка доли розовых марсиан $\hat{p} = 5 / 10 = 0{,}5$.

Предположим, что из всех 200 марсиан случайным образом выбрали 10. Распределение розовых и зеленых марсиан во всей совокупности, неизвестное исследователям, изображено в верхней части рис. 5.4. Закрашенные кружки соответствуют марсианам, попавшим в выборку. В нижней части рис. 5.4 показана информация, которой располагал бы исследователь, получивший такую выборку. Как видим, в выборке розовые и зеленые марсиане поделились поровну. Основываясь на этих данных, мы решили бы, что розовых марсиан столько же, сколько и зеленых, то есть их доля составляет 50%.

Исследователь мог бы извлечь другую выборку, например одну из представленных на рис. 5.5. Здесь выборочные доли розовых марсиан равны 30, 30, 10 и 20%. Как любая выборочная оценка, оценка доли (обозначим ее $\hat{p}$) отражает долю p в совокупности, но отклоняется от нее в силу случайности. Рассмот-

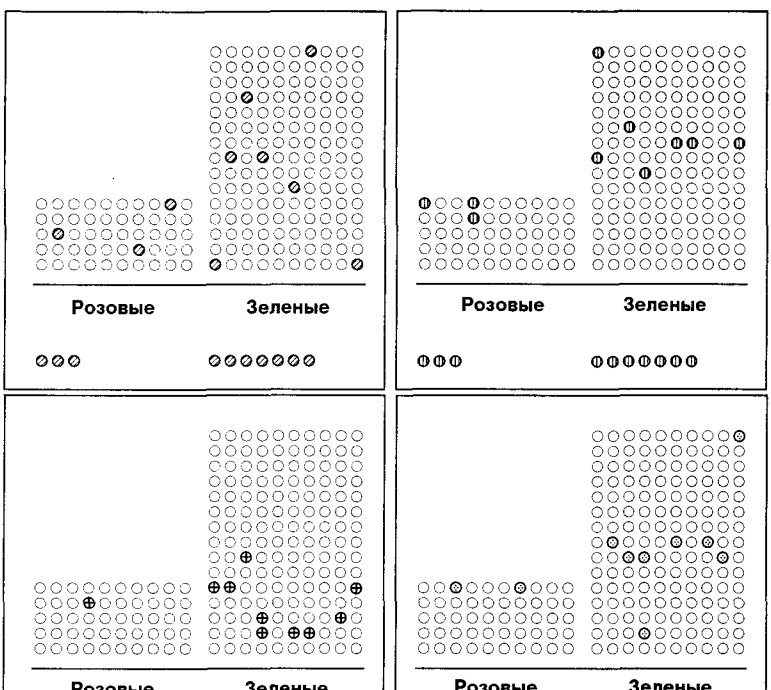


Рис. 5.5. Еще 4 случайные выборки из той же совокупности марсиан. Оценки доли розовых марсиан: 30, 30, 10 и 20%.

рим теперь не совокупность марсиан, а совокупность всех значений $\hat{p}$, вычисленных по выборкам объемом 10 каждая. (Из совокупности в 200 членов можно получить более 10^{16} таких выборок.) На рис. 5.6 приведены пять значений $\hat{p}$, вычисленных по пяти выборкам с рис. 5.4 и 5.5, и еще 20 значений, полученных на других случайных выборках того же объема. Среднее этих 25 значений составляет 30%. Это близко к истинной доле розовых марсиан — 25%. По аналогии со стандартной ошибкой среднего найдем *стандартную ошибку доли*. Для этого нужно охарактеризовать разброс выборочных оценок доли, то есть рассчитать

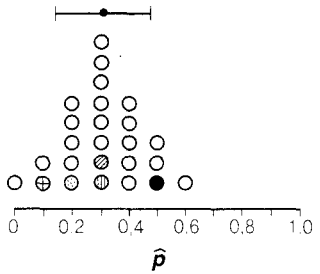


Рис. 5.6. Нанесем на график оценки доли розовых марсиан, полученные по выборке с рис. 5.4 и четырем выборкам с рис. 5.5. Добавим к ним еще 20 выборочных оценок. Получилось распределение выборочных оценок $\hat{p}$. Стандартное отклонение совокупности средних — это стандартная ошибка доли.

стандартное отклонение совокупности $\hat{p}$. В данном случае оно равно примерно 14%; в общем случае

$$\sigma_{\hat{p}} = \frac{\sigma}{\sqrt{n}},$$

где $\sigma_{\hat{p}}$ — стандартная ошибка доли, σ — стандартное отклонение, n — объем выборки. Поскольку $\sigma = \sqrt{p(1-p)}$, то

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}.$$

Заменив в приведенной формуле истинное значение доли ее оценкой $\hat{p}$, получим оценку стандартной ошибки доли:

$$s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Из центральной предельной теоремы (см. гл. 2) вытекает, что при достаточно большом объеме выборки выборочная оценка $\hat{p}$ приближенно подчиняется нормальному распределению, имеющему среднее p и стандартное отклонение $\sigma_{\hat{p}}$. Однако при значениях p, близких к 0 или 1, и при малом объеме выборки это не так. При какой численности выборки можно пользоваться приведенным способом оценки? Математическая статистика утверждает, что нормальное распределение служит хорошим при-

ближением, если и $n\hat{p}$ и $n(1-\hat{p})$ превосходят 5*. Напомним, что примерно 95% всех членов нормально распределенной совокупности находятся в пределах двух стандартных отклонений от среднего. Поэтому если перечисленные условия соблюдены, то с вероятностью 95% можно утверждать, что истинное значение p лежит в пределах $2s_{\hat{p}}$ от $\hat{p}$.

Вернемся на минуту к сравнению операционной летальности при галотановой и морфиновой анестезии. Напомним, что при использовании галотана летальность составила 13,1% (численность группы — 61 больной), а при использовании морфина — 14,9% (численность группы — 67 больных).

Стандартная ошибка доли для группы галотана

$$s_{\hat{p}_{\text{гал}}} = \sqrt{\frac{0{,}131(1-0{,}131)}{61}} = 0{,}043 = 4{,}3\%,$$

для группы морфина

$$s_{\hat{p}_{\text{мор}}} = \sqrt{\frac{0{,}149(1-0{,}149)}{67}} = 0{,}044 = 4{,}4\%.$$

Если учесть, что различие в летальности составило лишь 2%, то маловероятно, чтобы оно было обусловлено чем-нибудь, кроме случайного характера выборки.

Прежде чем двигаться дальше, перечислим те предпосылки, на которых основан излагаемый подход. Мы изучаем то, что в статистике принято называть *независимыми испытаниями Бернулли*. Эти испытания обладают следующими свойствами.

- Каждое отдельное испытание имеет ровно два возможных взаимно исключающих исхода.
- Вероятность данного исхода одна и та же в любом испытании.
- Все испытания независимы друг от друга.

В терминах совокупности и выборок эти свойства формулируются так.

---

* Если объем выборки недостаточен для использования нормального распределения, можно прибегнуть к помощи биномиального распределения. О биномиальном распределении см. J. H. Zar. Biostatistical analysis, 2nd ed. Prentice-Hall, Englewood Cliffs, N. J., 1984.

- Каждый член совокупности принадлежит одному из двух классов.
- Доля членов совокупности, принадлежащих одному классу, неизменна.
- Каждый член выборки извлекается из совокупности независимо от остальных.

СРАВНЕНИЕ ДОЛЕЙ

В предыдущей главе мы рассмотрели критерий Стьюдента t. Он вычисляется на основе выборочных средних и стандартной ошибки:

$$t = \frac{\text{Разность выборочных средних}}{\text{Стандартная ошибка разности выборочных средних}}.$$

Выборочная доля $\hat{p}$ аналогична выборочному среднему. Выражение для стандартной ошибки мы уже вывели. Теперь мы можем перейти к задаче сравнения долей, то есть к проверке нулевой гипотезы о равенстве долей. Для этого используется критерий z, аналогичный критерию Стьюдента t:

$$z = \frac{\text{Разность выборочных долей}}{\text{Стандартная ошибка разности выборочных долей}}.$$

Пусть $\hat{p}_1$ и $\hat{p}_2$ — выборочные доли. Поскольку стандартная ошибка — это стандартное отклонение всех возможных значений $\hat{p}$, полученных по выборкам заданного объема, и поскольку дисперсия разности равна сумме дисперсий, стандартная ошибка разности долей равна

$$s_{\hat{p}_1 - \hat{p}_2} = \sqrt{s_{\hat{p}_1}^2 + s_{\hat{p}_2}^2}.$$

Следовательно,

$$z = \frac{\hat{p}_1 - \hat{p}_2}{s_{\hat{p}_1 - \hat{p}_2}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{s_{\hat{p}_1}^2 + s_{\hat{p}_2}^2}}.$$

Если n_1 и n_2 — объемы двух выборок, то

$$s_{\hat{p}_1} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}} \text{ и } s_{\hat{p}_2} = \sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

Таким образом,

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}.$$

Итак, мы вывели формулу для критерия z. Вообще, этой буквой обозначаются величины со *стандартным нормальным распределением* (то есть нормальным распределением со средним $\mu = 0$ и стандартным отклонением $\sigma = 1$, см. табл. 6.4). С величиной z мы встретимся еще неоднократно. В данном случае нормальное распределение имеет место только при достаточно больших объемах выборок*.

Если при оценке дисперсии объединить наблюдения из обеих выборок, чувствительность критерия Стьюдента увеличится. Таким же способом можно повысить чувствительность критерия z. Действительно, если справедлива нулевая гипотеза, то обе выборочные доли $\hat{p}_1 = m_1/n_1$ и $\hat{p}_2 = m_2/n_2$ — это две оценки одной и той же доли p, которую мы, следовательно, можем оценить как

$$\hat{p} = \frac{m_1 + m_2}{n_1 + n_2}.$$

Тогда

$$s_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})}.$$

Отсюда имеем

$$s_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{s_{\hat{p}}^2}{n_1} + \frac{s_{\hat{p}}^2}{n_2}} = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

---

* Точнее говоря, когда значения $n\hat{p}$ и $n(1-\hat{p})$ больше 5. Если хотя бы для одной выборки это условие не выполняется, то критерий z неприменим и нужно воспользоваться точным критерием Фишера. Этот критерий мы рассмотрим чуть позже.

Подставляя полученную объединенную оценку в формулу для критерия z, имеем:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

О статистически значимом различии долей можно говорить, если значение z окажется «большим». С такой же ситуацией мы имели дело, рассматривая критерий Стьюдента. Отличие состоит в том, что t подчиняется распределению Стьюдента, а z — стандартному нормальному распределению. Соответственно, для нахождения «больших» значений z нужно воспользоваться стандартным нормальным распределением (рис. 2.5). Однако, поскольку при увеличении числа степеней свободы распределение Стьюдента стремится к нормальному, критические значения z можно найти в последней строке табл. 4.1. Для 5% уровня значимости оно составляет 1,96, для 1% — 2,58.

Поправка Йейтса на непрерывность

Нормальное распределение служит лишь приближением для распределения z. При этом оценка P оказывается заниженной и нулевая гипотеза будет отвергаться слишком часто. Причина состоит в том, что z принимает только дискретные значения, тогда как приближающее его нормальное распределение непрерывно. Для компенсации излишнего «оптимизма» критерия z введена *поправка Йейтса*, называемая также *поправкой на непрерывность*. С учетом этой поправки выражение для z имеет следующий вид:

$$z = \frac{\left|\hat{p}_1 - \hat{p}_2\right| - \frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

Поправка Йейтса слегка уменьшает значение z, уменьшая тем самым расхождение с нормальным распределением.

Галотан и морфин: операционная летальность

Теперь мы можем, наконец, сравнить операционную летальность при галотановой и морфиновой анестезии. Как вы помните, Конахан и соавт. исходили из предположения о том, что морфин в меньшей степени угнетает кровообращение, чем галотан, и потому предпочтительнее для общей анестезии. Действительно, при использовании морфина артериальное давление и сердечный индекс были выше, чем при использовании галотана, и различия эти статистически значимы. Однако выводы делать рано — ведь до сих пор не проанализированы различия операционной летальности, а именно этот показатель наиболее значим с практической точки зрения.

Итак, среди получавших галотан (1-я группа) умерли 8 больных из 61 (13,1%), а среди получавших морфин (2-я группа) — 10 из 67 (14,9%). Объединенная оценка доли умерших

$$\hat{p} = \frac{8+10}{61+67} = 0{,}141.$$

Величина $n\hat{p}$ для каждой из выборок равна соответственно $n_1\hat{p}_1 = 61 \times 0{,}141 = 8{,}6$ и $n_2\hat{p}_2 = 67 \times 0{,}149 = 9{,}4$. Оба значения больше 5*, поэтому можно воспользоваться критерием z. С учетом поправки Йейтса имеем:

$$z = \frac{|\hat{p}_1 - \hat{p}_2| - \frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}{\sqrt{\hat{p}_1(1-\hat{p}_1)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} =$$

$$= \frac{|0{,}131 - 0{,}149| - \frac{1}{2}\left(\frac{1}{61} + \frac{1}{67}\right)}{\sqrt{0{,}141(1-0{,}141)\left(\frac{1}{61} + \frac{1}{67}\right)}} = 0{,}04.$$

Это очень маленькая величина. Она гораздо ниже 1,96 — кри-

---

* Больше 5 и $n(1-\hat{p})$ — нетрудно показать, что если $\hat{p} < 0{,}5$, то $n(1-\hat{p}) > n\hat{p}$.

тического значения для 5% уровня значимости. Следовательно, хотя галотан и морфин действуют на кровообращение по-разному, нет никаких оснований говорить о различии операционной летальности.

Этот пример очень поучителен: мы убедились, сколь важно учитывать *исход* лечения. Организм устроен сложно, действие любого препарата многообразно. Если препарат положительно влияет на сердечно-сосудистую систему, то не исключено, что он отрицательно влияет, к примеру, на органы дыхания. Какой из эффектов перевесит и как это скажется на конечном результате — предвидеть трудно. Вот почему влияние препарата на любой показатель, будь то артериальное давление или сердечный индекс, нельзя считать доказательством его эффективности, пока не доказана *клиническая* эффективность. Иными словами, следует четко различать *показатели процесса* — всевозможные изменения биохимических, физиологических и прочих параметров, которые, *как мы полагаем*, играют положительную или отрицательную роль, — и *показатели результата*, обладающие реальной клинической значимостью. Так, изменения артериального давления и сердечного индекса под действием галотана и морфина — это показатели процесса, которые никак не сказались на показателе результата — операционной летальности. Если бы мы довольствовались наблюдением показателей процесса, то заключили бы, что морфин лучше галотана, хотя, как оказалось, выбор анестетика на летальность вообще не влияет.

Читая медицинские публикации или слушая аргументы сторонника того или иного метода лечения, следует прежде всего уяснить, о каких показателях идет речь — процесса или результата. Продемонстрировать воздействие некоторого фактора на процесс существенно легче, чем выяснить, влияет ли он на результат. Регистрация показателей процесса обычно проста и не занимает много времени. Напротив, выяснение результата, как правило, требует длительной кропотливой работы и нередко связано с субъективными проблемами измерений, особенно если речь идет о качестве жизни. И все же, решая, необходим ли предлагаемый метод лечения, нужно удостовериться, что он положительно влияет именно на показатели результата. Поверьте, больного и его семью прежде всего волнует результат, а не процесс.

Тромбоз шунта у больных на гемодиализе

Гемодиализ позволяет сохранить жизнь людям, страдающим хронической почечной недостаточностью. При гемодиализе кровь больного пропускают через искусственную почку — аппарат, удаляющий из крови продукты обмена веществ. Искусственная почка подсоединяется к артерии и вене больного: кровь из артерии поступает в аппарат и оттуда, уже очищенная, — в вену. Так как гемодиализ проводится регулярно, больному устанавливают артериовенозный шунт. В артерию и вену на предплечье вводят тефлоновые трубки; их концы выводят наружу и соединяют друг с другом. При очередной процедуре гемодиализа трубки разъединяют между собой и присоединяют к аппарату. После диализа трубки вновь соединяют, и кровь течет по шунту из артерии в вену. Завихрения тока крови в местах соединения трубок и сосудов приводят к тому, что шунт часто тромбируется. Тромбы приходится регулярно удалять, а в тяжелых случаях даже менять шунт. Руководствуясь тем, что аспирин препятствует образованию тромбов, Г. Хартер и соавт.* решили проверить, нельзя ли снизить риск тромбоза назначением небольших доз аспирина (160 мг/сут). Было проведено контролируемое испытание. Все больные, согласившиеся на участие в испытании и не имевшие противопоказаний к аспирину, были случайным образом разделены на две группы: 1-я получала плацебо, 2-я — аспирин. Ни врач, дававший больному препарат, ни больной не знали, был это аспирин или плацебо. Такой способ проведения испытания (он называется *двойным слепым*) исключает «подсуживание» со стороны врача или больного и, хотя технически сложен, дает наиболее надежные результаты. Исследование проводилось до тех пор, пока общее число больных с тромбозом шунта не достигло 24. Группы практически не различались по возрасту, полу и продолжительности лечения гемодиализом.

В 1-й группе тромбоз шунта произошел у 18 из 25 больных, во 2-й — у 6 из 19. Можно ли говорить о статистически значимом

---

* H. R. Harter, J. W. Burch, P. W. Majerus, N. Stanford, J. A. Delmez, C. B. Anderson, C. A. Weerts. Prevention of thrombosis in patients in hemodialysis by low-dose aspirin. *N. Engl. J. Med.*, 301:577—579, 1979.

различии доли больных с тромбозом, а тем самым об эффективности аспирина?

Прежде всего оценим долю больных с тромбозами в каждой из групп:

$$\hat{p}_1 = \frac{18}{25} = 0{,}72,$$

$$\hat{p}_2 = \frac{6}{19} = 0{,}32.$$

Проверим, можно ли применять критерий z: рассчитаем величины $n\hat{p}$ и $n(1-\hat{p})$ в каждой из групп:

$$n_1\hat{p}_1 = 18,\ n_1(1-\hat{p}_1) = 7$$

и

$$n_2\hat{p}_2 = 6,\ n_2(1-\hat{p}_2) = 13.$$

Как видим, все величины больше 5, поэтому критерий z применить можно.

Объединенная оценка доли больных с тромбозом

$$\hat{p} = \frac{6+18}{19+25} = 0{,}55.$$

Тогда

$$s_{\hat{p}_1-\hat{p}_2} = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = \sqrt{0{,}55(1-0{,}55)\left(\frac{1}{25}+\frac{1}{19}\right)} = 0{,}15.$$

Наконец, вычислим значение z

$$z = \frac{|\hat{p}_1-\hat{p}_2| - \frac{1}{2}\left(\frac{1}{25}+\frac{1}{19}\right)}{s_{\hat{p}_1-\hat{p}_2}} = \frac{|0{,}72-0{,}32|-0{,}05}{0{,}15} = 2{,}33.$$

По табл. 4.1 находим, что для 2% уровня значимости критическое значение z составляет 2,3263, то есть меньше, чем мы получили. А это значит, что снижение риска тромбоза шунта при приеме аспирина статистически значимо. Иными словами, если бы группы представляли собой две случайные выборки из одной

совокупности, то вероятность получить наблюдаемые (или большие) различия не превышала бы 2%.

ТАБЛИЦЫ СОПРЯЖЕННОСТИ: КРИТЕРИЙ χ^2

Рассмотренный выше метод хорошо работает, если качественный признак, который нас интересует, принимает два значения (тромбоз есть—нет, марсианин зеленый—розовый). Более того, поскольку метод является прямым аналогом критерия Стьюдента, число сравниваемых выборок также должно быть равно двум. Понятно, что и число значений признака, и число выборок может оказаться большим двух. Для анализа таких случаев нужен иной метод, аналогичный дисперсионному анализу. С виду этот метод, который мы сейчас изложим, сильно отличается от критерия z, но на самом деле между ними много общего.

Чтоб не ходить далеко за примером, начнем с только что разобранной задачи о тромбозе шунтов. Теперь мы будем рассматривать не *долю,* а *число* больных с тромбозом. Занесем результаты испытания в таблицу (табл. 5.1). Для каждой из групп укажем число больных с тромбозом и без тромбоза. У нас два признака: препарат (аспирин—плацебо) и тромбоз (есть—нет); в таблице указаны все их возможные сочетания, поэтому такая таблица называется таблицей сопряженности. В данном случае размер таблицы 2×2.

Посмотрим на клетки, расположенные на диагонали, идущей из верхнего левого в нижний правый угол. Числа в них заметно больше чисел в других клетках таблицы. Это наводит на мысль о связи между приемом аспирина и риском тромбоза.

Теперь взглянем на табл. 5.2. Это таблица *ожидаемых* чисел, которые мы получили бы, если бы аспирин не влиял на риск тромбоза. Как рассчитать ожидаемые числа, мы разберем чуть ниже, а пока обратим внимание на внешние особенности таблицы. Кроме немного пугающих дробных чисел в клетках можно заметить еще одно отличие от табл. 5.1 — это суммарные данные по группам в правом столбце и по тромбозам — в нижней строке. В правом нижнем углу — общее число больных в испытании. Об-

Таблица 5.1. Тромбозы шунта при приеме плацебо и аспирина

	Тромбоз есть	Тромбоза нет
Плацебо	18	7
Аспирин	6	13

ратите внимание, что, хотя числа в клетках на рис. 5.1 и 5.2 разные, суммы по строкам и по столбцам одинаковы.

Как же рассчитать ожидаемые числа? Плацебо получали 25 человек, аспирин — 19. Тромбоз шунта произошел у 24 из 44 обследованных, то есть в 54,55% случаев, не произошел — у 20 из 44, то есть в 45,45% случаев. Примем нулевую гипотезу о том, что аспирин *не влияет* на риск тромбоза. Тогда тромбоз должен с равной частотой 54,55% наблюдаться в группах плацебо и аспирина. Рассчитав, сколько составляет 54,55% от 25 и 19, получим соответственно 13,64 и 10,36. Это и есть ожидаемые числа больных с тромбозом в группах плацебо и аспирина. Таким же образом можно получить ожидаемые числа больных без тромбоза: в группе плацебо — 45,45% от 25, то есть 11,36, в группе аспирина — 45,45% от 19, то есть 8,64. Обратите внимание, что ожидаемые числа рассчитываются до второго знака после запятой — такая точность понадобится при дальнейших вычислениях.

Сравним табл. 5.1 и 5.2. Числа в клетках довольно сильно различаются. Следовательно, реальная картина отличается от той, которая наблюдалась бы, если бы аспирин не оказывал влияния на риск тромбоза. Теперь осталось построить критерий, который бы характеризовал эти различия одним числом, и затем найти его критическое значение, — то есть поступить так, как в случае критериев F, t или z.

Однако сначала вспомним еще один, уже знакомый нам при-

Таблица 5.2. Тромбозы шунта при приеме плацебо и аспирина: ожидаемые числа

	Тромбоз есть	Тромбоза нет	Всего
Плацебо	13,64	11,36	25
Аспирин	10,36	8,64	19
Всего	24	20	44

Таблица 5.3. Операционная летальность при галотановой и морфиновой анестезии

	Живы	Умерли	Всего
Галотан	53	8	61
Морфин	57	10	67
Всего	110	18	128

мер — работу Конахана по сравнению галотана и морфина, а именно ту часть, где сравнивалась операционная летальность. Соответствующие данные приведены в табл. 5.3. Форма таблицы такая же, что и табл. 5.1. В свою очередь, табл. 5.4, подобно табл. 5.2, содержит ожидаемые числа, то есть числа, вычисленные исходя из предположения, что летальность не зависит от анестетика. Из всех 128 оперированных в живых осталось 110, то есть 85,94%. Если бы выбор анестезии не оказывал влияния на летальность, то в обеих группах *доля* выживших была бы такой же и *число* выживших составило бы: в группе галотана — 85,94% от 61, то есть 52,42, в группе морфина — 85,94% от 67, то есть 57,58. Таким же образом можно получить и ожидаемые числа умерших. Сравним таблицы 5.3 и 5.4. В отличие от предыдущего примера, различия между ожидаемыми и наблюдаемыми значениями очень малы. Как мы выяснили раньше, различий в летальности нет. Похоже, мы на правильном пути.

Критерий χ^2 для таблицы 2×2

Критерий χ^2 (читается «хи-квадрат») не требует никаких предположений относительно параметров совокупности, из которой извлечены выборки, — это первый из *непараметрических* критериев, с которым мы знакомимся. Займемся его построением. Во-первых, как и всегда, критерий должен давать одно число,

Таблица 5.4. Операционная летальность при галотановой и морфиновой анестезии: ожидаемые числа

	Живы	Умерли	Всего
Галотан	52,42	8,58	61
Морфин	57,58	9,42	67
Всего	110	18	128

которое служило бы мерой отличия наблюдаемых данных от ожидаемых, то есть в данном случае различия между таблицей наблюдаемых и ожидаемых чисел. Во-вторых, критерий должен учитывать, что различие, скажем, в одного больного имеет большее значение при малом ожидаемом числе, чем при большом.

Определим критерий χ^2 следующим образом:

$$\chi^2 = \sum \frac{(O-E)^2}{E},$$

где O — наблюдаемое число в клетке таблицы сопряженности, E — ожидаемое число в той же клетке. Суммирование проводится по всем клеткам таблицы. Как видно из формулы, чем больше разница наблюдаемого и ожидаемого числа, тем больший вклад вносит клетка в величину χ^2. При этом клетки с малым ожидаемым числом вносят больший вклад. Таким образом, критерий удовлетворяет обоим требованиям — во-первых, измеряет различия и, во-вторых, учитывает их величину *относительно ожидаемых чисел.*

Применим критерий χ^2 к данным по тромбозам шунта. В табл. 5.1 приведены наблюдаемые числа, а в табл. 5.2 — ожидаемые.

$$\chi^2 = \sum \frac{(O-E)^2}{E} =$$

$$= \frac{(18-13{,}64)^2}{13{,}64} + \frac{(7-11{,}36)^2}{11{,}36} + \frac{(6-10{,}36)^2}{10{,}36} + \frac{(13-8{,}64)^2}{8{,}64} = 7{,}10.$$

Много это или мало? Испытаем наш новый критерий на данных по галотановой и морфиновой анестезии (табл. 5.3 и 5.4):

$$\chi^2 = \frac{(53-52{,}42)^2}{52{,}42} + \frac{(8-8{,}58)^2}{8{,}58} + \frac{(57-57{,}58)^2}{57{,}58} + \frac{(10-9{,}42)^2}{9{,}42} = 0{,}09.$$

Разница найденных значений χ^2 довольно велика: 7,10 в первом случае и 0,09 во втором, что соответствует тем впечатлениям, которые мы получили, сравнивая табл. 5.1 с 5.2 и 5.3 с 5.4. В первом случае мы получили «большое» значение χ^2, «большим» бы-

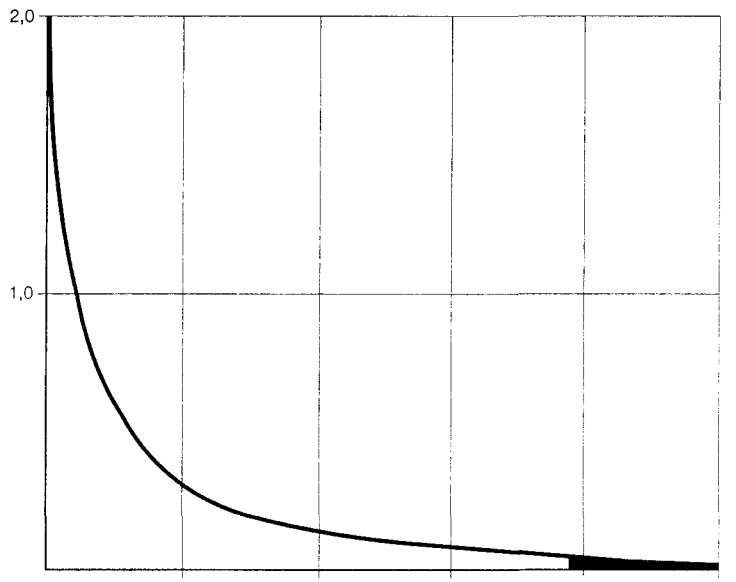


Рис. 5.7. Распределение χ^2 с 1 степенью свободы. Заштрихованная зона — это 5% наибольших значений.

ло и значение z, полученное по тем же данным. Можно показать, что для таблиц сопряженности размером 2×2 выполняется равенство $\chi^2 = z^2$.

Критическое значение χ^2 можно найти хорошо знакомым нам способом. На рис. 5.7 показано распределение возможных значений χ^2 для таблиц сопряженности размером 2×2 для случая, когда между изучаемыми признаками нет никакой связи. Величина χ^2 превышает 3,84 только в 5% случаев. Таким образом, 3,84 — критическое значение для 5% уровня значимости. В примере с тромбозом шунта мы получили значение 7,10, поэтому мы отклоняем гипотезу об отсутствии связи между приемом аспирина и образованием тромбов. Напротив, данные из табл. 5.3 хорошо согласуются с гипотезой об одинаковом влиянии галотана и морфина на послеоперационный уровень смертности.

Разумеется, как и все критерии значимости, χ^2 дает *вероятностную* оценку истинности той или иной гипотезы. На самом деле аспирин может и не оказывать влияния на риск тромбоза. На самом деле галотан и морфин могут по-разному влиять на операционную летальность. Но, как показал критерий, и то и другое *маловероятно*.

Применение критерия χ^2 правомерно, если *ожидаемое число в любой из клеток больше или равно 5**. Это условие аналогично условию применимости критерия z.

Критическое значение χ^2 зависит от размеров таблицы сопряженности, то есть от числа сравниваемых методов лечения (строк таблицы) и числа возможных исходов (столбцов таблицы). Размер таблицы выражается числом степеней свободы ν:

$$\nu = (r-1)(c-1),$$

где r — число строк, а c — число столбцов. Для таблиц размером 2×2 имеем $\nu = (2-1)(2-1) = 1$. Критические значения χ^2 для разных ν приведены в табл. 5.7.

Приведенная ранее формула для χ^2 в случае таблицы 2×2 (то есть при 1 степени свободы) дает несколько завышенные значения (сходная ситуация была с критерием z). Это вызвано тем, что теоретическое распределение χ^2 непрерывно, тогда как набор вычисленных значений χ^2 дискретен. На практике это приведет к тому, что нулевая гипотеза будет отвергаться слишком часто. Чтобы компенсировать этот эффект, в формулу вводят поправку Йейтса:

$$\chi^2 = \sum \frac{\left(|O-E| - \frac{1}{2} \right)^2}{E}.$$

Заметим, поправка Йейтса применяется только при $\nu = 1$, то есть для таблиц 2×2.

Применим поправку Йейтса к изучению связи между приемом аспирина и тромбозами шунта (табл. 5.1 и 5.2):

---

* В противном случае мы вынуждены использовать точный критерий Фишера.

$$\chi^2 = \frac{\left(|18-13{,}64|-\frac{1}{2}\right)^2}{13{,}64} + \frac{\left(|7-11{,}36|-\frac{1}{2}\right)^2}{11{,}36} +$$

$$+ \frac{\left(|6-10{,}36|-\frac{1}{2}\right)^2}{10{,}36} + \frac{\left(|13-8{,}64|-\frac{1}{2}\right)^2}{8{,}64} = 5{,}57.$$

Как вы помните, без поправки Йейтса значение χ^2 равнялось 7,10. Исправленное значение χ^2 оказалось меньше 6,635 — критического значения для 1% уровня значимости, но по-прежнему превосходит 5,024 — критическое значение для 2,5% уровня значимости.

Критерий χ^2 для произвольной таблицы сопряженности

Теперь рассмотрим случай, когда таблица сопряженности имеет число строк или столбцов, большее двух. Обратите внимание, что критерий z в таких случаях неприменим.

В гл. 3 мы показали, что занятия бегом уменьшают число менструаций*. Побуждают ли эти изменения обращаться к врачу? В табл. 5.5 приведены результаты опроса участниц исследования. Подтверждают ли эти данные гипотезу о том, что занятия бегом не влияют на вероятность обращения к врачу по поводу нерегулярности менструаций?

Из 165 обследованных женщин 69 (то есть 42%) обратились к врачу, остальные 96 (то есть 58%) к врачу не обращались. Если

Таблица 5.5. Частота обращения к врачу по поводу менструаций

Группа	Обращались	Не обращались	Всего
Контрольная	14	40	54
Физкультурницы	9	14	23
Спортсменки	46	42	88
Всего	69	96	165

* При этом мы для простоты вычислений размеры всех трех групп — контрольной, физкультурниц и спортсменок — полагали одинаковыми. Теперь мы воспользуемся настоящими данными.

Таблица 5.6. Частота обращения к врачу по поводу менструаций: ожидаемые числа

Группа	Обращались	Не обращались	Всего
Контрольная	22,58	31,42	54
Физкультурницы	9,62	13,38	23
Спортсменки	36,80	51,20	88
Всего	69	96	165

занятия бегом не влияют на вероятность обращения к врачу, то в каждой из групп к врачу должно было обратиться 42% женщин. В табл. 5.6 приведены соответствующие ожидаемые значения. Сильно ли отличаются от них реальные данные?

Для ответа на этот вопрос вычислим χ^2:

$$\chi^2 = \frac{(14-22{,}58)^2}{22{,}58} + \frac{(40-31{,}42)^2}{31{,}42} + \frac{(9-9{,}62)^2}{9{,}62} + \\ + \frac{(14-13{,}38)^2}{13{,}38} + \frac{(46-36{,}80)^2}{36{,}80} + \frac{(42-51{,}20)^2}{51{,}20} = 9{,}63.$$

Число строк таблицы сопряженности равно трем, столбцов — двум, поэтому число степеней свободы $\nu = (3-1)(2-1) = 2$. Если гипотеза об отсутствии межгрупповых различий верна, то, как видно из табл. 5.7, значение χ^2 превзойдет 9,21 не более чем в 1% случаев. Полученное значение больше. Тем самым, при уровне значимости 0,01 можно отклонить гипотезу об отсутствии связи между бегом и обращениями к врачу по поводу менструаций. Однако, выяснив, что связь существует, мы тем не менее не сможем указать, какая (какие) именно группы отличаются от остальных.

Итак, мы познакомились с критерием χ^2. Вот порядок его применения.

- Постройте по имеющимся данным таблицу сопряженности.
- Подсчитайте число объектов в каждой строке и в каждом столбце и найдите, какую долю от общего числа объектов составляют эти величины.
- Зная эти доли, подсчитайте с точностью до двух знаков после запятой ожидаемые числа — количество объектов, которое

попало бы в каждую клетку таблицы, если бы связь между строками и столбцами отсутствовала.

- Найдите величину χ^2, характеризующую различия наблюдаемых и ожидаемых значений. Если таблица сопряженности имеет размер 2×2, примените поправку Йейтса.
- Вычислите число степеней свободы, выберите уровень значимости и по табл. 5.7 определите критическое значение χ^2. Сравните его с полученным для вашей таблицы.

Как вы помните, для таблиц сопряженности размером 2×2 критерий χ^2 применим только в случае, когда все ожидаемые числа больше 5. Как обстоит дело с таблицами большего размера? В этом случае критерий χ^2 применим, если все ожидаемые числа не меньше 1 и доля клеток с ожидаемыми числами меньше 5 не превышает 20%. При невыполнении этих условий критерий χ^2 может дать ложные результаты. В таком случае можно собрать дополнительные данные, однако это не всегда осуществимо. Есть и более простой путь — объединить несколько строк или столбцов. Ниже мы покажем, как это сделать.

Преобразование таблиц сопряженности

В предыдущем разделе мы установили *существование* связи между занятием бегом и обращениями к врачу по поводу менструаций, или, что то же самое, существование различий между группами по частоте обращения к врачу. Однако мы не могли определить, *какие именно* группы отличаются друг от друга, а какие нет. С похожей ситуацией мы сталкивались в дисперсионном анализе. При сравнении нескольких групп дисперсионный анализ позволяет обнаружить сам факт существования различий, но не указывает выделяющиеся группы. Последнее позволяют сделать процедуры множественного сравнения, о которых мы говорили в гл. 4. Нечто похожее можно проделать и с таблицами сопряженности.

Глядя на табл. 5.5, можно предположить, что физкультурницы и спортсменки обращались к врачу чаще, чем женщины из контрольной группы. Различие между физкультурницами и спортсменками кажется незначительным.

Проверим гипотезу о том, что физкультурницы и спортсмен-

Таблица 5.7. Критические значения χ^2

	Уровень значимости							
ν	0,50	0,25	0,10	0,05	0,025	0,01	0,005	0,001
1	0,455	1,323	2,706	3,841	5,024	6,635	7,879	10,828
2	1,386	2,773	4,605	5,991	7,378	9,210	10,597	13,816
3	2,366	4,108	6,251	7,815	9,348	11,345	12,838	16,266
4	3,357	5,385	7,779	9,488	11,143	13,277	14,860	18,467
5	4,351	6,626	9,236	11,070	12,833	15,086	16,750	20,515
6	5,348	7,841	10,645	12,592	14,449	16,812	18,548	22,458
7	6,346	9,037	12,017	14,067	16,013	18,475	20,278	24,322
8	7,344	10,219	13,362	15,507	17,535	20,090	21,955	26,124
9	8,343	11,389	14,684	16,919	19,023	21,666	23,589	27,877
10	*9,342*	*12,549*	*15,987*	*18,307*	*20,483*	*23,209*	*25,188*	*29,588*
11	10,341	13,701	17,275	19,675	21,920	24,725	26,757	31,264
12	11,340	14,845	18,549	21,026	23,337	26,217	28,300	32,909
13	12,340	15,984	19,812	22,362	24,736	27,688	29,819	34,528
14	13,339	17,117	21,064	23,685	26,119	29,141	31,319	36,123
15	14,339	18,245	22,307	24,996	27,488	30,578	32,801	37,697
16	15,338	19,369	23,542	26,296	28,845	32,000	34,267	39,252
17	16,338	20,489	24,769	27,587	30,191	33,409	35,718	40,790
18	17,338	21,605	25,989	28,869	31,526	34,805	37,156	42,312
19	18,338	22,718	27,204	30,144	32,852	36,191	38,582	43,820
20	19,337	23,828	28,412	31,410	34,170	37,566	39,997	45,315
21	20,337	24,935	29,615	32,671	35,479	38,932	41,401	46,797
22	21,337	26,039	30,813	33,924	36,781	40,289	42,796	48,268
23	22,337	27,141	32,007	35,172	38,076	41,638	44,181	49,728
24	23,337	28,241	33,196	36,415	39,364	42,980	45,559	51,179
25	24,337	29,339	34,382	37,652	40,646	44,314	46,928	52,620
26	25,336	30,435	35,563	38,885	41,923	45,642	48,290	54,052
27	26,336	31,528	36,741	40,113	43,195	46,963	49,645	55,476
28	27,336	32,020	37,916	41,337	44,461	48,278	50,993	56,892
29	28,336	33,711	39,087	42,557	45,722	49,588	52,336	58,301
30	29,336	34,800	40,256	43,773	46,979	50,892	53,672	59,703
31	30,336	35,887	41,422	44,985	48,232	52,191	55,003	61,098
32	31,336	36,973	42,585	46,194	49,480	53,486	56,328	62,487
33	32,336	38,058	43,745	47,400	50,725	54,776	57,648	63,870
34	33,336	39,141	44,903	48,602	51,966	56,061	58,964	65,247
35	34,336	40,223	46,059	49,802	53,203	57,342	60,275	66,619
36	35,336	41,304	47,212	50,998	54,437	58,619	61,581	67,985
37	36,336	42,383	48,363	52,192	55,668	59,893	62,883	69,346
38	37,335	43,462	49,513	53,384	56,896	61,162	64,181	70,703
39	38,335	44,539	50,660	54,572	58,120	62,428	65,476	72,055
40	39,335	45,616	51,805	55,758	59,342	63,691	66,766	73,402

Таблица 5.7. Окончание

	Уровень значимости							
ν	0,50	0,25	0,10	0,05	0,025	0,01	0,005	0,001
41	40,335	46,692	52,949	56,942	60,561	64,950	68,053	74,745
42	41,335	47,766	54,090	58,124	61,777	66,206	69,336	76,084
43	42,335	48,840	55,230	59,304	62,990	67,459	70,616	77,419
44	43,335	49,913	56,369	60,481	64,201	68,710	71,893	78,750
45	44,335	50,985	57,505	61,656	65,410	69,957	73,166	80,077
46	45,335	52,056	58,641	62,830	66,617	71,201	74,437	81,400
47	46,335	53,127	59,774	64,001	67,821	72,443	75,704	82,720
48	47,335	54,196	60,907	65,171	69,023	73,683	76,969	84,037
49	48,335	55,265	62,038	66,339	70,222	74,919	78,231	85,351
50	49,335	56,334	63,167	67,505	71,420	76,154	79,490	86,661

J. H. Zar, Biostatistical Analysis, 2d ed, Prentice-Hall, Englewood Cliffs, N.J., 1984.

ки обращаются к врачу одинаково часто. Для этого выделим из исходной таблицы подтаблицу, содержащую данные по двум этим группам. В табл. 5.8 приведены наблюдаемые и ожидаемые числа; они довольно близки.

Размер таблицы 2×2. Поэтому вычислим χ^2 с поправкой Йейтса:

$$\chi^2 = \sum \frac{\left(|O-E|-\frac{1}{2}\right)^2}{E} =$$

$$= \frac{\left(|9-11{,}40|-\frac{1}{2}\right)^2}{11{,}40} + \frac{\left(|14-11{,}60|-\frac{1}{2}\right)^2}{11{,}60} +$$

$$+ \frac{\left(|46-43{,}60|-\frac{1}{2}\right)^2}{43{,}60} + \frac{\left(|42-44{,}40|-\frac{1}{2}\right)^2}{44{,}40} = 0{,}79.$$

Полученная величина значительно меньше критического значения. Поэтому гипотеза об отсутствии межгрупповых различий не отклоняется. Следовательно, эти группы можно объединить в одну. Полученную объединенную группу бегуний сравним с контрольной (табл. 5.9). На этот раз значение χ^2 равно 7,39, то

Таблица 5.8. Частота обращения к врачу по поводу менструаций (в скобках — ожидаемые числа)

Группа	Обращались	Не обращались	Всего
Физкультурницы	9 (11,40)	14 (11,60)	23
Спортсменки	46 (43,60)	42 (44,40)	88
Всего	55	56	111

Таблица 5.9. Частота обращения к врачу по поводу менструаций (в скобках — ожидаемые числа)

Группа	Обращались	Не обращались	Всего
Контрольная	14 (22,58)	40 (31,42)	54
Физкультурницы и спортсменки	55 (46,42)	56 (64,58)	111
Всего	69	96	165

есть больше критического значения 6,63, соответствующего уровню значимости 0,01.

Заметьте, мы выполнили два сравнения, используя одни и те же данные. Поэтому нужно применить поправку Бонферрони, умножив уровень значимости на 2. Исправленное значение уровня значимости $2 \times 0{,}01 = 0{,}02$. Итак, с уровнем значимости 0,02 мы заключаем, что физкультурницы не отличаются от спортсменок, но обе эти группы отличаются от женщин, не занимающихся бегом.

ТОЧНЫЙ КРИТЕРИЙ ФИШЕРА

Критерий χ^2 годится для анализа таблиц сопряженности 2×2, если ожидаемые значения в любой из ее клеток не меньше 5. Когда число наблюдений невелико, это условие не выполняется и критерий χ^2 неприменим. В этом случае используют *точный критерий Фишера*. Он основан на переборе всех возможных вариантов заполнения таблицы сопряженности при данной численности групп, поэтому чем она меньше, тем проще его применить.

Нулевая гипотеза состоит в том, что между лечением и исходом нет никакой связи. Тогда вероятность получить некоторую таблицу равна

Таблица 5.10. Обозначения, используемые в точном критерии Фишера

			Суммы по строкам
	O_{11}	O_{12}	R_1
	O_{21}	O_{22}	R_2
Суммы по столбцам	C_1	C_2	N

$$P = \frac{R_1!R_2!C_1!C_2!}{N!} \Big/ \frac{}{O_{11}!O_{12}!O_{21}!O_{22}!},$$

где R_1 и R_2 — суммы по строкам (число больных, лечившихся первым и вторым способом), C_1 и C_2 — суммы по столбцам (число больных с первым и вторым исходом), O_{11}, O_{12}, O_{21} и O_{22} — числа в клетках, N — общее число наблюдений (табл. 5.10). Восклицательный знак, как и всегда в математике, обозначает факториал*. Построив все остальные варианты заполнения таблицы, возможные при данных суммах по строкам и столбцам, по этой же формуле рассчитывают их вероятность. Вероятности, которые не превосходят вероятность исходной таблицы (включая саму эту вероятность), суммируют. Полученная сумма — это величина P для двустороннего варианта точного критерия Фишера.

В отличие от критерия χ^2, существуют одно- и двусторонний варианты точного критерия Фишера. К сожалению, в большинстве учебников описан именно односторонний вариант, он же обычно используется в компьютерных программах и приводится в статьях. Оно и не удивительно — ведь односторонний вариант дает меньшую величину P. Хуже то, что авторы не считают нужным хотя бы упомянуть, каким вариантом они пользовались. В табл. 5.11 показаны данные, которые получили Мак-Кинни и соавт.**, решив выяснить, насколько часто в статьях из двух

* Факториал числа — произведение всех целых чисел от этого числа до единицы: $n! = n \times (n-1) \times (n-2) \times \ldots \times 2 \times 1$. Например, $4! = 4 \times 3 \times 2 \times 1 = 24$. Факториал нуля равен единице.

** W. P. McKinney, M. J. Young, A. Harta, M. B. Lee. The inexact use of Ficher's exact test in six major medical journals. *JAMA*, 261:3430–3433, 1989.

Таблица 5.11. Частота указания варианта точного критерия Фишера в двух медицинских журналах

	Вариант критерия		
	Указан	Не указан	Всего
New England Journal of Medicine	1	8	9
Lancet	10	4	14
Всего	11	12	23

самых известных медицинских журналов указан вариант критерия. Выборка невелика, и критерий χ^2 применить нельзя. Поэтому для анализа использования точного критерия Фишера воспользуемся самим точным критерием Фишера. Из приведенной выше формулы для P следует, что вероятность при *тех же* значениях сумм по строкам и столбцам таблицы получить *такой же* набор чисел в клетках, что в табл. 5.11, равна

$$P = \frac{9!14!11!12!}{23!} = 0{,}00666.$$

Это небольшая вероятность. Теперь возьмем наименьшее из чисел в клетках (это единица на пересечении первой строки и первого столбца) и уменьшим его на 1. Числа в остальных клетках изменим так, чтобы суммы по строкам и столбцам остались прежними. Мы получили табл. 5.12. Соответствующая вероятность равна

$$P = \frac{9!14!11!12!}{23!} = 0{,}00027.$$

(Заметим, что числитель можно заново не вычислять, так как его значение зависит только от сумм по строкам и столбцам, которые не изменились.) Поскольку наименьшее число в клетке равно нулю, дальше уменьшать его невозможно. Таким образом, односторонний вариант точного критерия Фишера дает $P = 0{,}00666 + 0{,}00027 = 0{,}00695$.

Чтобы рассчитать значение двустороннего варианта точного критерия Фишера, нужно перебрать и все остальные возможные

Таблица 5.12.

	Вариант критерия		
	Указан	Не указан	Всего
New England Journal of Medicine	0	9	9
Lancet	11	3	14
Всего	11	12	23

варианты заполнения таблицы при условии неизменности сумм по строкам и столбцам. Получить все эти варианты несложно — надо только заметить, что при постоянных суммах по строкам и столбцам значения во всех четырех клетках полностью определяются значением в любой из них. Возьмем число все в той же левой верхней клетке и будем увеличивать его на 1, пересчитывая каждый раз числа в остальных клетках. В результате мы получим восемь вариантов заполнения (табл. 5.13). Для двух последних вариантов вероятность не превышает вероятности исходного варианта заполнения (0,00666), составляя соответственно 0,00242 и 0,00007. Таким образом, кроме исходного у нас есть еще три варианта «маловероятного» заполнения таблицы; просуммировав соответствующие вероятности и прибавив к ним вероятность исходного варианта, получим $P = 0{,}00666 + 0{,}00027 + 0{,}00242 + 0{,}00007 = 0{,}00944$. Это и есть значение двустороннего варианта точного критерия Фишера. Итак, различие частоты правильного использования точного критерия Фишера в журналах New England Journal of Medicine и Lancet статистически значимо ($P = 0{,}009$). В данном случае общий вывод при переходе от одностороннего к двустороннему варианту не изменился, однако так бывает далеко не всегда. Еще более грубая ошибка происходит, когда автор рассчитывает только вероятность получения исходной таблицы, пренебрегая построением остальных вариантов заполнения. Естественно, это приводит к сильному занижению P, то есть к «выявлению» различий там, где их нет.

В заключение изложим правила пользования точным критерием Фишера.

- Вычислите вероятность получить исходную таблицу.
- Постройте остальные возможные варианты заполнения таблицы при неизменных суммах по строкам и столбцам. Для

Таблица 5.13.

			Всего
	2	7	9
	9	5	14
Всего	11	12	23
	$P = 0{,}05330$		

			Всего
	6	3	9
	5	9	14
Всего	11	12	23
	$P = 0{,}12438$		

			Всего
	3	6	9
	8	6	14
Всего	11	12	23
	$P = 0{,}18657$		

			Всего
	7	2	9
	4	10	14
Всего	11	12	23
	$P = 0{,}02665$		

			Всего
	4	5	9
	7	7	14
Всего	11	12	23
	$P = 0{,}31983$		

			Всего
	8	1	9
	3	11	14
Всего	11	12	23
	$P = 0{,}00242$		

			Всего
	5	4	9
	6	8	14
Всего	11	12	23
	$P = 0{,}27985$		

			Всего
	9	0	9
	2	12	14
Всего	11	12	23
	$P = 0{,}00007$		

этого в одной из клеток проставьте все целые числа от нуля до максимально возможного, пересчитывая числа в остальных клетках так, чтобы суммы по строкам и столбцам оставались неизменными.

- Вычислите вероятности для всех полученных таблиц.
- Просуммируйте вероятность получить исходную таблицу и все вероятности, которые ее не превышают.

Итак, теперь мы умеем работать не только с количественными, но и с качественными признаками. Но вопрос, занимавший нас и в этой, и в предыдущих главах, был в сущности одним и тем же — как оценить статистическую значимость различий. В следующей главе мы взглянем на другую сторону медали. Именно, мы попытаемся понять, что означает *отсутствие* статистически значимых различий.

ЗАДАЧИ

5.1. Т. Бишоп (T. Bishop. High frequency neural modulation in dentistry. *J. Am. Dent. Assoc.*, 112:176—177, 1986) изучил эффективность высокочастотной стимуляции нерва в качестве обезболивающего средства при удалении зуба. Все больные подключались к прибору, но в одних случаях он работал, в других был выключен. Ни стоматолог, ни больной не знали, включен ли прибор. Позволяют ли следующие данные считать высокочастотную стимуляцию нерва действенным анальгезирующим средством?

	Прибор включен	Прибор выключен
Боли нет	24	3
Боль есть	6	17

5.2. Синдром внезапной детской смерти — основная причина смерти детей в возрасте от 1 недели до 1 года. Обычно смерть наступает на фоне полного здоровья незаметно, во сне, поэтому определение факторов риска имеет первостепенное значение. Считается, что синдром внезапной детской смерти чаще случается у недоношенных детей, негров, а также в семьях с низкими доходами. Н. Левак и соавт. (N. Lewak et al. Sudden infant death syndrome risk factors: prospective data review. *Clin. Pediatr.*, 18: 404—411, 1979) решили уточнить эти данные. Исследователи собрали сведения о 19047 детях, родившихся в одном из роддомов Окленда, штат Калифорния, с 1960 по 1967 г. Судьбу детей проследили до 1 года. Данных о 48 детях получить не удалось. От синдрома внезапной детской смерти умерли 44 ребенка. Данные о предполагаемых факторах риска представлены в табл. 5.14. Найдите признаки, связанные с риском синдрома внезапной детской смерти.

5.3. Могло ли повлиять отсутствие данных о 48 детях на результаты исследования? Если да, то как?

5.4. Р. Феннел и соавт. (R. Fennell et al. Urinary tract infections in children: effect of short course antibiotic therapy on recurrence rate in children with previous infections. *Clin. Pediatr.*, 19:121—124, 1980) сравнили эффективность трех антибиотиков при рецидивиру-

Таблица 5.14

Фактор		Синдром внезапной детской смерти	
		+	–
Возраст матери	До 25 лет	29	7301
	25 лет и старше	15	11241
Время от окончания предыдущей беременности	Менее 1 года	23	4694
	Более 1 года	11	7339
Планировалась ли беременность	Нет	23	7654
	Да	5	4253
Повторная беременность	Нет	36	12987
	Да	8	4999
Курение во время беременности	Да	24	5228
	Нет	10	9595
Посещения врача во время беременности	Менее 11 раз	31	10512
	11 раз или более	11	8154
Самый низкий гемоглобин во время беременности	Менее 12 мг%	26	12613
	12 мг% и более	7	2678
Раса	Белые	31	12240
	Негры	9	4323
	Другие	4	2153

По некоторым признакам данные отсутствуют, поэтому сумма в третьем столбце может оказаться меньше 44, а в четвертом — меньше 18 955.

ющей инфекции мочевых путей у девочек 3—16 лет. После короткого курса одного из антибактериальных препаратов (назначенного случайным образом) в течение года делали повторные посевы мочи. При выявлении бактериурии констатировали рецидив. Были получены следующие результаты.

	Рецидив	
	Есть	Нет
Ампициллин	20	7
Триметоприм/сульфаметоксазол	24	21
Цефалексин	14	2

Есть ли основания говорить о разной эффективности препаратов? Если да, то какой лучше?

5.5. А. О'Нил и соавт. (A. O'Neil et al. A waterborn epidemic of acute infectious non-bacterial gastroenteritis in Alberta, Canada. *Can. J. Public Health*, 76:199—203, 1985) недавно сообщили о вспышке гастроэнтерита в маленьком канадском городке. Исследователи предположили, что источником инфекции была водопроводная вода. Они исследовали зависимость между количеством выпитой воды и числом заболевших. Какие выводы можно сделать из приводимых данных?

Количество выпитой воды, стаканов в день	Число заболевших	Число не заболевших
Менее 1	39	121
От 1 до 4	265	258
5 и более	265	146

5.6. Как правило, качество исследования выше, а соответствие собираемых данных поставленному вопросу точнее, если данные собираются специально для этого исследования после его планирования. Р. и С. Флетчеры (R. Fletcher, S. Fletcher. Clinical research in general medical journals: a 30-year perspective. *N. Engl. J. Med.*, 301:180—183, 1979) исследовали 612 работ, случайным образом выбранных из журналов Journal of American Medical Association, Lancet и New England Journal of Medicine, чтобы определить, собирали ли их авторы свои данные до или после планирования исследования. Вот что удалось обнаружить:

	1946	1956	1966	1976
Число рассмотренных работ	151	149	157	155
Процент работ, где данные собирали				
после планирования исследования	76	71	49	44
до планирования исследования	24	29	51	56

Оцените статистическую значимость различия долей. Если различия есть, то можно ли сказать, что положение меняется к лучшему?

5.7. Одна из причин инсульта — окклюзия сонной артерии. Чтобы выяснить, какое лечение — медикаментозное или хирургическое — дает в этом случае лучшие результаты, У. Филдс и соавт. (W. Fields et al. Joint study of extracranial arterial occlusion, V: Progress report of prognosis following surgery or nonsurginal treatment for transient ishemic attacks and cervical carotid artery lesions. *JAMA*, 211:1993—2003, 1970) сравнили долгосрочный прогноз у леченных двумя методами.

	Повторный инсульт или смерть	
Лечение	Да	Нет
Хирургическое	43	36
Медикаментозное	53	19

Можно ли говорить о превосходстве одного из видов лечения?

5.8. В диагностике ишемической болезни сердца используют нагрузочную пробу: с помощью физической нагрузки вызывают ишемию миокарда, которую выявляют на ЭКГ. Существует другой метод: ишемию вызывают внутривенным введением дипиридамола, а выявляют с помощью эхокардиографии. Ф. Латтанци и соавт. (F. Lattanzi et al. Inhibition of dipyridamole-induced ishemia by antianginal therapy in humans: correlation with exercise electrocardiography. *Circulation*, 83:1256—1262, 1991) сравнили результаты двух методов у больных, получавших и не получавших антиангинальную терапию. Результаты приведены в таблице.

Без антиангинальной терапии

		Дипиридамол + эхокардиография	
		+	–
Нагрузка + ЭКГ	+	38	2
	–	14	3

На фоне антиангинальной терапии

		Дипиридамол + эхокардиография	
		+	–
Нагрузка + ЭКГ	+	21	6
	–	16	14

Оцените различия между результатами двух методов.

5.9. Д. Сакетт и М. Гент (D. Sackett, M. Gent. Controversy in counting and attributing events in clinical trials. *N. Engl. J. Med.*, 301:1410—1412, 1979) сделали важное замечание относительно методики сбора данных в исследовании результатов медикаментозного и хирургического лечения окклюзии сонной артерии (задача 5.7). Так как изучался «долгосрочный прогноз», в исследование включали только тех больных, которые не умерли и у которых не было повторного инсульта во время госпитализации. В результате из рассмотрения были исключены 15 оперированных (5 из них умерли, а у 10 инсульт произошел вскоре после операции) и только 1 больной, лечившийся медикаментозно. Если учесть и этих 16 больных, то данные примут такой вид:

	Повторный инсульт или смерть	
Лечение	Да	Нет
Хирургическое	58	36
Медикаментозное	54	19

Что теперь можно сказать о предпочтительности одного из видов лечения? Какое сравнение более верно — с учетом этих 16 больных или без их учета (как в задаче 5.7)? Почему?

5.10. Распространенность болезни X равна 10%. Болезнью Y страдает 1000 человек, болезнью Z — также 1000 человек. Болезнь X с равной вероятностью поражает страдающих болезнями Y и Z. Вероятность госпитализации при этих болезнях разная: для болезни X она составляет 40%, Y — 50%, Z — 20%. Посмотрим, сколько больных с разными сочетаниями болезней окажется в больнице.

Из 1000 человек, страдающих болезнью Y, болезнь X имеют 10%, то есть 100 человек. Из них 50% (50 человек) будут госпитализированы в связи с болезнью Y, из оставшихся 50 человек в связи с болезнью X госпитализируют 40%, то есть 20 человек. Таким образом, в больнице окажется 70 больных с сочетанием болезней Y и X.

Из 900 человек, страдающих болезнью Y, но не X, будут госпитализированы 50%, то есть 450 человек.

Такой же расчет для болезни Z показывает, что в больницу

попадет 52 человека с сочетанием болезней Z и X, а с болезнью Z, но не X, — 180 человек.

Исследователь, работающий в больнице, в которую попали все госпитализированные, обнаружил следующую связь.

	С болезнью X	Без болезни X
Болезнь Y	70	450
Болезнь Z	52	180

Оцените статистическую значимость различий частоты болезни X среди страдающих болезнями Y и Z. Можно ли по этим данным судить о связи болезней Y и Z с болезнью X? (Приведенный пример заимствован из работы: D. Mainland. The risk of fallacious conclusions from autopsy data on the incidence of diseases with applications to heart disease. *Am. Heart J.*, 45:644—654, 1953.)

Что значит «незначимо»: чувствительность критерия

До сих пор мы занимались оценкой вероятности нулевой гипотезы, то есть предположения об отсутствии эффекта экспериментального воздействия. Вероятность нулевой гипотезы (P) мы оценивали с помощью различных критериев значимости — F, t, q, q', z и χ^2. Если значение критерия превышало критическое, нулевую гипотезу отклоняли. При этом мы совершенно справедливо утверждали, что нашли *статистически значимые различия*. Если значение критерия оказывалось меньше критического, говорили об *отсутствии статистически значимых различий*. И это тоже справедливо. К сожалению, обычно этим не ограничиваются. Не обнаружив различий, исследователь считает это доказательством их отсутствия. А это уже совершенно неверно. Прежде чем сделать вывод об отсутствии различий, следует выяснить, была ли *чувствительность* критерия достаточной, чтобы их обнаружить.

Чувствительностью* называется способность критерия обнаружить различия. Чувствительность зависит от величины раз-

* С этим понятием мы уже встречались в гл. 3 и 4; другое название чувствительности — мощность.

личий, от разброса данных и от объема выборки. Наиболее важен объем выборок: чем он больше, тем чувствительнее критерий. При достаточно больших выборках малейшее различие оказывается статистически значимым. И наоборот, если выборки малы, даже большие различия статистически незначимы. Зная эти закономерности, можно заранее определить численность выборок, необходимую для выявления эффекта.

ЭФФЕКТИВНЫЙ ДИУРЕТИК

Разбирая критерий Стьюдента, мы использовали пример, в котором препарат, предположительно обладавший диуретическим действием, в действительности не увеличивал диурез. Сейчас рассмотрим обратный пример. Исследуемый препарат — на самом деле диуретик. Он увеличивает суточный диурез в среднем с 1200 до 1400 мл. На рис. 6.1А показано распределение суточного диуреза для всех 200 членов совокупности при приеме плацебо, а на рис. 6.1Б — при приеме этого препарата.

Теперь представим себе исследователя, который, разумеется, не может наблюдать всю совокупность. Случайным образом он выбирает две группы по 10 человек в каждой, дает 1-й группе плацебо, а 2-й – препарат (диуретик), после чего измеряет суточный диурез в обеих группах. На рис. 6.1В представлены результаты этих измерений. В 1-й группе средний суточный диурез составил 1180 мл (стандартное отклонение 144 мл), а во 2-й группе — 1400 мл (стандартное отклонение 245 мл). Оценим различия по критерию Стьюдента.

Объединенная оценка дисперсии равна

$$s^2 = \frac{1}{2}(s_1^2 + s_2^2) = \frac{1}{2}(144^2 + 245^2) = 40381 = 201^2.$$

Значение t равно

$$t = \frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\frac{s^2}{n_2} + \frac{s^2}{n_1}}} = \frac{1400 - 1180}{\sqrt{\frac{201^2}{10} + \frac{201^2}{10}}} = 2{,}447,$$

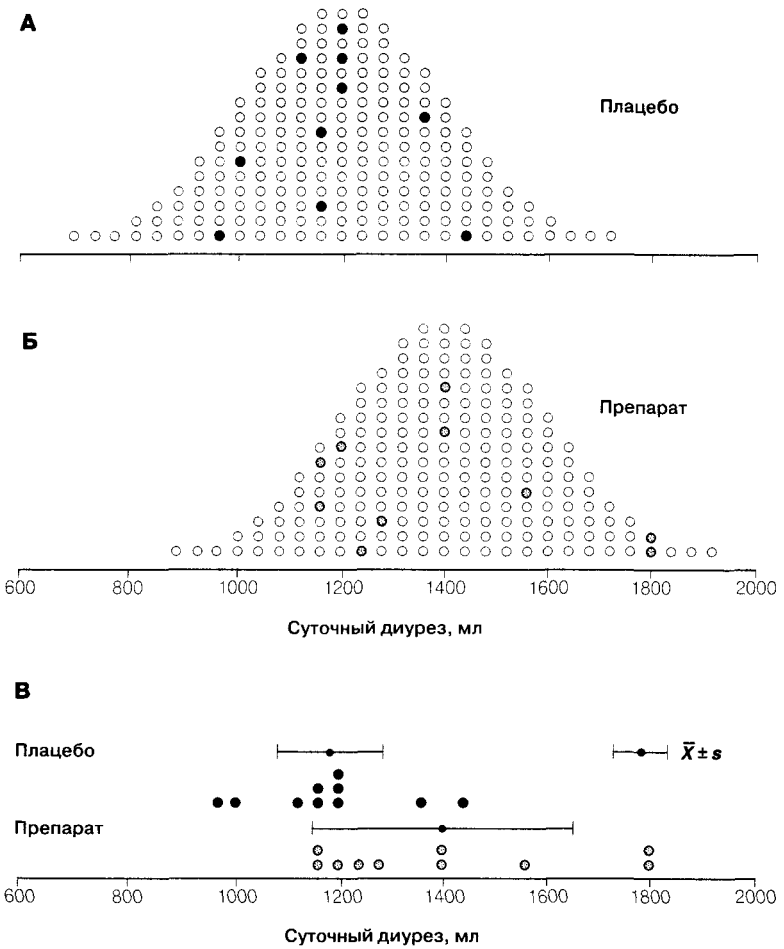


Рис. 6.1. Исследование диуретического эффекта нового препарата. **А.** Суточный диурез в совокупности из 200 человек после приема плацебо. Десять человек, попавшие в выборку, помечены черным. **Б.** Суточный диурез в той же совокупности после приема препарата. Средний диурез увеличился на 200 мл. Десять человек, попавшие в выборку, помечены штриховкой. **В.** Такими видит данные исследователь; $t = 2{,}447$. Это больше критического значения t для 18 степеней свободы (2,101) и 5% уровня значимости, поэтому можно заключить, что различия статистически значимы, то есть препарат обладает диуретическим действием.

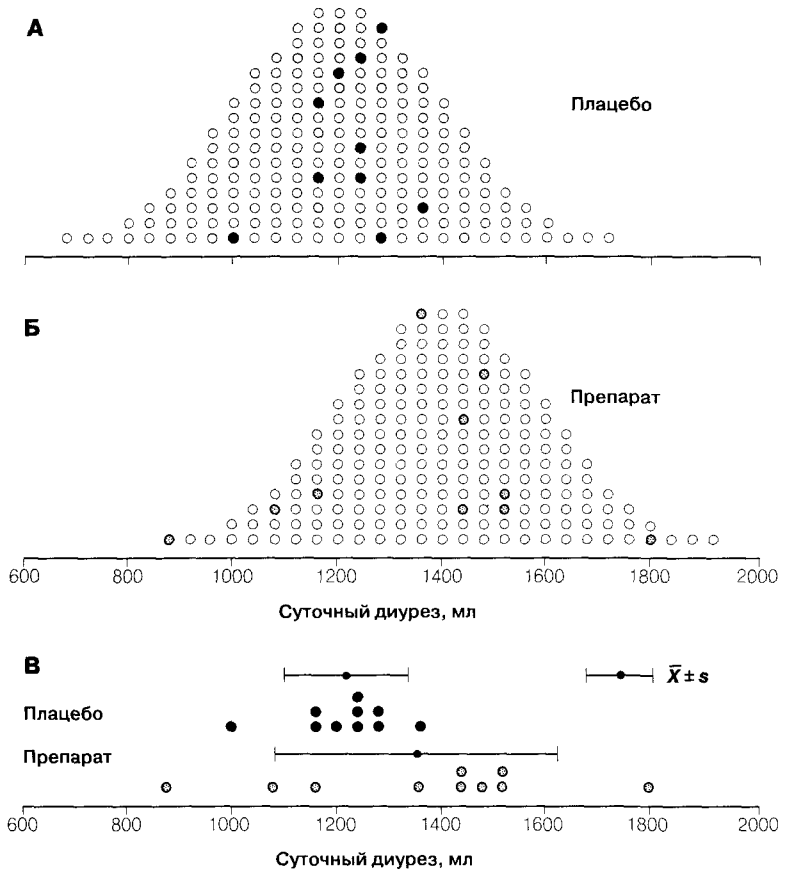


Рис. 6.2. А и Б. Та же совокупность, что и на рис. 6.1, но в выборку попали другие люди. **В.** Изменился и результат, который наблюдает исследователь. Теперь $t = 1{,}71$, что меньше критического значения. В данном случае исследователю не повезло — ему придется признать, что статистически значимых различий не выявлено, то есть диуретическое действие препарата не доказано, — тогда как в действительности оно есть.

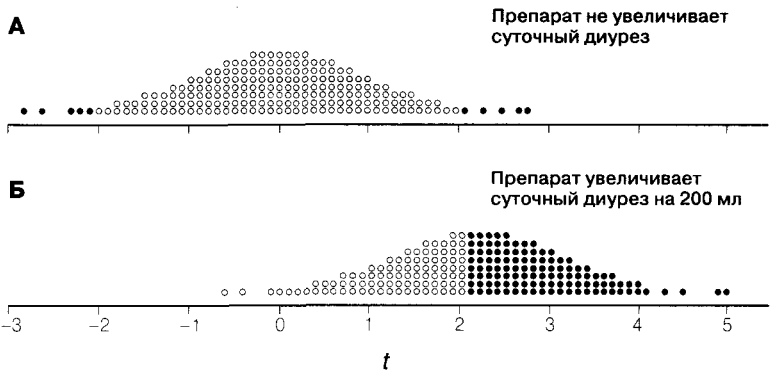


Рис. 6.3. А. Такое распределение мы получим, извлекая пары случайных выборок по 10 человек в каждой из *одной и той же совокупности* и каждый раз вычисляя t (см. рис. 4.5А). Только 5% значений по абсолютной величине превышают 2,1 (помечены черным). Таким образом, 2,1 — критическое значение для 5% уровня значимости. **Б.** Теперь будем извлекать пары выборок из *разных* совокупностей, средний диурез в которых различается на 200 мл (рис. 6.1А и Б). Распределение значений t сместилось вправо. Критическое значение превышено в 111 случаях из 200. Следовательно, вероятность получить правильное заключение об эффективности препарата составляет 55%.

что превышает 2,101 — критическое значение при уровне значимости 0,05 и числе степеней свободы $2(n-1)=18$. Поэтому нулевая гипотеза будет отклонена, а препарат будет назван эффективным диуретиком. Как это и есть на самом деле.

Конечно, исследователь мог бы набрать и другие две группы, например представленные на рис. 6.2. На этот раз средний суточный диурез — 1216 мл в контрольной группе и 1368 мл в группе, получавшей препарат. Стандартное отклонение составляет соответственно 97 и 263 мл, а объединенная оценка дисперсии $1/2(97^2+263^2)=198^2$. Теперь значение t:

$$t=\frac{1368-1216}{\sqrt{\frac{198^2}{10}+\frac{198^2}{10}}}=1{,}71,$$

что меньше 2,101. Нулевую гипотезу отклонить нельзя, хотя мы-то знаем, что она неверна! Какова вероятность такой ситуации?

Для ответа на этот вопрос повторим мысленные эксперименты, подобные тем, что мы проделали в гл. 4 (см. рис. 4.5). Тогда мы строили распределение величины *t* для случая, когда сравниваемые группы представляли собой случайные выборки из одной и той же совокупности. Это распределение показано на рис. 6.3А. Теперь построим распределение *t* для случая, когда выборки извлекаются из *разных* совокупностей. Из двух совокупностей, показанных на рис. 6.2, можно извлечь более 10^{27} выборок объемом в 10 человек; ограничимся пока двумястами. Результат показан на рис. 6.3Б. В 111 случаях из 200 значение *t* оказалось не меньше критического значения 2,101. Итак, в этом случае (то есть при *этих* величине эффекта, дисперсии и численности групп) вероятность отклонить нулевую гипотезу (то есть найти различие) составляет 111/200 = 0,55. Можно оценить и вероятность *не отклонить нулевую гипотезу* (то есть не найти существующих различий). Это 1 – 0,55 = 0,45, то есть 45%. Как видим, шансы обнаружить и не обнаружить диуретический эффект были примерно равны.

ДВА РОДА ОШИБОК

В медицине для характеристики диагностических проб часто используют два показателя: чувствительность и специфичность. Чувствительность — это вероятность положительного результата у больного; она характеризует способность пробы выявлять болезнь. Специфичность — это вероятность отрицательного результата у здорового; можно сказать, что она характеризует способность пробы выявлять отсутствие болезни.

Диагностические пробы и критерии значимости во многом схожи. Диагностические пробы выявляют болезни, критерии значимости выявляют различия. Можно сказать, что с третьей главы по пятую мы занимались специфичностью критериев значимости. В этой главе мы рассматриваем чувствительность, то есть способность критерия выявлять различия. Иногда свойства критериев значимости описывают в несколько иных терминах: не вероятностью правильного результата, а вероятностью ошибки.

Если мы ошибочно отклоняем нулевую гипотезу, то есть на-

Таблица 6.1. Ошибки критериев значимости

	В действительности	
По результатам применения критерия	Различия есть	Различий нет
Различия выявлены	Истинноположительный результат, $1-\beta$	Ложноположительный результат (ошибка I рода), α
Различий не выявлено	Ложноотрицательный результат (ошибка II рода), β	Истинноотрицательный результат, $1-\alpha$

ходим различия там, где их нет, то это называется *ошибкой I рода*. Максимальная приемлемая вероятность ошибки I рода называется *уровнем значимости* и обозначается α. С этой величиной мы уже много раз встречались; обычно α принимают равной 0,05 (то есть 5%), однако можно взять и какой-нибудь другой уровень значимости, например 0,1 или 0,01.

Если мы не отклоняем нулевую гипотезу, когда она не верна, то есть не находим различий там, где они есть, то это — *ошибка II рода*. Ее вероятность обозначается β. Ясно, что вероятность обнаружить различия, то есть *чувствительность* критерия, равна $1-\beta$. В нашем примере с диуретиком $\beta = 0{,}45$ и $1-\beta = 0{,}55$, то есть чувствительность критерия при данных условиях составляет 55%.

Все, что мы узнали об ошибках критериев значимости, кратко представлено в таблице 6.1.

ЧЕМ ОПРЕДЕЛЯЕТСЯ ЧУВСТВИТЕЛЬНОСТЬ?

Естественно, мы заинтересованы в том, чтобы по возможности уменьшить вероятность ошибки II рода, то есть повысить чувствительность критерия. Для этого нужно знать, от чего она зависит. В принципе, эта задача похожа на ту, что решалась применительно к ошибкам I рода, но за одним важным исключением. Чтобы оценить чувствительность критерия, нужно задать *величину различий*, которую он должен выявлять. Эта величина определяется задачами исследования. В примере с диуретиком чувствительность была невелика — 55%. Но, может быть, исследова-

тель просто не считал нужным выявлять прирост диуреза с 1200 до 1400 мл/сут, то есть всего на 17%?

С увеличением *разброса* данных повышается вероятность ошибок обоих типов. Как мы вскоре увидим, величину различий и разброс данных удобнее учитывать совместно, рассчитав отношение величины различий к стандартному отклонению.

Чувствительность диагностической пробы можно повысить, снизив ее специфичность — аналогичное соотношение существует между уровнем значимости и чувствительностью критерия. Чем выше *уровень значимости* (то есть чем меньше α), тем ниже чувствительность.

Как мы уже говорили, важнейший фактор, который влияет на вероятность ошибок как I, так и II рода, — это *объем выборок*. С ростом объема выборок вероятность ошибок уменьшается. Практически это очень важно, поскольку прямо связано с планированием эксперимента.

Прежде чем перейти к подробному рассмотрению факторов, влияющих на чувствительность критерия, перечислим их еще раз.

- Уровень значимости α. Чем меньше α, тем ниже чувствительность.
- Отношение величины различий к стандартному отклонению. Чем больше это отношение, тем чувствительнее критерий.
- Объем выборок. Чем больше объем, тем выше чувствительность критерия.

Уровень значимости

Чтобы получить наглядное представление о связи чувствительности критерия с уровнем значимости, вернемся к рис. 6.3. Выбирая уровень значимости α, мы тем самым задаем критическое значение t. Это значение мы выбираем так, чтобы доля превосходящих его значений — *при условии, что препарат не оказывает эффекта,* — была равна α (рис. 6.3А). Чувствительность критерия есть доля тех значений критерия, которые превосходят критическое *при условии, что лечение дает эффект* (рис.6.3Б). Как видно из рисунка, если изменить критическое значение, изменится и эта доля.

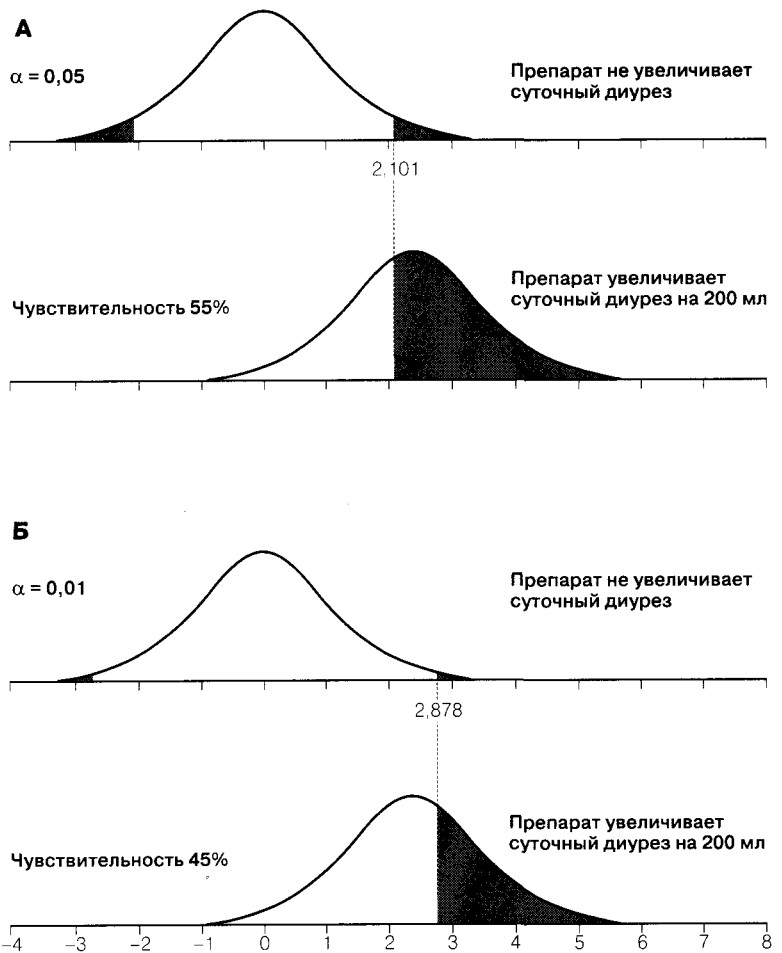


Рис. 6.4. Выбирая уровень значимости α, мы тем самым определяем критический уровень *t*. Чем меньше α, тем выше критический уровень и тем ниже чувствительность. **А.** Уровень значимости α = 0,05, критическое значение *t* = 2,101, чувствительность 55%. **Б.** Теперь уровень значимости α = 0,01, критическое значение *t* возросло до 2,878 и чувствительность снизилась до 45%.

Рассмотрим подробнее, как это происходит. На рис. 6.4А изображено распределение значений критерия Стьюдента. Отличие от рис. 6.3 состоит в том, что теперь это распределение, полученное для всех 10^{27} возможных пар выборок. Верхний график – это распределение значений t для случая, когда препарат *не обладает* диуретическим действием. Предположим, мы выбрали уровень значимости 0,05, то есть приняли $\alpha = 0{,}05$. В этом случае критическое значение равно 2,101, то есть мы отвергаем нулевую гипотезу и признаем различия статистически значимыми при $t > +2{,}101$ или $t < -2{,}101$. Соответствующие области на графике заштрихованы, а критическое значение изображено вертикальной пунктирной линией, спускающейся к нижнему графику, на котором изображено распределение t для случая, когда препарат *обладает* диуретическим действием, а именно увеличивает суточный диурез на 200 мл. По форме нижний график такой же, как верхний, но сдвинут на 200 мл вправо. Доля значений t, превышающих критическое значение 2,101 (заштрихованная область), составляет 0,55. Итак чувствительность критерия в данном случае 55%; а вероятность ошибки второго рода $\beta = 1 - 0{,}55 = 0{,}45$, то есть 45%.

А теперь взглянем на рис. 6.4Б. На нем изображены те же самые распределения значений t. Отличие в выбранном уровне значимости — $\alpha = 0{,}01$. Критическое значение t повысилось до 2,878, пунктирная линия сместилась вправо и отсекает от нижнего графика только 45%. Таким образом, при переходе от 5% к 1% уровню значимости чувствительность снизилась с 55 до 45%. Соответственно, вероятность ошибки II рода повысилась до $1 - 0{,}45 = 0{,}55$.

Итак, снижая α, мы снижаем риск отвергнуть верную нулевую гипотезу, то есть найти различия (эффект) там, где их нет. Но тем самым мы снижаем и чувствительность — вероятность выявить имеющиеся на самом деле различия.

Величина различий

Рассматривая влияние уровня значимости, мы принимали величину различий постоянной: наш препарат увеличивал суточный диурез с 1200 до 1400 мл, то есть на 200 мл. Теперь примем

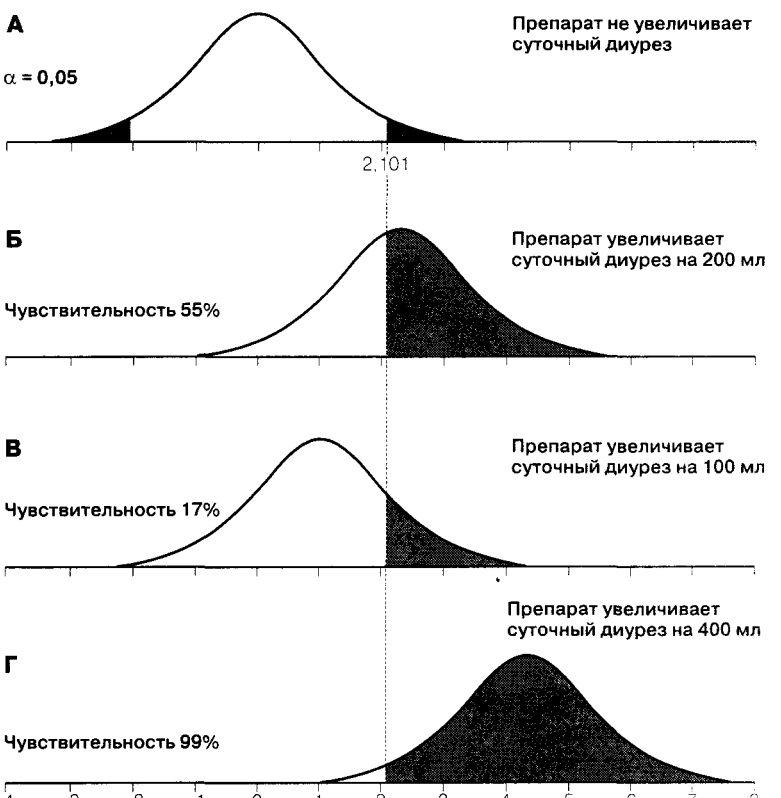


Рис. 6.5. *Чем больше величина различий, тем сильнее распределение t сдвигается вправо и тем выше чувствительность.*

постоянным уровень значимости $\alpha = 0{,}05$ и посмотрим, как чувствительность критерия зависит от величины различий. Понятно, что большие различия выявить легче, чем маленькие. Рассмотрим следующие примеры. На рис. 6.5А изображено распределение значений *t* для случая, когда исследуемый препарат не обладает диуретическим действием. Заштрихованы 5% наибольших по абсолютной величине значений *t*, расположенных левее –2,101 или правее +2,101. На рис. 6.5Б изображено распределение значений *t* для случая, когда препарат увеличивает суточный

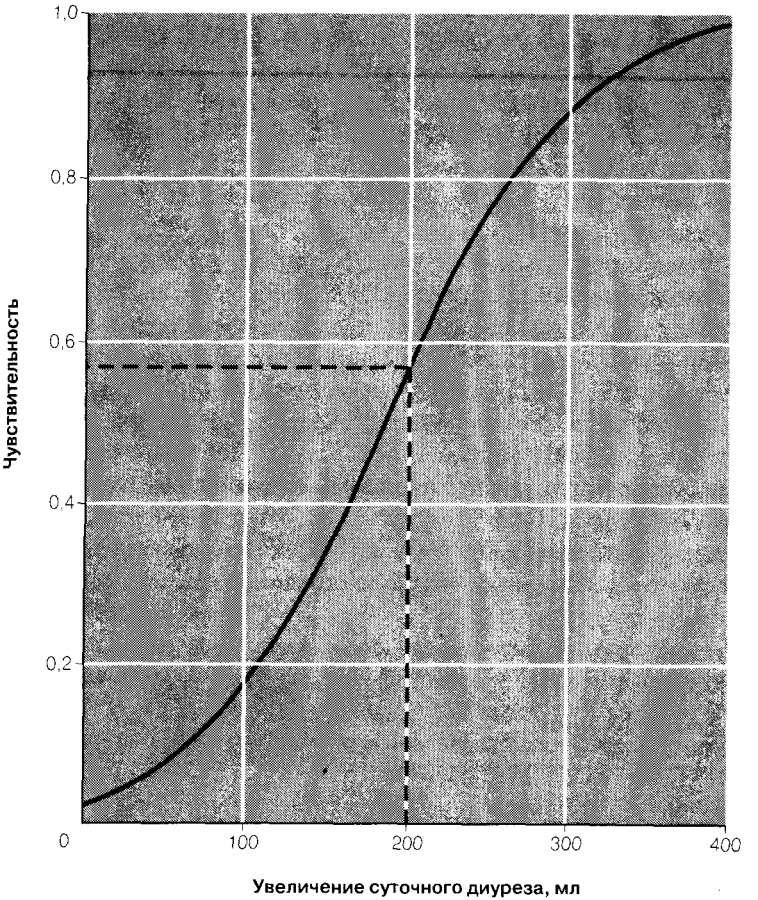


Рис. 6.6. Чувствительность критерия Стьюдента как функция от величины различий при объеме выборок 10 человек и уровне значимости $\alpha = 0{,}05$. Пунктирная линия показывает, как пользоваться графиком. Для величины различий 200 мл чувствительность составляет 0,55.

диурез в среднем на 200 мл (эту ситуацию мы уже рассматривали). Выше правого критического значения лежит 55% возможных значений *t*: чувствительность равна 0,55. Далее, на рис. 6.5В представлено распределение значений *t* для случая, когда препарат увеличивает диурез в среднем на 100 мл. Теперь только 17% значений *t* превышает 2,101. Тем самым, чувствительность критерия равна лишь 0,17. Иными словами, эффект будет обнаружен менее чем в одном из каждых пяти сравнений контрольной и экспериментальной групп. Наконец, рис. 6.5Г представляет случай увеличения диуреза на 400 мл. В критическую область попало 99% значений *t*. Чувствительность критерия равна 0,99: различия будут выявлены почти наверняка.

Повторяя этот мысленный эксперимент, можно определить чувствительность критерия для всех возможных значений эффекта, от нулевого до «бесконечного». Нанеся результаты на график, мы получим рис. 6.6, где чувствительность критерия показана как функция от величины различий. По этому графику можно определить, какой будет чувствительность при той или иной величине эффекта. Пользоваться графиком пока что не очень удобно, ведь он годится только для этих численности групп, стандартного отклонения и уровня значимости. Вскоре мы построим другой график, более подходящий для планирования исследования, но сначала нужно подробнее разобраться с ролью разброса значений и численности групп.

Разброс значений

Чувствительность критерия возрастает с ростом наблюдаемых различий; с ростом разброса значений чувствительность, напротив, снижается.

Напомним, что критерий Стьюдента *t* определяется следующим образом:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}},$$

где $\overline{X}_1$ и $\overline{X}_2$ — средние, s — объединенная оценка стандартного

отклонения σ, n_1 и n_2 — объемы выборок. Заметьте, что $\bar{X}_1$ и $\bar{X}_2$ — это оценки двух (различных) средних — μ_1 и μ_2. Для простоты допустим, что объемы обеих выборок равны, то есть $n_1 = n_2$. Тогда вычисленное значение t есть оценка величины

$$t' = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}}} = \frac{\mu_1 - \mu_2}{\sigma\sqrt{\frac{2}{n}}}.$$

Обозначим δ (греческая буква «дельта») величину эффекта, то есть разность средних: $\delta = \mu_1 - \mu_2$, тогда

$$t' = \frac{\delta}{\sigma\sqrt{\frac{2}{n}}} = \frac{\delta}{\sigma}\sqrt{\frac{n}{2}}.$$

Таким образом, t' зависит от отношения величины эффекта к стандартному отклонению.

Рассмотрим несколько примеров. Стандартное отклонение в исследуемой нами совокупности составляет 200 мл (см. рис. 6.1). В таком случае увеличение суточного диуреза на 200 или 400 мл равно соответственно одному или двум стандартным отклонениям. Это очень заметные изменения. Если бы стандартное отклонение равнялось 50 мл, то те же самые изменения диуреза были бы еще более значительными, составляя соответственно 4 и 8 стандартных отклонений. Наоборот, если бы стандартное отклонение равнялось, например, 500 мл, то изменение диуреза в 200 мл составило бы 0,4 стандартного отклонения. Обнаружить такой эффект было бы непросто да и вряд ли вообще стоило бы.

Итак, на чувствительность критерия влияет не абсолютная величина эффекта, а ее отношение к стандартному отклонению. Обозначим его φ (греческая «фи»); это отношение $\varphi = \delta/\sigma$ называется *параметром нецентральности*.

Объем выборки

Мы узнали о двух факторах, которые влияют на чувствительность критерия: уровень значимости α и параметр нецентральности φ. Чем больше α и чем больше φ, тем больше чувстви-

тельность. К сожалению, влиять на φ мы не можем вовсе, а что касается α, то его увеличение повышает риск отвергнуть верную нулевую гипотезу, то есть найти различия там, где их нет. Однако есть еще один фактор, который мы можем, в определенных пределах, менять по своему усмотрению, не жертвуя уровнем значимости. Речь идет об объеме выборок (численности групп). *С увеличением объема выборки чувствительность критерия увеличивается.*

Существуют две причины, в силу которых увеличение объема выборки увеличивает чувствительность критерия. Во-первых, увеличение объема выборки увеличивает число степеней свободы, что, в свою очередь, уменьшает критическое значение. Во-вторых, как видно из только что полученной формулы

$$t' = \frac{\delta}{\sigma}\sqrt{\frac{n}{2}},$$

значение t растет с ростом объема выборки n (это справедливо и для многих других критериев).

На рис 6.7А воспроизведены распределения с рис. 6.4А. Верхний график соответствует случаю, когда препарат не обладает диуретическим действием, нижний — когда препарат увеличивает суточный диурез на 200 мл. Численность каждой из групп составляет 10 человек. На рис 6.7Б приведены аналогичные распределения. Отличие в том, что теперь в каждую группу входило не 10, а 20 человек. Раз объем каждой из групп равен 20, число степеней свободы равно $\nu = 2(20-1) = 38$. Из таблицы 4.1 находим, что критическое значение t при 5% уровне значимости равно 2,024 (в случае выборок объемом 10 оно равнялось 2,101). С другой стороны, увеличение объема выборок привело к увеличению значений критерия. В результате уже не 55, а 87% значений t превышают критическое значение. Итак, увеличение численности групп с 10 до 20 человек привело к повышению чувствительности с 0,55 до 0,87.

Перебирая все возможные объемы выборок, можно построить график чувствительности критерия как функции от численности групп (рис. 6.8). С увеличением объема чувствительность

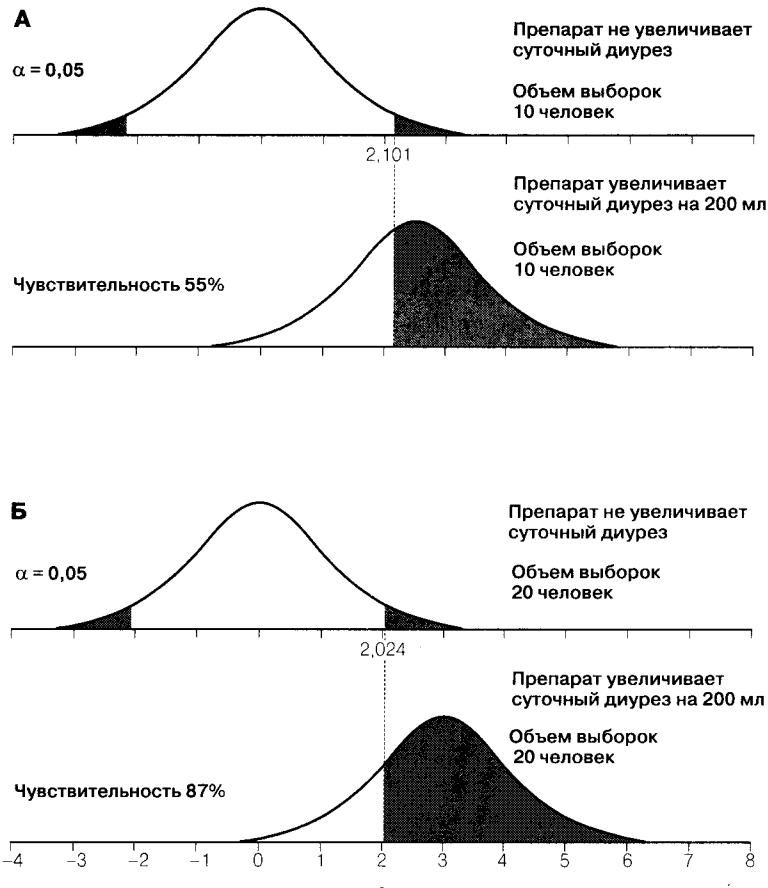


Рис. 6.7. Увеличение объема выборки повышает чувствительность по двум причинам. Во-первых, увеличивается число степеней свободы, и критическое значение *t* уменьшается. Во-вторых, при той же величине различий получаются более высокие значения *t*.

растет. Сначала она растет ускоренно, затем, начиная с некоторого объема выборки, рост замедляется.

Расчет чувствительности — важнейшая составная часть планирования медицинских исследований. Теперь, познакомившись с наиболее важным фактором, определяющим чувствительность, мы готовы решить эту задачу.

Как определить чувствительность критерия?

На рис. 6.9 чувствительность критерия Стьюдента представлена как функция от параметра нецентральности $\varphi = \delta/\sigma$ при уровне значимости $\alpha = 0{,}05$. Четыре кривые соответствуют четырем объемам выборок.

Подразумевается, что выборки имеют равный объем. Что делать, если это не так? Если вы обратились к рис. 6.9 при *планировании* исследования (что весьма разумно), то нужно учесть следующее. При заданной общей численности обследованных именно равная численность групп обеспечивает максимальную чувствительность. Значит, равную численность групп и следует запланировать. Если же вы решили рассчитать чувствительность *после* проведения исследования, когда, не найдя статистически значимых различий, вы хотите определить, в какой степени это можно считать доказательством отсутствия эффекта, — тогда следует принять численность обеих групп равной меньшей из них. Такой расчет даст несколько заниженную оценку чувствительности, но убережет вас от излишнего оптимизма.

Применим кривые с рис. 6.9 к примеру с диуретиком (см. рис. 6.1). Мы хотим вычислить чувствительность критерия Стьюдента при уровне значимости $\alpha = 0{,}05$. Стандартное отклонение равно 200 мл. Какова вероятность выявить увеличение суточного диуреза на 200 мл?

$$\varphi = \frac{\delta}{\sigma} = \frac{200}{200} = 1.$$

Численность контрольной и экспериментальной групп равна десяти. Выбираем на рис. 6.9 соответствующую кривую и находим, что чувствительность критерия равна 0,55.

До сих пор мы говорили о чувствительности критерия Стью-

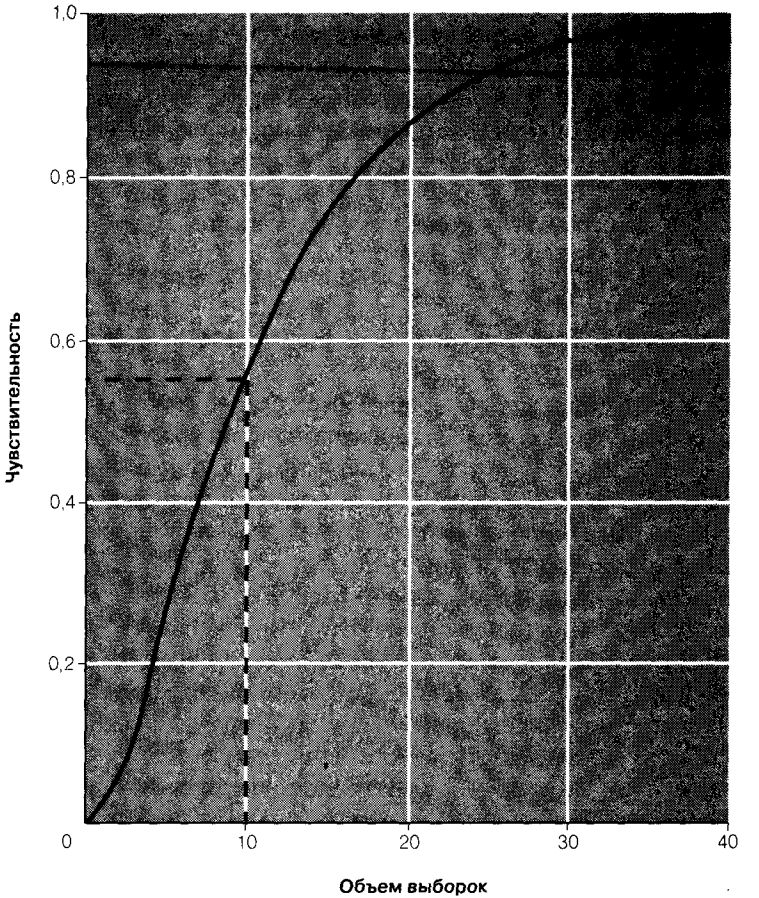


Рис. 6.8. Чувствительность критерия Стьюдента как функция от объема выборок при величине различий 200 мл, уровне значимости $\alpha = 0{,}05$ и стандартном отклонении $\sigma = 200$ мл. При объеме выборок 10 человек чувствительность составляет 0,55.

дента. Можно рассчитать чувствительность и других критериев. Определяется она теми же самыми факторами, но ход вычислений будет несколько иным.

Галотан и морфин при операциях на открытом сердце

В гл. 4 мы сравнили сердечный индекс при галотановой и морфиновой анестезии (см. табл. 4.2) и не нашли статистически значимых различий. (Напомним, что сердечный индекс — это отношение минутного объема сердца к площади поверхности тела.) Однако группы были малы — 9 и 16 человек. Средняя величина сердечного индекса в группе галотана равнялась 2,08 л/мин/м^2; в группе морфина 1,75 л/мин/м^2, то есть на 16% меньше. Даже если бы различия были статистически значимыми, вряд ли столь небольшая разница представляла бы какой-либо практический интерес.

Поэтому поставим вопрос так: какова была вероятность выявить разницу в 25%? Объединенная оценка дисперсии $s^2 = 0{,}89$, значит, стандартное отклонение равно 0,94 л/мин/м^2. Двадцать пять процентов от 2,08 л/мин/м^2 — это 0,52 л/мин/м^2.

Тем самым,

$$\varphi = \frac{\delta}{\sigma} = \frac{0{,}52}{0{,}94} = 0{,}553.$$

Поскольку численности групп не совпадают, для оценки чувствительности выберем меньшую из них — 9. Из рис. 6.9 следует, что в таком случае чувствительность критерия — 0,16. Шансы выявить даже 25% различия были весьма малы.

Подведем итоги.

- Чувствительность критерия есть вероятность отвергнуть ложную гипотезу об отсутствии различий.
- На чувствительность критерия влияет уровень значимости: чем меньше α, тем ниже чувствительность.
- Чем больше величина эффекта, тем больше чувствительность.
- Чем больше объем выборки, тем больше чувствительность.
- Для разных критериев чувствительность вычисляется по-разному.

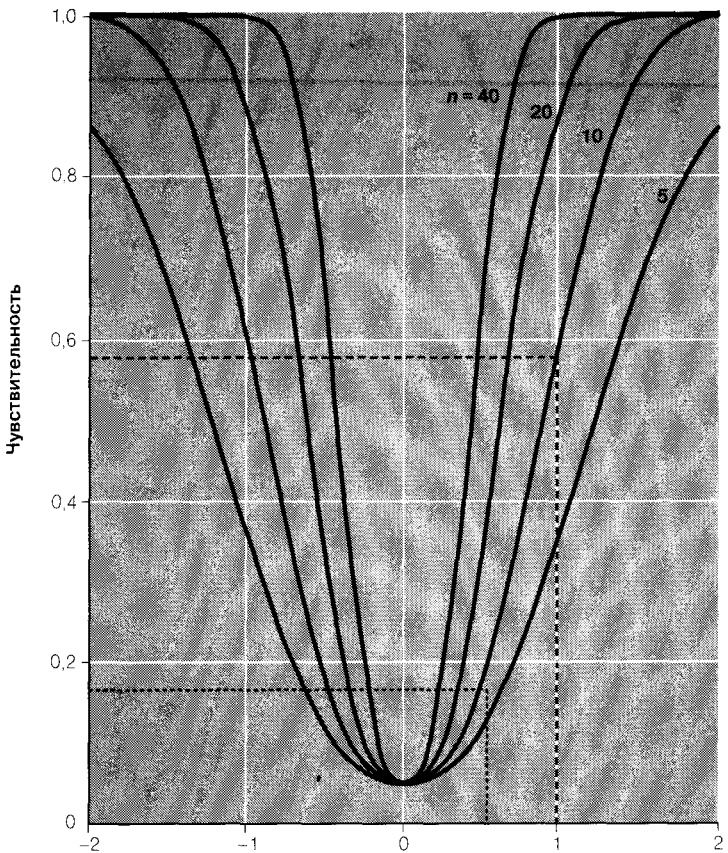


Рис. 6.9. Чувствительность критерия Стьюдента как функция от параметра нецентральности φ при уровне значимости $\alpha = 0{,}05$ для разных объемов выборок *n*. Параметр нецентральности — это отношение величины различий к стандартному отклонению в совокупности: $\varphi = \delta/\sigma$. Пунктирные линии показывают, как пользоваться графиками. Если, например, величина различий $\delta = 200$ мл, стандартное отклонение $\sigma = 200$ мл, то $\varphi = 1$. Для объема выборок $n = 10$ чувствительность составляет 0,55. При $\varphi = 0{,}55$ и $n = 9$ чувствительность — всего лишь 0,16.

ЧУВСТВИТЕЛЬНОСТЬ ДИСПЕРСИОННОГО АНАЛИЗА

Чувствительность дисперсионного анализа* определяется теми же факторами, что чувствительность критерия Стьюдента, похож и способ ее вычисления. Для расчета нам понадобятся следующие данные: число групп, их численность, уровень значимости и величина различий. Что понимать под величиной различий, если число групп больше двух? В качестве величины различий δ используют минимальную величину различий между любыми двумя группами. Параметр нецентральности рассчитывают по формуле:

$$\varphi = \frac{\delta}{\sigma}\sqrt{\frac{n}{2k}},$$

где σ — стандартное отклонение в совокупности, k — число групп, n — численность каждой из них**. Есть другой способ, несколько более сложный. Если μ_i — среднее в i-й группе, то

$$\varphi = \sqrt{\frac{\Sigma(\mu_i - \mu)^2}{k\sigma^2}},$$

где

$$\mu = \frac{\Sigma\mu_i}{k}$$

есть среднее по всем группам.

Определив параметр нецентральности и зная межгрупповое число степеней свободы $\nu_{меж} = k - 1$, чувствительность находят по графикам, где она представлена как функция от параметра нецентральности. На рис. 6.10 изображены графики для $\nu_{меж} = 2$, графики для других значений $\nu_{меж}$ вы найдете в приложении Б.

---

* Во вводном курсе этот раздел можно пропустить без ущерба для понимания последующего материала.

** Численность групп предполагается равной. Как и в случае критерия Стьюдента, именно равная численность групп обеспечивает максимальную чувствительность при заданной общей численности обследованных.

Те же графики можно использовать и для определения численности групп, обеспечивающей необходимую чувствительность. Это сложнее, чем в случае критерия Стьюдента, так как теперь n входит и в параметр нецентральности φ, и в выражение для числа степеней свободы $\nu_{\text{вну}}$. Поэтому значение n приходится подбирать путем последовательного приближения. Сначала вы произвольно выбираете начальное значение n и вычисляете чувствительность. В зависимости от найденного значения чувствительности вы изменяете n, после чего повторяете вычисление. Эта процедура повторяется до тех пор, пока значение чувствительности не окажется достаточно близким к нужному.

Бег и менструации

Чтобы получше разобраться с тем, как вычислить чувствительность и объем выборки при дисперсионном анализе, обратимся к примеру с влиянием бега на частоту менструаций, который мы разбирали в гл. 3 (рис. 3.9). Сейчас нас интересует, какова вероятность выявить различие в одну менструацию в год ($\delta = 1$). Число групп $k = 3$; стандартное отклонение $\sigma = 2$. Численность каждой из групп $n = 26$. Уровень значимости выбираем: $\alpha = 0{,}05$. Найдем параметр нецентральности:

$$\varphi = \frac{1}{2}\sqrt{\frac{26}{2 \times 3}} = 1{,}04.$$

Межгрупповое число степеней свободы $\nu_{\text{меж}} = k - 1 = 3 - 1 = 2$ и внутригрупповое $\nu_{\text{вну}} = k(n-1) = 3(26-1) = 75$. По рис. 6.10 находим, что чувствительность составит около 0,30.

Результат обескураживающий, что вообще характерно для расчетов чувствительности. Положим, нам хотелось бы иметь чувствительность равной 0,80. Какая численность групп нужна для этого? В том, что объем $n = 26$ слишком мал, мы только что убедились. Из рис. 6.10 мы видим, что параметр нецентральности должен быть приблизительно равен 2. Для $n = 26$ он близок к 1. Значит, численность групп должна быть такой, чтобы параметр нецентральности увеличился вдвое. При вычислении φ из численности групп n извлекается квадратный корень, поэтому чис-

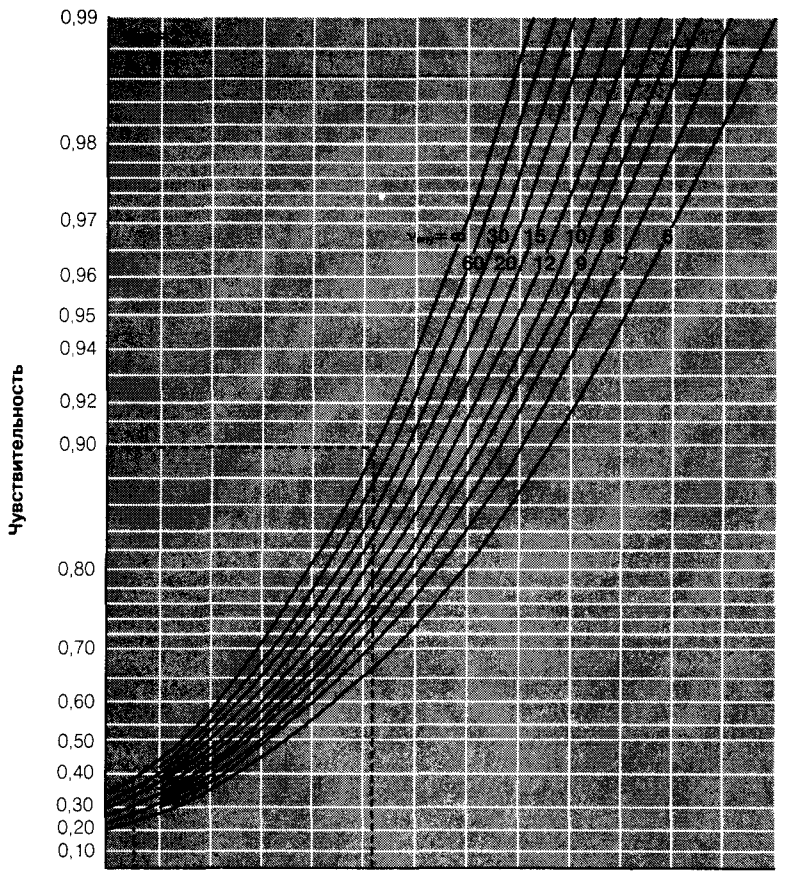


Рис. 6.10. Чувствительность дисперсионного анализа как функция от параметра нецентральности φ при уровне значимости $\alpha = 0{,}05$ и межгрупповом числе степеней свободы $\nu_{меж} = 2$. В приложении Б вы найдете аналогичные графики для других значений α и $\nu_{меж}$.

E. S. Pearson, H. O. Hartley. Charts for the power function for analysis of variance tests, derived from the non-central F distribution. *Biometrika*, 38:112–130, 1951.

ленность групп должна увеличиться в $2^2 = 4$ раза. Таким образом, нужно, чтобы в каждую из групп входило по 100 человек. Тогда

$$\varphi = \frac{1}{2}\sqrt{\frac{100}{2\times 3}} = 2{,}04$$

и $\nu_{\text{вну}} = k(n-1) = 3(100-1) = 297$. По рис. 6.10 находим, что в этом случае чувствительность составит 0,88, то есть даже больше, чем мы хотели. Поскольку стандартное отклонение может оказаться больше, чем мы думали, некоторый избыток чувствительности нам не помешает, однако резонно спросить, где же и на какие средства мы наберем такие группы. Нельзя ли хоть немного сократить их численность? Попробуем $n = 75$. Тогда

$$\varphi = \frac{1}{2}\sqrt{\frac{75}{2\times 3}} = 1{,}77$$

и $\nu_{\text{вну}} = 3(75-1) = 222$. Рис. 6.10 показывает, что теперь чувствительность равна 0,80.

Таким образом, для того чтобы при уровне значимости $\alpha = 0{,}05$ с вероятностью 80% обнаружить в трех группах различие в одну менструацию в год, когда стандартное отклонение предположительно составляет 2 менструации в год, нужно набрать группы по 75 человек.

ЧУВСТВИТЕЛЬНОСТЬ ТАБЛИЦ СОПРЯЖЕННОСТИ*

Графиками с рис. 6.10 (и из приложения Б) можно воспользоваться для нахождения чувствительности и объема выборки при работе с таблицами сопряженности**. Сначала нужно решить, какое минимальное различие вы хотели бы обнаружить. В случае таблиц сопряженности это означает, что вам нужно заполнить клетки не-

---

* Во вводном курсе этот раздел можно опустить.

** Таблицу сопряженности 2×2 можно рассматривать как задачу сравнения двух долей. Как в этом случае вычислить чувствительность и объем выборки, вы поймете, решив задачу 6.6. Более подробно этот вопрос изложен в работе: A. F. Feinstein. Clinical biostatistics. Mosby, St. Louis, 1977.

Таблица 6.2. Обозначения, используемые при вычислении чувствительности критерия χ^2

p_{11}	p_{12}	R_1
p_{21}	p_{22}	R_2
p_{31}	p_{32}	R_3
C_1	C_2	1,00

которыми долями. В таблице 6.2 приведены обозначения, используемые при вычислении чувствительности таблицы сопряженности, для примера взята таблица 3×2. Здесь p_{ij} – доля в i-й строке j-го столбца, например p_{11} – доля всех наблюдений в левой верхней клетке, p_{12} — доля наблюдений в правой верхней клетке, и так далее. Сумма всех долей составляет 1. Суммы по строкам обозначаются R_i, по столбцам — C_j. Параметр нецентральности задается формулой

$$\varphi = \sqrt{\frac{N}{(r-1)(c-1)+1}\sum\frac{(p_{ij}-R_iC_j)^2}{R_iC_j}},$$

где r — число строк, c — число столбцов и N — общее число наблюдений. Зная значение φ и число степеней свободы $\nu_{\text{вну}} = \infty$ и $\nu_{\text{меж}} = (r-1)(c-1)$, чувствительность можно определить по кривым с рис. 6.10.

Для нахождения объема выборки, при котором достигается требуемая чувствительность, воспользуемся обратной процедурой. Именно, сначала по рис. 6.10 найдем значение параметра нецентральности φ для заданной чувствительности и числа степеней свободы $\nu_{\text{меж}} = (r-1)(c-1)$ и $\nu_{\text{вну}} = \infty$. А теперь найдем объем выборки, разрешив приведенную выше формулу относительно N:

$$N = \frac{\varphi^2[(r-1)(c-1)+1]}{\sum\frac{(p_{ij}-R_iC_j)^2}{R_iC_j}}.$$

Бег и менструации

Дейл и соавт. изучали не только то, как занятия бегом влияют на частоту менструаций, но и то, какая доля женщин обращалась к

врачу. (Этот пример мы подробно рассмотрели в гл. 5, см. табл. 5.5.) Допустим, мы хотим выявить различия не меньшие, чем в табл. 6.3. Уровень значимости $\alpha = 0{,}05$, общее число обследованных $N = 165$. Расссчитаем сначала сумму

$$\sum \frac{(p_{ij} - R_i C_j)^2}{R_i C_j} = \frac{(0{,}025 - 0{,}250 \times 0{,}350)^2}{0{,}250 \times 0{,}350} +$$

$$+ \frac{(0{,}225 - 0{,}250 \times 0{,}650)^2}{0{,}250 \times 0{,}650} + \frac{(0{,}100 - 0{,}300 \times 0{,}350)^2}{0{,}300 \times 0{,}350} +$$

$$+ \frac{(0{,}200 - 0{,}300 \times 0{,}650)^2}{0{,}300 \times 0{,}650} + \frac{(0{,}225 - 0{,}450 \times 0{,}350)^2}{0{,}450 \times 0{,}350} +$$

$$+ \frac{(0{,}225 - 0{,}450 \times 0{,}650)^2}{0{,}450 \times 0{,}650} = 0{,}114.$$

Тогда

$$\varphi = \sqrt{\frac{165}{(3-1)(2-1)+1} 0{,}114} = 2{,}50.$$

По рис 6.10 находим, что для $\varphi = 2{,}50$ при $\nu_{меж} = (r-1)(c-1) = (3-1)(2-1) = 2$ и $\nu_{вну} = \infty$ степенях свободы и уровне значимости $\alpha = 0{,}05$ чувствительность равна 0,98.

ПРАКТИЧЕСКИЕ ТРУДНОСТИ

Нетрудно рассчитать чувствительность критерия задним числом, когда и стандартное отклонение, и величина эффекта уже известны. К сожалению, мы не знаем эти параметры, когда планируем исследование. Стандартное отклонение можно примерно оценить по литературным данным или проведя предварительное исследование. Величину эффекта узнать заранее невозможно (обычно ее оценка и является целью исследования). Поэтому при расчете чувствительности нужно указать *минимальную* величину эффекта, которую мы хотим выявить. Немногие решаются поведать миру о том, какова же эта величина, поэтому чувствительность очень редко рассчитывают заранее. Между тем де-

Таблица 6.3. Предполагаемые доли женщин, обращавшихся к врачу по поводу нерегулярности менструаций

	Обращались к врачу		
Группа	Да	Нет	Всего
Контроль	0,025	0,225	0,250
Физкультурницы	0,100	0,200	0,300
Спортсменки	0,225	0,225	0,450
Всего	0,350	0,650	1,00

лать это совершенно необходимо: иначе мы рискуем проводить исследования, заведомо обреченные на неуспех.

Если после проведения исследования эффект обнаружен, то чувствительность уже неважна. В противном случае — если эффекта не выявлено — она приобретает первостепенное значение. В самом деле, если мы не обнаружили статистически значимых различий при чувствительности 80%, то с высокой вероятностью можно утверждать, что различий действительно нет. Иными словами, мы получили *отрицательный результат*. Если же чувствительность составляла 25%, то мы просто не получили никакого результата. Обычно данные, необходимые для определения чувствительности, содержатся в статье, поэтому читатель может сам провести расчет.

ЗАЧЕМ ВЫЧИСЛЯТЬ ЧУВСТВИТЕЛЬНОСТЬ?

Ранее, в 4 гл., мы разобрали распространенную ошибку, состоящую в многократном применении критерия Стьюдента. В терминах этой главы можно сказать, что многократное применение критерия Стьюдента увеличивает ошибку I рода. На практике же это означает, что нам сообщают о «статистически значимых различиях» там, где их в действительности нет. Теперь, познакомившись с методами определения чувствительности критерия и убедившись, насколько малой она нередко оказывается, мы можем судить о причинах этого явления. Многие исследования не имели бы никаких шансов на успех, если бы завершались одним единственным сравнением. Конечно, проще сравнить группы по целому ряду лабораторных показателей, чем сделать числен-

ность групп достаточной для выявления разницы в летальности. С другой стороны, пренебрежение оценкой чувствительности приводит к тому, что во вполне корректно (в остальном) проведенном исследовании клинически значимый эффект остается невыявленным из-за слишком малой численности групп.

Теперь мы получили достаточное представление о чувствительности, чтобы избежать этих ловушек. Мы узнали о том, как можно оценить чувствительность критерия по данным, приведенным в публикации, и как самому вычислить нужный объем выборок, чтобы обнаружить эффект заданной величины. Результаты таких вычислений часто разочаровывают, поскольку оказывается, что численность групп должна быть огромной (особенно в сравнении с тем обычно небольшим числом больных, которые участвуют в клинических исследованиях)*. Как бы то ни было, мы должны отдавать себе отчет в ограниченности наших возможностей. Однако заведомо несостоятельные исследования все же проводятся. Вряд ли авторы сознательно замалчивают недостаток чувствительности, рассчитывая, что благодаря эффекту множественных сравнений «что-нибудь найдется». На самом деле большинство из них просто никогда ничего не слышали о чувствительности критериев.

Фрейман и соавт.** изучили 71 публикацию*** по результатам контролируемых испытаний, проведенных в 1960—1977 гг., в которых исследуемый метод лечения не дал статистически значимого ($P < 0{,}05$) улучшения исхода. Лишь в 20% работ численность групп была достаточной, чтобы обнаружить снижение частоты неблагоприятных исходов (смерть, осложнение и т. п.) на 25% с

---

* По данным Р. А. и С. У. Флетчеров (R. A. Fletcher, S. W. Fletcher. Clinical research in general medical journals: a 30-year perspective. *N. Engl. J. Med.*, 301:180—183, 1979), изучавших работы, опубликованные в Journal of the American Medical Association, Lancet и New England Journal of Medicine, в период с 1946 по 1976 г. медиана численности группы составляла от 16 до 36 человек.

** J. A. Freiman, T. C. Chalmers, H. Smith Jr., R. R. Kuebler. The importance of beta, the type II error and sample size in the design and interpretation of the randomized controlled trial. *N. Engl. J. Med.*, 299:690—694, 1978.

*** В журналах Lancet, New England Journal of Medicine, Journal of the American Medical Association.

вероятностью 50%. Только в *одной* статье говорилось, что уровень значимости и чувствительность были определены до начала исследования, 14 статей содержали указания на желательность большей численности групп.

Пятнадцать лет спустя аналогичное исследование провели Моэр и соавт., рассмотрев публикации по результатам контролируемых испытаний в тех же журналах за 1990 г. Число публикаций по этой теме по сравнению с 1975 г. возросло вдвое, однако доля отрицательных результатов осталась прежней — около 27%. Доля исследований, обеспечивающих достаточную чувствительность, оказалась примерно той же, что и в работе Фреймана и соавт., однако расчет численности групп обнаружен уже в трети статей. Итак, некоторый прогресс налицо, хотя ситуация все же оставляет желать лучшего. Как и во всем, что касается применения статистических методов, полностью полагаться на авторов пока нельзя. Прежде чем принять вывод о неэффективности того или иного метода лечения, читателю следует самостоятельно оценить чувствительность примененного критерия.

Что же все-таки делать с работами, не обнаружившими эффекта из-за недостаточной численности групп*? Нужно ли мах-

---

* Необходимость заранее определять численность групп ставит исследователей перед нелегким выбором: мириться с высоким риском не получить результат или проводить дорогостоящее широкомасштабное исследование. Эта проблема в значительной мере снимается *методами последовательного анализа*. При последовательном анализе численность групп не определяется заранее: вместо этого больных включают в исследование по одному. Дождавшись наступления того или иного исхода, выбирают одно из трех: 1) принять гипотезу об отсутствии эффекта, 2) отвергнуть гипотезу либо 3) включить еще одного больного. Последовательный анализ обычно обеспечивает те же величины α и β, что и обычные методы, при меньшей численности групп. Применять на каждом шаге критерий Стьюдента было бы неправильно: из-за эффекта множественных сравнений мы получили бы чрезмерно «оптимистическое» значение P. Последовательный анализ требует применения специальных методов оценки статистической значимости, которые изложены в главе «Sequential analysis» книги W. J. Dixon, F. J. Massey. Introduction to Statistical Analysis, McGraw-Hill, New York, 1969.

нуть рукой на полученные результаты или из них можно извлечь нечто полезное? Оказывается, можно. Для этого следует отказаться от альтернативной логики «эффект есть — эффекта нет» и вместо этого оценить величину эффекта и степень неопределенности этой оценки, то есть рассчитать доверительный интервал, чем мы и займемся в следующей главе.

ЗАДАЧИ

6.1. Используя данные табл. 4.2, вычислите чувствительность критерия Стьюдента, способного обнаружить 50% различие наилучшего сердечного индекса между галотановой и морфиновой анестезией.

6.2. По тем же данным определите, какова должна быть численность групп, чтобы с вероятностью 80% обнаружить 25% различие в наилучшем сердечном индексе.

6.3. Используя данные табл. 4.2, определите чувствительность критерия Стьюдента для выявления изменения среднего артериального давления и общего периферического сосудистого сопротивления на 25%.

6.4. В задаче 3.5 мы не обнаружили влияния внутривенного введения тетрагидроканнабинолов на антибактериальную защиту у крыс. Допустим, минимальное снижение, которое мы хотим выявить, составляет 20%, уровень значимости $\alpha = 0{,}05$. Какова чувствительность критерия Стьюдента?

6.5. По тем же данным определите, какой должна быть численность групп, чтобы обеспечить выявление снижения антибактериальной защиты на 20% с вероятностью 90% (уровень значимости $\alpha = 0{,}05$).

6.6. Какой должна быть численность групп, чтобы с вероятностью 90% обнаруживать снижение летальности с 90 до 30%. Уровень значимости $\alpha = 0{,}05$. При решении вам пригодятся табличные значения стандартного нормального распределения (табл. 6.4).

6.7. Используя данные из задачи 3.2, найдите вероятность обнаружить снижение максимальной объемной скорости середины выдоха на 0,25 л/с при уровне значимости $\alpha = 0{,}05$.

Таблица 6.4. Процентили стандартного нормального распределения

Отклонение z от среднего (в стандартных отклонениях)	Площадь слева от z	Площадь справа от z
−2,5	0,0062	0,9938
−2,4	0,0082	0,9918
−2,3	0,0107	0,9893
−2,2	0,0139	0,9861
−2,1	0,0179	0,9821
−2,0	0,0228	0,9772
−1,9	0,0287	0,9713
−1,8	0,0359	0,9641
−1,7	0,0446	0,9554
−1,6	0,0548	0,9452
−1,5	0,0668	0,9332
−1,4	0,0808	0,9192
−1,3	0,0968	0,9032
−1,2	0,1151	0,8849
−1,1	0,1357	0,8643
−1,0	0,1587	0,8413
−0,9	0,1841	0,8159
−0,8	0,2119	0,7881
−0,7	0,2420	0,7580
−0,6	0,2743	0,7267
−0,5	0,3085	0,6975
−0,4	0,3446	0,6554
−0,3	0,3821	0,6179
−0,2	0,4207	0,5793
−0,1	0,4602	0,5398
0,0	0,5000	0,5000
0,1	0,5398	0,4602
0,2	0,5793	0,4207
0,3	0,6179	0,3821
0,4	0,6554	0,3446
0,5	0,6975	0,3085
0,6	0,7267	0,2743
0,7	0,7580	0,2420

Таблица 6.4. Окончание

Отклонение z от среднего (в стандартных отклонениях)	Площадь слева от z	Площадь справа от z
0,8	0,7881	0,2119
0,9	0,8159	0,1841
1,0	0,8413	0,1587
1,1	0,8643	0,1357
1,2	0,8849	0,1151
1,3	0,9032	0,0968
1,4	0,9192	0,0808
1,5	0,9332	0,0668
1,6	0,9452	0,0548
1,7	0,9554	0,0446
1,8	0,9641	0,0359
1,9	0,9713	0,0287
2,0	0,9772	0,0228
2,1	0,9821	0,0179
2,2	0,9861	0,0139
2,3	0,9893	0,0107
2,4	0,9918	0,0082
2,5	0,9938	0,0062

6.8. Используя данные из задачи 3.3, найдите вероятность обнаружить увеличение уровня липопротеидов высокой плотности на 5 и 10 мг%. Уровень значимости $\alpha = 0,05$.

6.9. По тем же данным определите, какой должна быть численность групп, чтобы изменение в 5 мг% можно было обнаружить с вероятностью 80% при уровне значимости $\alpha = 0,05$.

6.10. В задаче 5.4 сравнивали частоту рецидивов инфекции мочевых путей после короткого курса того или иного антибактериального препарата. Допустим, минимальные различия, которые мы хотим выявить, таковы: в группах ампициллина и триметоприма/сульфаметоксазола рецидив наступает у двух третей девочек, в группе цефалексина — у одной трети. Какой была бы чувствительность таблицы сопряженности при численности групп, указанной в задаче 5.4? Уровень значимости $\alpha = 0,05$.

6.11. Каким должен быть объем выборки, чтобы в задаче 6.10 чувствительность составила 80%?

Глава 7

Доверительные интервалы

До сих пор мы занимались в основном нахождением различий между группами, не слишком интересуясь *величиной* этих различий. Мы формулировали нулевую гипотезу, то есть предполагали, что экспериментальные группы — это просто две случайные выборки из одной и той же совокупности. Затем мы оценивали вероятность получить наблюдаемые различия при условии, что нулевая гипотеза верна. Если эта вероятность была мала, мы отвергали нулевую гипотезу и делали вывод, что различия статистически значимы. При таком подходе мы всегда получаем только качественный результат: либо отклоняем нулевую гипотезу, либо не отклоняем, либо признаем различия статистически значимыми, либо не признаем. Количественная оценка различий от нас ускользает. Между тем, как мы выяснили в предыдущей главе, вероятность выявления различий зависит не только от их величины, но и от численности групп. Сколь угодно малые различия при достаточно большой численности групп могут оказаться статистически значимыми, или, как пишут в диссертаци-

ях, «высоко достоверными». При этом речь может идти о разнице в несколько миллиметров ртутного столба.

Характеристика, которая дополняет и даже заменяет качественное суждение (значимо—незначимо), — это *доверительный интервал*. В гл. 2 мы уже встречались с этим понятием, хотя и не применяли этот термин. Тогда мы выяснили, что истинное среднее в 95% случаев лежит на расстоянии не больше двух ошибок среднего от выборочного среднего. Промежуток длиной в четыре ошибки среднего — это и есть 95% доверительный интервал. Смысл доверительного интервала из этого примера достаточно ясен: мы не знаем точно, чему равна некоторая величина, но можем указать интервал, в котором она находится (с заданной вероятностью). В этой главе мы научимся определять доверительные интервалы для разных величин, в том числе для разности средних (величины эффекта) и доли. Мы покажем, что доверительный интервал можно использовать вместо обычных критериев значимости*. Доверительные интервалы используют также для определения границ нормы лабораторного показателя.

ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ ДЛЯ РАЗНОСТИ СРЕДНИХ

В гл. 4 мы определили критерий Стьюдента как

$$t = \frac{\text{Разность выборочных средних}}{\text{Стандартная ошибка разности выборочных средних}}.$$

Вычислив t, его сравнивают с критическим значением t_α для заданного уровня значимости α. Для двух случайных выборок из одной совокупности вероятность получить значение t, по абсолютной величине превышающее t_α, весьма мала (а именно, не превышает α; напомним, что уровень значимости α — это максимальная приемлемая вероятность ошибочно признать существование различий там, где их нет). Поэтому, получив «боль-

---

* Существует мнение, что только доверительные интервалы и нужно использовать. Эта точка зрения кратко изложена в работе: K. J. Rothman. A show of confidence. *N. Engl. J. Med.*, 299:1362—1363, 1978.

шое» значение t, мы делаем вывод о статистической значимости различий.

Для случайных выборок, извлеченных из одной совокупности, распределение всех возможных значений t (распределение Стьюдента) симметрично относительно среднего, равного нулю (см. рис. 4.5). Если же выборки извлечены из двух совокупностей с *разными* средними, то распределение всех возможных значений t будет иметь среднее, отличное от нуля (см. рис. 6.3 и 6.5).

Формулу для t можно видоизменить так, чтобы распределение t было *всегда* симметрично относительно нуля:

$$t = \frac{\text{Разность выборочных средних} - \text{Разность истинных средних}}{\text{Стандартная ошибка разности выборочных средних}}.$$

Заметим, что если обе выборки извлечены из одной совокупности, то разность истинных средних равна нулю и в этом случае новая формула совпадает с предыдущей.

Вот математическая запись новой формулы:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{\bar{X}_1 - \bar{X}_2}}.$$

Поскольку истинных средних (то есть средних по совокупности) мы не знаем, то и вычислить значение t по этой формуле мы не можем. Но эта формула и не предназначена для нахождения t. Она позволяет сделать другое — оценить разность $\mu_1 - \mu_2$, то есть истинную величину различий. Для этого вместо вычисления t выберем его подходящее значение и, подставив в формулу, вычислим величину $\mu_1 - \mu_2$. Как выбрать «подходящее» значение?

По определению 100α процентов всех возможных значений t расположены левее $-t_\alpha$ или правее $+t_\alpha$. Остальные $100(1-\alpha)$ процентов значений t попадают в интервал от $-t_\alpha$ до $+t_\alpha$. Например, 95% значений t находится в интервале от $-t_{0,05}$ до $+t_{0,05}$. (Критические значения t, в частности $t_{0,05}$, можно найти по табл. 4.1.) Значит, в $100(1-\alpha)$ процентах всех случаев

$$-t_\alpha < \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{\bar{X}_1 - \bar{X}_2}} < +t_\alpha.$$

Преобразуя это неравенство, получаем

$$(\bar{X}_1 - \bar{X}_2) - t_\alpha s_{\bar{X}_1 - \bar{X}_2} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + t_\alpha s_{\bar{X}_1 - \bar{X}_2}.$$

Таким образом, разность истинных средних отличается от разности выборочных средних менее чем на произведение t_α и стандартной ошибки разности выборочных средних. Это неравенство задает *доверительный интервал* для разности средних $\mu_1 - \mu_2$. К примеру, 95% доверительный интервал для разности средних определяется неравенством

$$(\bar{X}_1 - \bar{X}_2) - t_{0,05} s_{\bar{X}_1 - \bar{X}_2} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + t_{0,05} s_{\bar{X}_1 - \bar{X}_2}.$$

В этот интервал разность истинных средних попадет в 95% случаев.

Этот способ определения доверительного интервала, как и критерий Стьюдента, на котором он основан, можно применять только тогда, когда совокупность имеет хотя бы приближенно нормальное распределение*.

Эффективный диуретик

На рис. 6.1 показан суточный диурез в совокупности из 200 человек после приема плацебо (рис. 6.1А) и диуретика (рис. 6.1Б). Средний диурез при приеме плацебо составил $\mu_п = 1200$ мл, при приеме диуретика — $\mu_д = 1400$ мл. Таким образом, препарат увеличивает суточный диурез на $\mu_д - \mu_п = 1400 - 1200 = 200$ мл. Как обычно, исследователь вынужден довольствоваться выборками, по которым он и оценивает величину эффекта. На рис. 6.1 помимо известных нам, но не исследователю, данных по совокупности приведены данные, полученные по двум выборкам, в каждую из которых входило по 10 человек. В контрольной группе средний диурез составил 1180 мл, а в группе, получавшей диуретик, — 1400 мл. Среднее увеличение диуреза в *данном* опыте:

$$\bar{X}_д - \bar{X}_п = 1400 - 1180 = 220 \text{ мл}.$$

Как и всякая выборочная оценка, подверженная влиянию

* Доверительные интервалы можно определять и в случае множественных сравнений. Подробнее об этом см.: J. H. Zar. Biostatistical analysis, 2nd ed, Prentice-Hall, Englewood Cliff, N. J., 1984, p. 191—192, 195.

случая, эта величина отличается от истинного увеличения суточного диуреза, равного 200 мл. И если бы мы, основываясь на выборочных данных, сказали, что препарат увеличивает суточный диурез в среднем на 220 мл, то упустили бы из виду неопределенность, присущую выборочной оценке. Правильнее будет рассчитать доверительный интервал — он покажет не одно число, скорее всего не совпадающее с истинным, а диапазон чисел, куда истинное попадает почти наверняка (например, с вероятностью 95%).

Вычислим сначала объединенную оценку дисперсии. По ней мы сможем найти стандартную ошибку разности средних. Стандартные отклонения у принимавших диуретик и плацебо составили соответственно 245 и 144 мл. В обеих группах было по 10 человек. Объединенная оценка дисперсии

$$s^2 = \frac{1}{2}(s_д^2 + s_п^2) = \frac{1}{2}(245^2 + 144^2) = 201^2.$$

Стандартная ошибка разности средних

$$s_{\bar{X}_д - \bar{X}_п} = \sqrt{\frac{s^2}{n_д} + \frac{s^2}{n_п}} = \sqrt{\frac{201^2}{10} + \frac{201^2}{10}} = 89{,}9.$$

Для определения 95% доверительного интервала найдем по табл. 4.1 значение $t_{0,05}$. Объем каждой из выборок $n = 10$. Поэтому число степеней свободы $\nu = 2(n-1) = 2(10-1) = 18$. Соответствующее табличное значение $t_{0,05}$ равно 2,101.

Теперь можно вычислить 95% доверительный интервал для среднего изменения диуреза:

$$(\bar{X}_д - \bar{X}_п) - t_{0,05} s_{\bar{X}_д - \bar{X}_п} < \mu_д - \mu_п < (\bar{X}_д - \bar{X}_п) + t_{0,05} s_{\bar{X}_д - \bar{X}_п},$$

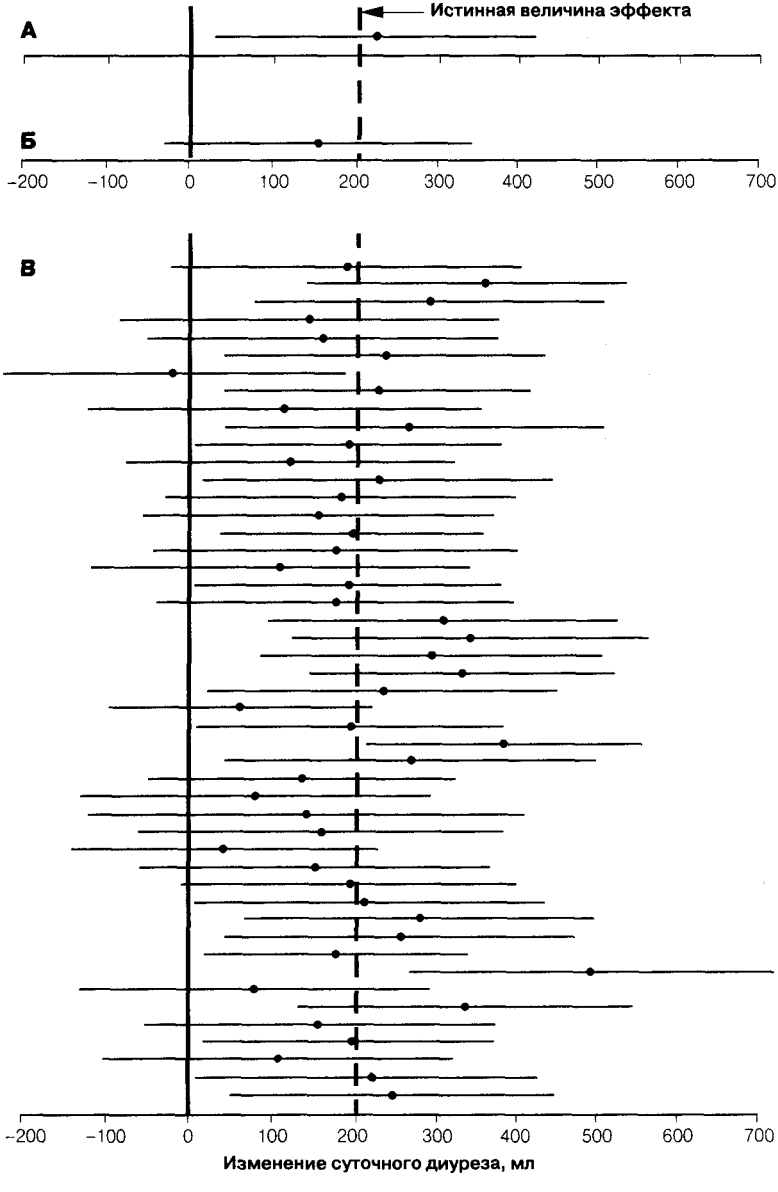
то есть

$$220 - 2{,}101 \times 89{,}9 < \mu_д - \mu_п < 220 + 2{,}101 \times 89{,}9$$

и окончательно:

$$31 < \mu_д - \mu_п < 409.$$

Таким образом, 95% доверительный интервал среднего изменения диуреза составляет 31—409 мл. Иными словами, выбо-


Истинная величина эффекта
А
Б
В
-200
-100
0
100
200
300
400
500
600
700
Изменение суточного диуреза, мл

рочные данные позволяют с 95% надежностью утверждать, что препарат увеличивает диурез более чем на 31 мл, но менее чем на 409 мл. Как и следовало ожидать, истинное значение 200 мл находится в этом интервале.

Первый из рассчитанных нами доверительных интервалов изображен на рис. 7.1А.

Другие выборки

Понятно, что в нашем распоряжении могли оказаться совершенно другие выборки. Ранее мы видели, что разные выборки дают разные оценки среднего и стандартного отклонения. Точно так же по разным выборкам мы будем получать разные доверительные интервалы. (И не удивительно — ведь доверительный интервал рассчитывают по среднему и стандартному отклонению.) Мы вычислили интервал по выборкам с рис 6.1. Для другой пары выборок — например с рис. 6.2 — доверительный интервал будет другим. Вычислим его.

Суточный диурез в группе плацебо составил в среднем 1216 мл, а в группе, получавшей диуретик, — 1368 мл. Стандартные отклонения — 97 и 263 мл соответственно. Увеличение среднего диуреза при приеме препарата $\bar{X}_д - \bar{X}_п = 1368 - 1216 = 152$ мл. Находим объединенную оценку дисперсии:

$$s^2 = \frac{1}{2}(97^2 + 263^2) = 198^2$$

Рис. 7.1. Новый взгляд на испытания диуретика. **А.** 95% доверительный интервал изменения диуреза, вычисленный по данным с рис. 6.1В. Интервал содержит истинную величину изменения (+200 мл) и не содержит нуля. Последнее говорит о том, что изменение диуреза статистически значимо. **Б.** Такой же доверительный интервал, вычисленный по данным с рис. 6.2В. Он тоже содержит истинную величину изменения диуреза, но он содержит также и ноль: статистически значимого изменения диуреза не выявлено. **В.** Еще сорок восемь 95% доверительных интервалов для пар выборок, извлеченных из той же пары совокупностей (рис. 6.1А и Б). Теперь у нас в общей сложности 50 доверительных интервалов. Из них 3 не содержат истинного значения и 27 не содержат нуля. Если бы мы построили 95% доверительные интервалы по всем возможным парам выборок, то доля не содержащих истинного значения составила бы 5%, а доля не содержащих нуля — 55%, что соответствует чувствительности критерия.

и стандартную ошибку разности средних:

$$s_{\bar{X}_д - \bar{X}_п} = \sqrt{\frac{198^2}{10} + \frac{198^2}{10}} = 89.$$

Тогда 95% доверительный интервал для среднего изменения суточного диуреза:

$$152 - 2{,}101 \times 89 < \mu_д - \mu_п < 152 + 2{,}101 \times 89,$$

$$-35 < \mu_д - \mu_п < 339.$$

Этот интервал (рис. 7.1Б) отличается от полученного ранее. Однако и он содержит истинное среднее увеличение диуреза — 200 мл. Если бы в нашем распоряжении была только выборка с рис. 6.2, мы бы сказали, что на 95% уверены в том, что препарат увеличивает средний диурез на величину, меньшую 339 и большую −35 мл. Заметьте, на сей раз доверительный интервал включает и отрицательные значения. Тем самым, выборочные данные не противоречат тому, что «диуретик» в действительности может уменьшать диурез. Значение этого интересного обстоятельства мы разберем позже, когда будем обсуждать использование доверительных интервалов для проверки гипотез.

Пока что мы определили доверительные интервалы для двух пар выборок из совокупности, изображенной на рис. 6.1. На самом деле число возможных пар выборок превышает 10^{27}. На рис. 7.1В показаны 95% доверительные интервалы для 48 из них. Теперь у нас в общей сложности 50 доверительных интервалов. Еще раз убедившись, что разные выборки дают разные доверительные интервалы, заметим, что большинство из них — точнее 47 из 50 — содержат истинное значение, показанное на рис. 7.1 вертикальной пунктирной линией. Если бы мы перебрали все возможные выборки, то доля 95% доверительных интервалов, содержащих истинное значение, составила бы в точности 95%.

ИНТЕРВАЛ ШИРЕ — ДОВЕРИЯ БОЛЬШЕ

Мы только что убедились, что 95% доверительный интервал может и не содержать истинного значения, однако, как правило, он

его содержит — а именно, в 95% случаев. Вообще, истинное значение содержат k процентов k-процентных доверительных интервалов. Иными словами, k — это *вероятность* того, что интервал содержит истинное значение. От этой вероятности k зависит ширина интервала. Взглянем еще раз на рис. 7.1. Если мы хотим, чтобы больше интервалов перекрывало истинное значение, нам придется их расширить. Чем больше k, тем шире k-процентный доверительный интервал. Для примера вычислим, в дополнение к 95%, еще и 90 и 99% доверительные интервалы для двух выборок с рис. 6.1. Разность средних и стандартная ошибка разности средних у нас уже есть, осталось только по табл. 4.1 найти новые значения t_α (по-прежнему число степеней свободы $\nu = 18$).

Для 90% доверительного интервала находим $t_{0,10} = 1{,}734$.

Тогда:

$$220 - 1{,}734 \times 89{,}9 < \mu_д - \mu_п < 220 + 1{,}734 \times 89{,}9,$$

$$64 < \mu_д - \mu_п < 376.$$

По сравнению с 95%, 90% доверительный интервал более узкий (рис. 7.2). Неужели волшебным образом наши знания о величине $\mu_д - \mu_п$ стали более точными? Разумеется, нет. Сужение доверительного интервала досталось нам ценой снижения вероятности того, что он действительно содержит истинное значение.

Для вычисления 99% доверительного интервала находим в табл. 4.1 критическое значение $t_{0,01} = 2{,}878$. Тогда интервал имеет вид

$$220 - 2{,}878 \times 89{,}9 < \mu_д - \mu_п < 220 + 2{,}878 \times 89{,}9,$$

то есть

$$-39 < \mu_д - \mu_п < 478.$$

Это самый широкий доверительный интервал из трех изображенных на рис. 7.2.

Подведем итоги. Приводя k-процентный доверительный интервал, мы сообщаем, во-первых, в каких пределах находится истинное значение неизвестной нам величины и, во-вторых — с какой вероятностью k. Например, говоря: «95% доверительный

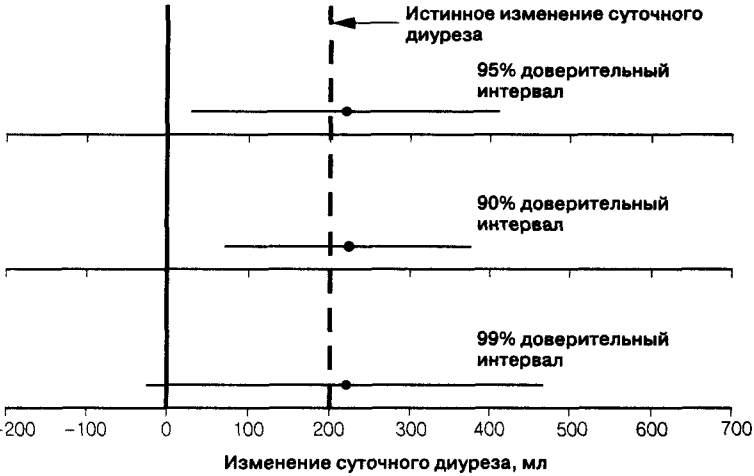


Рис. 7.2. Три доверительных интервала одной и той же разности средних (см. рис. 6.1). 99% доверительный интервал самый широкий, 90% — самый узкий. Истинная разность средних (изменение суточного диуреза) показана вертикальной пунктирной линией.

интервал 31—409 мл», имеют в виду следующее: «Вероятность того, что истинное значение лежит в пределах 31—409 мл, составляет 95%». Не исключено, к сожалению, что вам не повезет и истинное значение окажется вне доверительного интервала. С 95% доверительными интервалами такое случается в 5% случаев. Желая застраховаться от подобной ошибки, вы можете рассчитать 99% доверительный интервал. Однако учтите, что он окажется шире 95% доверительного интервала. Вообще, чем больше k (вероятность того, что доверительный интервал содержит истинное значение), тем больше ширина интервала.

ПРОВЕРКА ГИПОТЕЗ С ПОМОЩЬЮ ДОВЕРИТЕЛЬНЫХ ИНТЕРВАЛОВ

Доверительные интервалы можно использовать для оценки статистической значимости различий. Это и не удивительно, ведь нахождение доверительного интервала имеет общую базу с тра-

дициионными методами проверки гипотез. И там и тут мы встречаем разность выборочных средних, ее стандартную ошибку и распределение Стьюдента.

Истинная разность средних может находиться в любой точке доверительного интервала, поэтому если доверительный интервал содержит ноль, то мы не можем отвергнуть возможность того, что $\mu_1 - \mu_2 = 0$, то есть нулевую гипотезу. С другой стороны, нахождение истинной разности средних вне доверительного интервала маловероятно. Поэтому, если доверительный интервал не содержит нуля, справедливость нулевой гипотезы о равенстве средних маловероятна. Можно сформулировать следующее правило.

Если $100(1-\alpha)$*-процентный доверительный интервал разности средних не содержит нуля, то различия статистически значимы* ($P < \alpha$)*; напротив, если этот интервал содержит ноль, то различия статистически не значимы* ($P > \alpha$).

Применим это правило к двум только что рассмотренным примерам. На рис. 7.1А 95% доверительный интервал не содержит нуля, поэтому, как и при использовании критерия Стьюдента, мы заключаем, что препарат увеличивает диурез (уровень значимости $\alpha = 0{,}05$). Напротив, 95% доверительный интервал на рис. 7.1Б содержит ноль. Значит, в данном случае мы не можем отвергнуть гипотезу об отсутствии эффекта. К такому же выводу мы пришли раньше, используя критерий Стьюдента.

Из пятидесяти 95% доверительных интервалов на рис. 7.1 двадцать три содержат ноль. Следовательно, 23/50 = 44% соответствующих выборок не дают оснований говорить о статистически значимых различиях (то есть о наличии эффекта) при уровне значимости $1 - 0{,}95 = 0{,}05$. Если бы в нашем распоряжении были все возможные доверительные интервалы, мы увидели бы, что 45% из них содержат ноль. Это значит, что в 45% случаев мы не сможем отвергнуть гипотезу об отсутствии эффекта, то есть совершим ошибку II рода. Следовательно, как и прежде (см. рис. 6.4), $\beta = 0{,}45$, а чувствительность критерия равна $1 - 0{,}45 = 0{,}55$.

Говоря о «статистически значимых различиях», всегда полезно привести еще и доверительный интервал — это даст возможность судить о величине эффекта. Если статистическая значимость обнаружена благодаря большому объему выборки, а не величине эффекта, доверительный интервал укажет на это. Другими сло-

вами, использование доверительных интервалов позволяет среди статистически значимых эффектов выделить те, которые сами по себе слишком слабы, чтобы иметь клиническое значение.

Предположим, мы должны оценить эффективность гипотензивного препарата. Мы набираем две группы по 100 человек в каждой — контрольную, которой даем плацебо, и экспериментальную, которой даем препарат. Пусть в экспериментальной группе диастолическое давление составило в среднем $\bar{X}_э = 81$ мм рт. ст. (стандартное отклонение 11 мм рт. ст.), а в контрольной — $\bar{X}_к = 85$ мм рт. ст. (стандартное отклонение 9 мм рт. ст.). Для оценки статистической значимости различий воспользуемся критерием Стьюдента.

Объединенная оценка дисперсии составляет

$$s^2 = \frac{1}{2}(11^2 + 9^2) = 10^2,$$

откуда

$$t = \frac{\bar{X}_э - \bar{X}_к}{s_{\bar{X}_э - \bar{X}_к}} = \frac{81-85}{\sqrt{\frac{10^2}{100} + \frac{10^2}{100}}} = -2{,}83.$$

Это значение по абсолютной величине больше критического значения $t_{0,01} = 2{,}601$ для уровня значимости 0,01 и числа степеней свободы $\nu = 2(n-1) = 198$ (см. табл. 4.1). Таким образом, снижение диастолического артериального давления статистически значимо ($P < 0{,}01$).

Мы обнаружили статистически значимый эффект. Но какова его *клиническая* значимость? Вычислим 95% доверительный интервал для разности средних. Так как при 198 степенях свободы $t_{0,05}$ равно 1,972 (см. табл. 4.1), доверительный интервал имеет вид

$$-4 - 1{,}972 \times 1{,}41 < \mu_э - \mu_к < -4 + 1{,}972 \times 1{,}41,$$

то есть

$$-6{,}8 < \mu_э - \mu_к < -1{,}2$$

Таким образом, с вероятностью 95% препарат снижает артериальное давление на 1,2—6,8 мм рт. ст. Этот эффект невелик, особенно если сравнить его со стандартными отклонениями (9 и

11 мм рт. ст.). Итак, гипотензивный эффект выражен слабо, а его статистическая значимость обусловлена исключительно большой численностью групп.

Приведенный пример наглядно показывает, почему, знакомясь с исследованием эффективности того или иного препарата, важно знать не только уровень значимости, но и величину эффекта. Авторы публикаций редко балуют читателя доверительными интервалами, но обычно все же указывают численность групп, средние величины и их стандартные ошибки. В таких случаях нужно самостоятельно рассчитать стандартные отклонения (произведение стандартной ошибки среднего на квадратный корень из численности группы) и построить доверительный интервал. Этого часто достаточно, чтобы понять, имеет исследование сугубо академическую или еще и практическую ценность.

ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ ДЛЯ СРЕДНЕГО

Продолжим рассматривать разнообразные применения доверительных интервалов. Найдем доверительный интервал для среднего. Определив выборочное среднее $\overline{X}$, мы понимаем, разумеется, что это всего лишь выборочная оценка истинного среднего μ, которое, впрочем, скорее всего находится где-то поблизости. «Где-то поблизости» можно охарактеризовать количественно, то есть указать интервал, в котором с заданной вероятностью k находится истинное среднее. Это и будет k-процентный доверительный интервал для среднего.

Приближенный способ вычисления этого интервала изложен в гл. 2: примерно в 95% случаев выборочное среднее уклоняется от истинного не более чем на две стандартные ошибки среднего. Осталось внести некоторые уточнения.

Ранее мы выяснили, что величина

$$t = \frac{\text{Разность выборочных средних} - \text{Разность истинных средних}}{\text{Стандартная ошибка разности выборочных средних}}$$

подчиняется распределению Стьюдента. Можно показать, что

$$t = \frac{\text{Выборочное среднее} - \text{Истинное среднее}}{\text{Стандартная ошибка среднего}}$$

также подчиняется распределению Стьюдента. Математическая запись для последней величины выглядит так:

$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}}.$$

Дальнейший вывод аналогичен выводу доверительного интервала для разности истинных средних. Опустив промежуточные этапы, приведем формулу $100(1-\alpha)$-процентного доверительного интервала для среднего:

$$\bar{X} - t_\alpha s_{\bar{X}} < \mu < \bar{X} + t_\alpha s_{\bar{X}},$$

где t_α — критическое значение t для уровня значимости α и числа степеней свободы $\nu = n - 1$ (n — объем выборки).

Смысл доверительного интервала для среднего совершенно аналогичен смыслу доверительного интервала для разности средних. Приводя k-процентный доверительный интервал среднего, мы утверждаем, что вероятность того, что истинное среднее находится в этом интервале, равна k. Иными словами, если получить все возможные выборки из некоторой совокупности и для каждой рассчитать k-процентный доверительный интервал, то доля интервалов, содержащих среднее по совокупности (истинное среднее), составит k.

Вычислить доверительный интервал несложно, однако — если объем выборки достаточно велик — можно пользоваться и приведенным выше «правилом двух стандартных ошибок». Для выборок, имеющих объем от 20 и выше, $t_{0,05}$ приблизительно равно 2 (см. табл. 4.1), и мы получим достаточно точный результат. Если же объем выборки меньше 20, доверительный интервал окажется зауженным, а наше представление о точности, с какой мы можем судить об истинном среднем, — преувеличенным.

ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ ДЛЯ РАЗНОСТИ ДОЛЕЙ

Изложенные способы вычисления доверительных интервалов

нетрудно приспособить для разности долей. В гл. 5 мы определили критерий z как

$$z = \frac{\text{Разность выборочных долей}}{\text{Стандартная ошибка разности выборочных долей}}.$$

Величина z имеет приблизительно нормальное распределение; в гл. 5 мы использовали z для проверки гипотезы о равенстве двух выборочных долей (или, что то же самое, для оценки статистической значимости различий выборочных долей). Можно показать, что даже если в совокупностях, из которых извлечены выборки, доли различны, то отношение

$$z = \frac{\text{Разность выборочных долей} - \text{Разность истинных долей}}{\text{Стандартная ошибка разности выборочных долей}}$$

приближенно следует нормальному распределению — при условии, что объемы выборок достаточно велики.

Если p_1 и p_2 — истинные доли в каждой из совокупностей, а $\hat{p}_1$ и $\hat{p}_2$ — выборочные оценки этих долей, то

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{s_{\hat{p}_1 - \hat{p}_2}}.$$

В $100(1-\alpha)$ процентах случаев z по абсолютной величине не превышает z_α, то есть

$$-z_\alpha < \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{s_{\hat{p}_1 - \hat{p}_2}} < z_\alpha.$$

Преобразовав это неравенство, мы получим формулу для $100(1-\alpha)$-процентного интервала для разности истинных долей:

$$(\hat{p}_1 - \hat{p}_2) - z_\alpha s_{\hat{p}_1 - \hat{p}_2} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + z_\alpha s_{\hat{p}_1 - \hat{p}_2}.$$

Как вы помните, распределение Стьюдента с увеличением числа степеней свободы стремится к нормальному. Поэтому z_α можно найти в табл. 4.1 — в строке, соответствующей бесконечному числу степеней свободы.

Чаще всего используют 95% доверительный интервал, в этом случае $z_\alpha = z_{0,05} = 1{,}96$.

Галотан и морфин: операционная летальность

В гл. 5 мы сравнивали операционную летальность при галотановой и морфиновой анестезии и не нашли статистически значимых различий. Посмотрим, каков 95% доверительный интервал для различия летальностей.

В группе галотана умерли 8 оперированных из 61, доля умерших $\hat{p}_1 = 8/61 = 0{,}13$. В группе морфина умерли 10 из 67, $\hat{p}_2 = 0{,}15$. Разность долей равна $\hat{p}_1 - \hat{p}_2 = 0{,}13 - 0{,}15 = -0{,}02$. Объединенная оценка доли

$$\hat{p} = \frac{8+10}{61+67} = 0{,}14$$

и стандартная ошибка разности

$$s_{\hat{p}_1-\hat{p}_2} = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} =$$

$$= \sqrt{0{,}14(1-0{,}14)\left(\frac{1}{61}+\frac{1}{67}\right)} = 0{,}062 = 6{,}2\%.$$

Тем самым, 95% доверительный интервал для различия летальности имеет вид:

$$(\hat{p}_1 - \hat{p}_2) - z_{0{,}05} s_{\hat{p}_1-\hat{p}_2} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + z_{0{,}05} s_{\hat{p}_1-\hat{p}_2},$$

то есть

$$-0{,}020 - 1{,}960 \times 0{,}062 < p_1 - p_2 < -0{,}020 + 1{,}960 \times 0{,}062$$

и окончательно

$$-0{,}142 < p_1 - p_2 < 0{,}102.$$

Итак, с вероятностью 95% можно утверждать, что истинная величина различия попадает в интервал между −14,2 и 10,2%. Вычисленный доверительный интервал содержит ноль, поэтому различия летальности статистически не значимы*.

---

* При использовании поправки Йейтса нужно раздвинуть границы доверительного интервала, соответственно уменьшив нижнюю и увеличив верхнюю на величину $(1/n_1 + 1/n_2)/2$.

Тромбоз шунта у больных на гемодиализе

В гл. 5 мы рассмотрели влияние аспирина на риск тромбоза шунта у больных на гемодиализе. Доля больных с тромбозом в группе плацебо составила 72%, а в группе, получавшей аспирин, — 32%. Мы уже убедились, что это различие статистически значимо. Однако мы не можем утверждать, что «аспирин снижает риск тромбоза на 40%», — правильнее будет указать доверительный интервал для снижения риска. Стандартную ошибку разности долей мы уже рассчитали в гл. 5, она составляет 0,15. Поэтому 95% доверительный интервал для истинной разности долей имеет вид

$$0{,}40 - 1{,}96 \times 0{,}15 < p_{\text{п}} - p_{\text{а}} < 0{,}40 + 1{,}96 \times 0{,}15,$$

то есть

$$0{,}11 < p_{\text{п}} - p_{\text{а}} < 0{,}69.$$

Таким образом, в вероятностью 95% можно утверждать, что прием аспирина снижает риск тромбоза на величину от 11 до 69%.

Отрицателен ли «отрицательный» результат?

В гл. 6 мы познакомились со статьей Фреймана и соавт. Они рассмотрели 71 медицинскую публикацию, в которых исследуемый метод лечения не дал статистически значимого снижения частоты неблагоприятных исходов (под неблагоприятным исходом в разных статьях понимали смерть, осложнения и т. п.). Фрейман и соавт. обнаружили, что в большинстве работ численность групп была слишком мала, чтобы обеспечить достаточную чувствительность. Неужели столь огромный труд пропал даром? Попробуем получить из этих работ хоть какую-то информацию.

На рис. 7.3 представлены 90% доверительные интервалы величины эффекта (разность долей неблагоприятных исходов в контрольной и экспериментальной группах). Статистически значимых различий не было выявлено ни в одном случае, поэтому все они содержат ноль. Посмотрим на верхнюю границу доверительных интервалов. Можно заметить, что во многих случаях она отличается от нуля всего на несколько процентов. Иными словами, с вероятностью 90% мы можем утверждать, что эффект, если и существует, весьма незначителен. Дальнейшие исследования

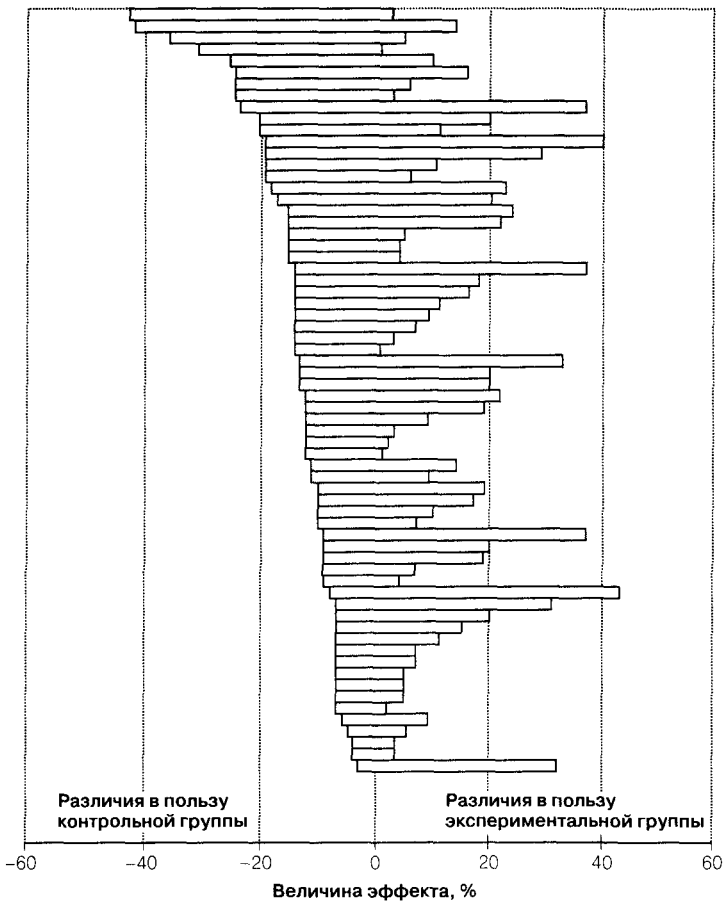


Рис. 7.3. 90% доверительные интервалы величины эффекта в 71 клиническом испытании. Здесь величина эффекта — это разность долей больных с неблагоприятным исходом в контрольной и экспериментальной группах. Поскольку статистически значимого эффекта не было выявлено ни в одном случае, все доверительные интервалы содержат ноль. Видно, что некоторые доверительные интервалы довольно сильно смещены в сторону положительных значений — возможно, при большем числе больных различия достигли бы статистической значимости. В других случаях верхняя граница интервала превышает ноль всего на несколько процентов. Можно сделать вывод, что если соответствующие методы лечения и дают эффект, то очень незначительный.

соответствующих методов лечения вряд ли перспективны. Верхняя граница некоторых интервалов простирается до 30% и даже до 40%. Напомним, что с вероятностью 90% мы можем утверждать, что истинная величина находится внутри доверительного интервала, но где именно — определить невозможно. Поэтому не исключено, что соответствующие методы лечения все же эффективны и при большей численности групп это удалось бы доказать. Если мы решим повторить испытание, то при его планировании стоит учесть полученные оценки. Было бы неразумно, например, рассчитывать чувствительность и численность групп, полагая, что величина эффекта достигнет 50%.

ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ ДЛЯ ДОЛИ

Если объем выборки достаточно велик, то доверительный интервал для доли можно приближенно вычислить, используя нормальное распределение*.

Когда выборка мала (а в медицинских исследованиях так оно обычно и бывает), приближение нормальным распределением недопустимо. В таких случаях приходится вычислять точные значения доверительных интервалов, используя биномиальное распределение. Чтобы не обременять читателя вычислительными тонкостями, мы чуть позже приведем графический способ нахождения доверительных интервалов по малым выборкам. Заметим, что при оценке долей по выборкам небольшого объема расчет доверительного интервала особенно желателен. Причина в том, что, если выборка мала, изменение признака *даже у одного* из ее членов приведет к резкому изменению долей.

Итак, при достаточно большом объеме выборки величина

$$z = \frac{\text{Наблюдаемая доля} - \text{Истинная доля}}{\text{Стандартная ошибка долей}}$$

приближенно следует нормальному распределению (см. табл. 6.4).

---

* Как говорилось в гл. 5, для этого нужно, чтобы и np и $n(1-p)$ были больше 5 (здесь n — объем выборки, p —доля).

Математическая запись для z:

$$z = \frac{\hat{p} - p}{s_{\hat{p}}}.$$

Отсюда уже знакомым способом получаем формулу для $100(1-\alpha)$-процентного доверительного интервала для истинной доли:

$$\hat{p} - z_\alpha s_{\hat{p}} < p < \hat{p} + z_\alpha s_{\hat{p}}.$$

Доля статей, содержащих статистические ошибки

Как видно из рис. 1.3, доля статей с ошибками в применении статистических методов за последние несколько десятков лет составляет 40—60%. Глядя на график, можно подумать, что доля эта с годами снижается. Однако рассмотрены были далеко не все статьи, поэтому точки — это всего лишь оценки истинной доли. Построим 95% доверительный интервал для последней точки — может быть, наше впечатление изменится.

Последняя точка соответствует периоду с января по март 1976 г. Из оригинальных статей, опубликованных в этот период, С. Гор и соавт.* рассмотрели 77, статистические ошибки были обнаружены в 32. Выборочная доля составляет $\hat{p} = 32/77 = 0{,}42$, ее стандартная ошибка

$$s_{\hat{p}} = \sqrt{\frac{0{,}42(1-0{,}42)}{77}} = 0{,}056.$$

Тогда 95% доверительный интервал имеет вид

$$0{,}42 - 1{,}96 \times 0{,}056 < p < 0{,}42 + 1{,}96 \times 0{,}056,$$

то есть

$$0{,}31 < p < 0{,}53.$$

В этот интервал попадают обе оценки, сделанные в 60-х го-

---

* S. M. Gore, I. G. Jones, E. C. Rytter. Misuse of statistical methods: critical assessment of articles in BMJ from January to March 1976. *Br. Med. J.*, 1(6053):85—87, 1977.

дах. Вряд ли это позволяет утверждать, что ситуация меняется к лучшему.

Ошибки плодят ошибки. Авторы обзоров, опираясь на неверные данные оригинальных статей, делают неверные выводы, которые воспринимаются читателями как последнее слово медицинской науки. Насколько широко распространено это явление? На несостоятельные данные оригинальных статей опирались авторы 5 из 62 обзорных статей, рассмотренных Гор. Таким образом,

$$\hat{p} = \frac{5}{62} = 0{,}081,$$

$$s_{\hat{p}} = \sqrt{\frac{0{,}081(1-0{,}081)}{62}} = 0{,}035.$$

Тогда 95% доверительный интервал для доли обзорных статей, содержащих необоснованные выводы, имеет вид:

$$0{,}081 - 1{,}960 \times 0{,}035 < p < 0{,}081 + 1{,}960 \times 0{,}035.$$

То есть это интервал от 1,2 до 15%.

Точные доверительные интервалы для долей

Часто объем выборки или наблюденная доля слишком малы, чтобы использовать приближение с помощью нормального распределения*. В подобных случаях следует воспользоваться точным распределением. Это так называемое *биномиальное распределение*. Оно чрезвычайно важно для медицинских исследова-

* Причина, позволившая нам (в этой главе и гл. 5) использовать нормальное распределение вместо биномиального, состоит в том, что с ростом объема выборки биномиальное распределение стремится к нормальному. Это следует из сформулированной в гл. 2 центральной предельной теоремы. Более подробное изложение можно найти в: W. J. Dixon, F. J. Massey. Introduction to statistical analysis, McGraw-Hill, New York, 1983, sec. 13-5, Binomial distribution: proportion, и B. W. Broun, Jr., M. Hollander. Statistics: a biomedical introduction, Wiley, New York, 1977, Chap. 7, Statistical Inference for Dichotomous Variable.

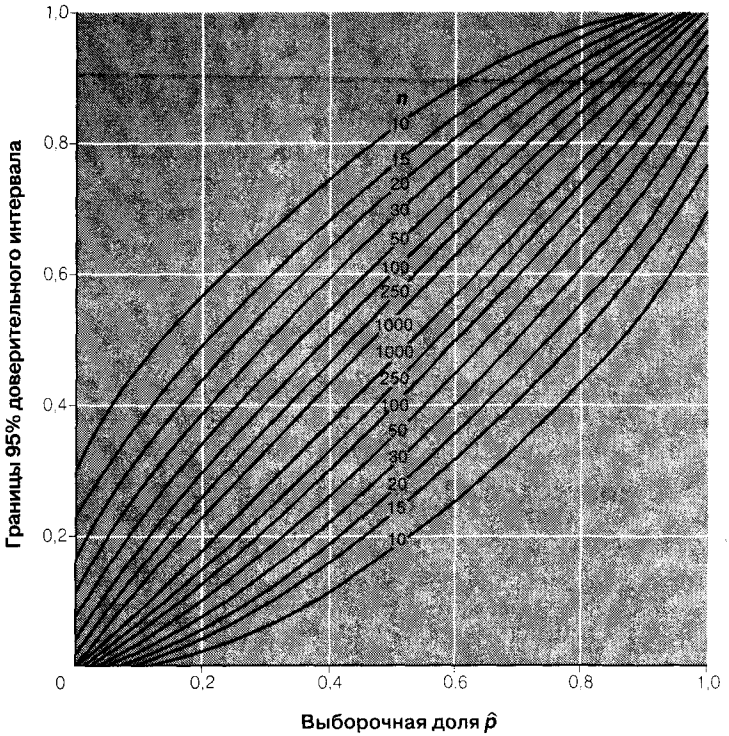


Рис. 7.4. 95% доверительные интервалы для долей, вычисленные на основании биномиального распределения. Найдите на горизонтальной оси точку, соответствующую выборочной доле. Проведите через эту точку вертикальную линию. Границы доверительного интервала — это вертикальные координаты точек пересечения этой линии с парой кривых, соответствующих объему выборки n.

ний, в которых часто приходится иметь дело с редкими событиями и выборками малого объема.

Сначала покажем, к чему приводит неправомерное использование метода, основанного на нормальном распределении. Рассмотрим пример, в котором $np < 5$, то есть нарушено одно из условий применимости нормального распределения. Испытывая новый препарат, мы дали его 30 добровольцам, и, к счастью, ни у

одного из них препарат не оказал побочного действия. Выборочная оценка риска побочного действия

$$\hat{p} = \frac{0}{30} = 0\%.$$

Вряд ли можно на этом основании гарантировать, что препарат *никогда* не окажет побочного действия. Чтобы получить более реалистичную оценку, вычислим 95% доверительный интервал для p.

Какие результаты даст расчет, основанный на использовании нормального распределения? Имеем $\hat{p} = 0$, поэтому

$$s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0(1-0)}{30}} = 0.$$

Тем самым, 95% доверительный интервал состоит из единственной точки — нуля. Возможно, это неплохо для рекламы нового препарата, но, увы, противоречит здравому смыслу.

Обратимся теперь к рис. 7.4. Чтобы определить доверительный интервал, основанный на биномиальном распределении, нужно сначала найти на горизонтальной оси точку, соответствующую выборочной доле $\hat{p}$. Затем нужно провести из нее перпендикуляр и посмотреть, где его пересекает пара кривых, помеченных числом, равным объему выборки. Вертикальные координаты точек пересечения — это и есть границы 95% доверительного интервала. В нашем примере $\hat{p} = 0$ и $n = 30$. Нижняя граница доверительного интервала — 0, верхняя — около 0,1. Тем самым с вероятностью 95% мы можем утверждать, что риск побочного действия не превысит 10%.

Предположим, что в одном случае из 30 препарат все-таки оказал побочное действие. Тогда $\hat{p} = 1/30 = 0{,}033$ и

$$s_{\hat{p}} = \sqrt{\frac{0{,}033(1-0{,}033)}{30}} = 0{,}033.$$

Используя нормальное приближение, мы получили бы

$$0{,}033 - 1{,}96 \times 0{,}033 < p < 0{,}033 + 1{,}96 \times 0{,}033,$$

то есть

$-0{,}032 < p < 0{,}098.$

Понятно, что ни в каком случае доля не может быть *отрицательной* величиной, хотя величина интервала, как окажется, определена правильно.

Какой интервал даст биномиальное распределение? По рис. 7.4 находим, что это интервал от 0 до примерно 0,13. Обратите внимание, что он не сильно отличается от интервала, найденного для $\hat{p} = 0$. Так и должно быть, ведь различие между отсутствием осложнений и одним осложнением весьма незначительно.

Заметьте, что чем меньше объем выборки, тем сильнее он влияет на величину доверительного интервала. Предположим, мы бы дали препарат не 30, а 10 добровольцам. Тогда нижний предел 95% доверительного интервала, конечно, остался бы нулем, но верхний был бы уже не 13, а 33%.

ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ ДЛЯ ЗНАЧЕНИЙ*

До сих пор нас интересовали доверительные интервалы для тех или иных *параметров* распределения, например среднего μ или доли p. Нередко, однако, нужен доверительный интервал для самих *значений измеряемого признака*. Например, мы хотим оценить диапазон, в который будет попадать 95% всех значений. Особенно часто подобные задачи возникают при определении границ нормы какого-нибудь лабораторного показателя. Обычно доверительный интервал значений определяют как *выборочное среднее плюс-минус два стандартных отклонения*. Если мы имеем дело с нормальным распределением и объем выборки достаточно велик (больше 100 человек), то правило двух стандартных отклонений дает верный результат. Как быть, если в нашем распоряжении не 100, а менее двух десятков человек, что довольно типично для клинических исследований? Разумеется, об определении границ нормы по столь малой выборке нечего и думать. Тем не менее оценку доверительного интервала можно получить и тут. Однако от правила двух стандартных отклонений

---

* Описанные ниже методы применимы только к данным, приближенно подчиняющимся нормальному распределению.

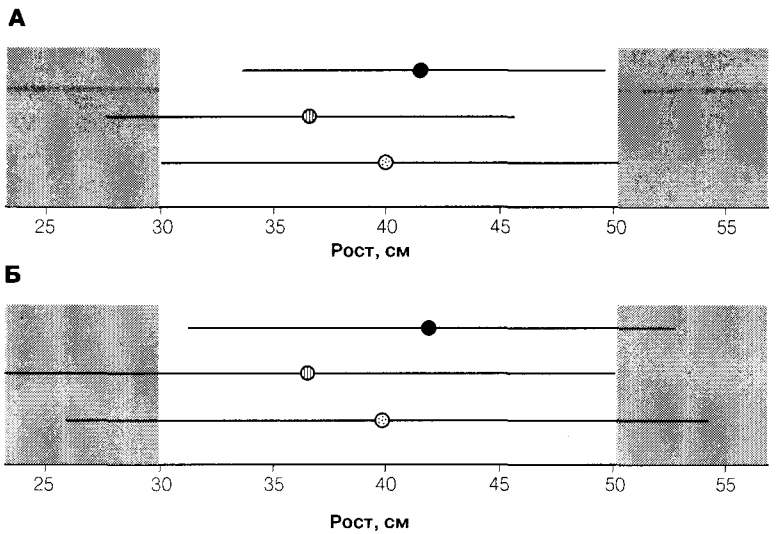


Рис. 7.5. 95% доверительные интервалы для роста марсиан, вычисленные по трем выборкам с рис. 2.6. **А.** В качестве доверительного интервала использовали среднюю величину плюс-минус два стандартных отклонения. Результат оставляет желать лучшего: два интервала из трех не покрывают истинного интервала, заключающего 95% значений. **Б.** Доверительные интервалы определили как среднее плюс-минус произведение $K_{0,05}$ на стандартное отклонение. Ситуация улучшилась — теперь истинный интервал покрывают два интервала.

придется отказаться: при малых выборках интервал получается слишком узким.

Рассмотрим пример. На рис. 2.6 представлены распределение по росту всех 200 ныне живущих марсиан, а также три случайные выборки по 10 марсиан в каждой. Рост 95% всех марсиан лежит в пределах от 31 до 49 см. Средний рост марсианина — 40 см, стандартное отклонение — 5 см. Три выборки, изображенные в нижней части рисунка, дают следующие оценки среднего роста: 41,5, 36 и 40 см. Выборочные стандартные отклонения — соответственно 3,8, 5 и 5 см. Применим к этим выборочным оценкам правило двух стандартных отклонений. Полученные доверительные интервалы изображены на рис. 7.5А. Как видим, в двух из трех случаев интервалы не покрывают 95% всех членов совокупности.

Причина, в общем, понятна. Выборочное среднее и выбо-

рочное стандартное отклонение — не более чем *оценки* истинного среднего и стандартного отклонения. Точность этих оценок при малом объеме выборок невелика. Ошибка в оценке одного параметра накладывается на ошибку в оценке другого — в результате шансы получить правильный результат и вовсе низки. Рассмотрим выборку на рис. 2.6В. Нам повезло — оценка стандартного отклонения совпала с истинным его значением 5 см. Однако оценка среднего оказалась заниженной — 36 см вместо 40 см. Поэтому интервал смещен относительно истинного среднего и накрывает менее 95% всех значений.

Учитывая приблизительность оценок по выборкам небольшого объема, нужно брать интервал, более широкий, чем плюс-минус два стандартных отклонения (при выборках большого объема такая страховка не нужна). Этот интервал вычисляют по формуле

$$\bar{X} - K_\alpha s < X < \bar{X} + K_\alpha s,$$

где $\bar{X}$ — выборочное среднее, s — выборочное стандартное отклонение, а K_α — коэффициент, который зависит от доли f членов совокупности, которые должны попасть в доверительный интервал, от вероятности того, что они действительно туда попали $1-\alpha$ и от объема выборки n. Этот коэффициент играет примерно ту же роль, что t_α или z_α. Для вычисления 95% доверительного интервала нужно определить $K_{0,05}$; зависимость $K_{0,05}$ от объема выборки для различных значений f показана на рис. 7.6.

Заметим, что K_α больше, чем t_α (как t_α больше, чем z_α), поскольку учитывает не только значение среднего, но и неопределенность оценок среднего и стандартного отклонения*.

При объеме выборки от 5 до 25, типичном для медицинских исследований, K_α должен быть существенно больше двух. Если бы в рассматриваемом случае мы взяли интервал в плюс-минус два стандартных отклонения от среднего, то он покрыл бы заметно менее 95% совокупности. На рис. 7.5Б изображены 95% доверительные интервалы для роста 95% членов совокупности

---

* Вывод формулы для K_α, показывающий его связь с доверительными интервалами для среднего и стандартного отклонения, можно найти, например, в работе: A. E. Lewis, Biostatistics, Reinhold, New York, 1966, Chap. 12. Tolerance limits and indices of discrimination.

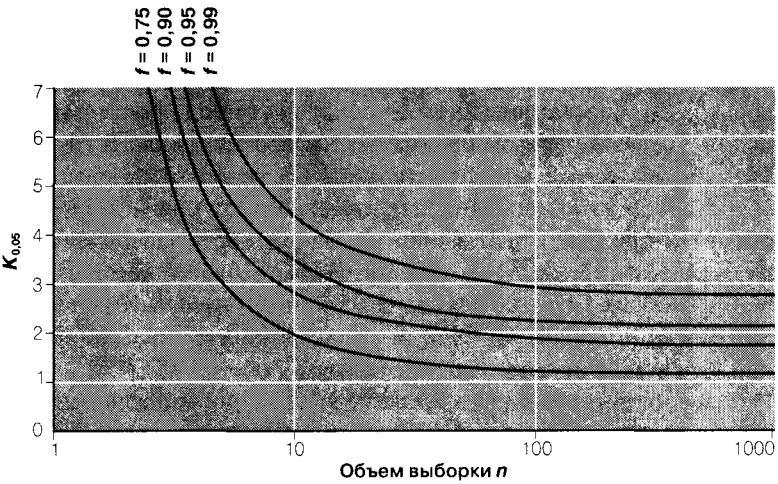


Рис. 7.6. Коэффициент $K_{0,05}$ зависит от объема выборки и от доли членов совокупности f, которые должны попадать в 95% доверительный интервал.

марсиан, построенные по трем выборкам с рис. 2.6. Теперь все три интервала покрывают не менее 95% членов совокупности.

Применение правила двух стандартных отклонений к выборкам небольшого объема приводит к зауживанию доверительного интервала значений. Упомянем еще об одной распространенной ошибке. Как говорилось в гл. 2, многие путают стандартную ошибку среднего со стандартным отклонением. Найдя интервал «выборочное среднее плюс-минус две стандартные ошибки среднего», они уверены, что в него попадет 95% совокупности (тогда как на самом деле 95% составляет вероятность, что в интервал попадет среднее по совокупности). В результате интервал допустимых значений оказывается еще более зауженным.

ЗАДАЧИ

7.1. По данным из задачи 2.6 найдите 90 и 95% доверительные интервалы для среднего числа авторов статей, опубликованных в медицинских журналах за 1946, 1956, 1966 и 1976 гг.

7.2. Ранее (задача 3.1) мы познакомились с исследованием

Ч. О'Херлихи и Г. Мак-Дональда (C. O'Herlihy, H. MacDonald. Influence of preinduction prostaglandin E_2 vaginal gel on cervical ripening and labor. *Obstet. Gynecol.*, 54:708—710, 1979). Как выяснилось, гель с простагландином E_2 сокращает продолжительность родов. Позволяет ли он избежать кесарева сечения? В группе, получавшей гель с простагландином E_2, кесарево сечение потребовалось 15% женщин, в контрольной группе — 23,9%. В обеих группах было по 21 женщине. Найдите 95% доверительные интервалы для доли рожениц, которым требуется кесарево сечение в обеих группах. Найдите 95% доверительный интервал для разности долей. Можно ли утверждать, что простагландин снижает вероятность кесарева сечения?

7.3. По данным задачи 3.1 найдите 95% доверительный интервал для разности средней продолжительности родов у получавших гель с простагландином E_2 и получавших плацебо. Позволяет ли вычисленный доверительный интервал утверждать, что различия статистически значимы?

7.4. По данным задачи 5.1 найдите 95% доверительные интервалы для долей больных, которые не чувствовали боли при включенном и выключенном приборе. Можно ли по этим интервалам оценить статистическую значимость различий?

7.5. По данным задачи 3.2 найдите 95% доверительные интервалы для каждой из групп. В чем заключаются различия между группами?

7.6. По данным задачи 5.6 найдите 95% доверительные интервалы для доли работ, где данные были получены до планирования исследования.

7.7. По данным задачи 2.2 найдите 95% доверительные интервалы для 90 и 95% значений. Результаты представьте на одном рисунке с исходными данными.

Глава 8

Анализ зависимостей

Самый первый из рассмотренных нами примеров (рис. 1.2) был посвящен вопросу об эффективности диуретика. Пяти людям дали разные дозы препарата, измерили диурез и увидели, что чем больше доза, тем больше диурез. В дальнейшем оказалось, что этот результат не отражает реальной картины и что никакой связи между дозой и диурезом на самом деле нет. Тогда мы еще не знали о методах анализа зависимостей. Им посвящена эта глава. Мы узнаем, как с помощью *уравнения регрессии* выразить связь между дозой диуретика и диурезом (так называемый *регрессионный анализ*) и как с помощью *коэффициента корреляции* измерить силу этой связи.

Подобно тому как мы поступали в предыдущих главах, рассмотрим сначала уравнение регрессии для *совокупности*, а затем выясним, как оценивать его параметры по *выборке*. В гл. 3 и 4 мы брали нормально распределенную совокупность, находили параметры распределения (среднее μ и стандартное отклонение σ), затем находили выборочные оценки этих параметров ($\bar{X}$ и s) и

использовали их для оценки значимости *различий между группами*, например получавших препарат и не получавших. Теперь мы также будем иметь дело с нормально распределенной совокупностью, но группа будет только одна. Интересовать же нас будет *связь между двумя количественными признаками,* характеризующими членов этой группы, например между дозой препарата и эффектом, ростом и весом. Мы ограничимся случаем *линейной* зависимости *двух* переменных*.

Сколько весит марсианин?

Итак, начнем с совокупности. Совокупность марсиан нами уже достаточно хорошо изучена, особенно что касается роста. Но ведь мы их еще и взвешивали! Разберемся, как связаны вес и рост. Вы, конечно, помните, что на Марсе живет 200 марсиан. В гл. 2 мы обнаружили, что их рост подчиняется нормальному распределению со средним $\mu = 40$ см и стандартным отклонением $\sigma = 5$ см. Оказывается, что вес марсиан тоже подчиняется нормальному распределению с параметрами $\mu = 12$ г и $\sigma = 2{,}5$ г. Но самое замечательное, что отчетливо видно на рис. 8.1, — это зависимость веса от роста. Как правило, *чем больше рост марсианина, тем больше вес, причем эта зависимость линейна.*

Посмотрим, сколько весят марсиане, чей рост равен 32 см. Таких марсиан четверо, а их вес равен соответственно 7,1; 7,8; 8,3 и 8,8 г. Таким образом, средний вес марсиан ростом 32 см равен 8 г. Восемь марсиан ростом 46 см весят 13,7; 14,5; 14,8; 15,0; 15,1; 15,2; 15,3 и 15,8 г. Их средний вес 15 г. Если для каждого значения роста мы подсчитаем соответствующий ему средний вес, то окажется, что найденные значения лежат на *прямой линии,* как изображено на рис. 8.2.

Теперь, выбрав какой-то рост, мы всегда сможем примерно определить вес марсианина этого роста. Точнее, мы сможем оп-

---

* Линейная зависимость y от x определяется формулой $y = \alpha + \beta x$. Возможна нелинейная зависимость, например $y = \alpha + \beta x^2$. Возможна и множественная зависимость, когда определяющих признаков более одного, например $y = \alpha + \beta x + \gamma z$. Она рассматривается в книге S. Glantz, B. Slinker. Primer of applied regression and analysis of variance. McGraw-Hill, New York, 1990.

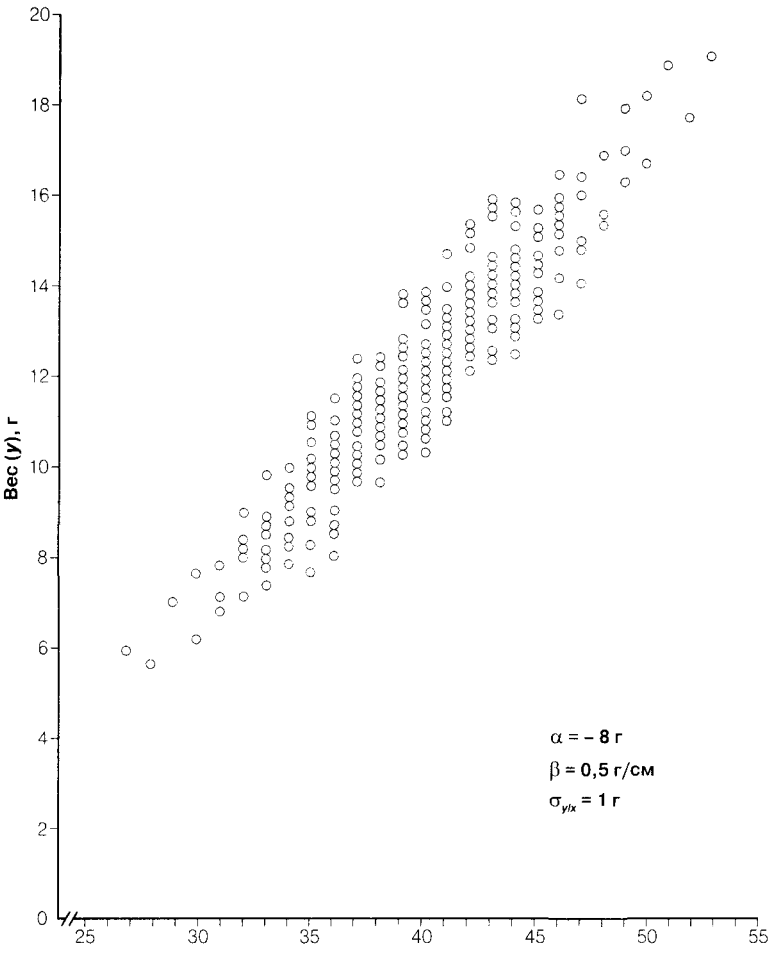


Рис. 8.1. Рост и вес марсиан. Как известно, число обитателей Марса составляет 200; каждый из них был измерен и взвешен, результат нанесен на график в виде кружка. Распределение марсиан по росту и по весу нормально. Более того, средний вес марсиан определенного роста связан с ростом линейной зависимостью; разброс значений веса для всех ростов одинаков. Чтобы к совокупности можно было применить регрессионный анализ, она должна обладать всеми этими свойствами.

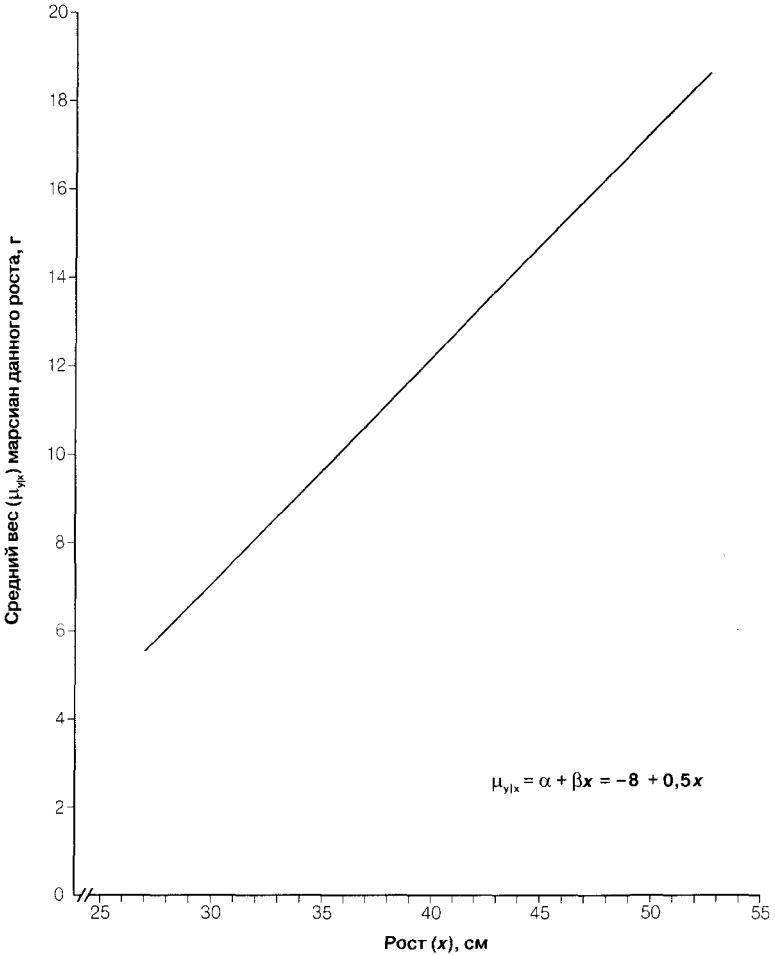


Рис. 8.2. Если рассчитать средний вес марсиан разного роста и нанести полученные значения на график, окажется, что они образуют прямую линию. Иначе говоря, средний вес марсиан линейно зависит от роста.

ределить *средний* вес марсиан этого роста, поскольку для каждого роста существует определенный разброс веса. Разброс этот, кстати, можно оценить, рассчитав стандартное отклонение веса для каждого роста. Оказывается, какой бы рост мы ни взяли, стандартное отклонение веса составит 1 г, что заметно меньше стандартного отклонения веса для всей, не разделенной по весам, совокупности марсиан.

УРАВНЕНИЕ РЕГРЕССИИ

Прежде чем перейти к обобщению этих закономерностей, дадим несколько определений. В уравнении регрессии одна из переменных, x, называется *независимой переменной*, а другая, y, — *зависимой*. Набор значений y, соответствующих определенному значению x, обозначим $y|x$.

В примере с марсианами рост мы будем рассматривать как независимую переменную, а вес — как зависимую. Понятно, что это не означает, что одна переменная действительно *определяет* другую. Просто по значению одного признака мы *предсказываем* значение второго. В условиях эксперимента мы произвольно меняем независимую переменную и смотрим, как меняется зависимая. При этом речь действительно идет о зависимости, то есть о причинной связи. В прочих же случаях выявление *статистической* связи двух переменных указывает на возможность *причинной* связи, но не доказывает ее. Разобраться в причинах и следствиях вообще невозможно чисто статистическими методами. Необходимо, в частности, найти биологический механизм, порождающий выявленную связь. Например, эпидемиологические данные о связи пассивного курения с заболеваемостью ишемической болезнью сердца еще не доказывают, что пассивное курение способствует развитию ИБС. Может быть, и то и другое — следствие какой-либо неизвестной причины, например нервной обстановки в рабочем коллективе. Однако экспериментальные данные* о том, что пассивное курение и отдельные компоненты та-

* О том, как анализировать совокупность эпидемиологических и экспериментальных данных для выявления причинных связей, можно прочесть в работах: S. A. Glantz, W. W. Parmley. Passive smoking and

бачного дыма вызывают поражение сердца у лабораторных животных, говорят в пользу именно причинной связи.

Вернемся к нашим марсианам. Для каждого значения независимой переменной x (в нашем примере это рост) рассчитаем среднее значение зависимой переменной y (вес). Это среднее в точке x обозначим $\mu_{y|x}$. Тогда обнаруженная нами линейная зависимость описывается уравнением

$$\mu_{y|x} = \alpha + \beta x.$$

Здесь α — значение y в точке $x = 0$ (коэффициент сдвига), β — коэффициент наклона*. В нашем примере при увеличении роста на 1 см средний вес увеличивается на 0,5 г, поэтому $\beta = 0{,}5$. Хотя представить марсиан весом –8 г не легче, чем ростом 0 см, тем не менее для прямой с рис. 8.2 имеем $\alpha = -8$ г. Таким образом, прямая средних (для каждого роста) весов задается формулой

$$\mu_{y|x} = -8 + 0{,}5x.$$

Теперь посмотрим, как распределены веса марсиан одного роста. В данном случае это *нормальное распределение* со средним $\mu_{y|x}$ и стандартным отклонением $\sigma_{y|x}$. Но этого еще недостаточно для применения методов, которые мы рассмотрим ниже. Помимо нормальности распределения требуется, чтобы $\sigma_{y|x}$ было *одинаковым* для разных x. Иначе говоря разброс значений зависимой случайной переменной y должен быть неизменным при любом значении независимой переменной x. В нашем примере это условие выполняется.

Итак, значения переменных должны удовлетворять следующим условиям.

- Среднее значение $\mu_{y|x}$ линейно зависит от x.
- Для любого значения x значения $y|x$ распределены нормально.
- Стандартное отклонение $\sigma_{y|x}$ одинаково при всех значениях x.

Функция, задающая зависимость $\mu_{y|x}$ от x, определяется па-

---

heart disease: epidemiology, physiology, and biochemistry. *Circulation*, 83:1—12, 1991 и S. A. Glantz, W. W. Parmley. Passive smoking and heart disease: mechanisms and risk. *JAMA*, 273:1047—1053, 1995.

* Эти обозначения совпадают с обозначениями ошибок I и II рода. Будем надеятся, что это не породит путаницы.

раметрами α и β. Разброс значений $y|x$ в точке x задается стандартным отклонением $\sigma_{y|x}$. Оценим эти параметры.

ОЦЕНКА ПАРАМЕТРОВ УРАВНЕНИЯ РЕГРЕСИИ ПО ВЫБОРКЕ

В реальной жизни редко удается получить данные обо всей совокупности, и исследователю приходится довольствоваться выборками. Допустим, мы располагали бы данными не о всех марсианах, а только о десяти. На рис. 8.3А они показаны черными кружками среди 190 своих собратьев. На рис. 8.3Б данные показаны так, как их видит исследователь, изучивший эту выборку. Что можно сказать о совокупности, основываясь на этих выборочных данных?

Похоже, что в этом случае исследователю повезло. Зависимость веса от роста в выборке выглядит примерно так же, как и в совокупности в целом. Но ведь выборка может вводить в заблуждение. Вспомним пример с рис. 1.2. В выборке из 5 человек диурез отчетливо увеличивался с ростом дозы препарата (рис 1.2А), тогда как на самом деле никакой зависимости не было (рис 1.2Б). Какова вероятность ошибочного заключения? Как мы скоро увидим, эта задача сводится к *оценке параметров уравнения регрессии α и β по выборке*.

Метод наименьших квадратов

Сейчас нам предстоит оценить параметры уравнения регрессии α и β. Обозначим их выборочные оценки соответственно a и b. Найти наилучшие оценки этих параметров — это то же самое, что провести наилучшую прямую через имеющиеся точки, поскольку $y = a + bx$ — это уравнение прямой. Какую прямую считать наилучшей? Посмотрим на рис. 8.4. На нем изображены 4 прямые. Прямая I явно не годится — все точки оказались по одну сторону от нее. Прямая II немного лучше, она хотя бы пересекает область, где находятся наши точки. Однако она слишком круто устремляется вверх. Какая из прямых III и IV является лучшей, сказать трудно. Почему прямая II кажется лучше прямой I, а прямая III — лучше прямой II? Очевидно, прямая тем лучше,

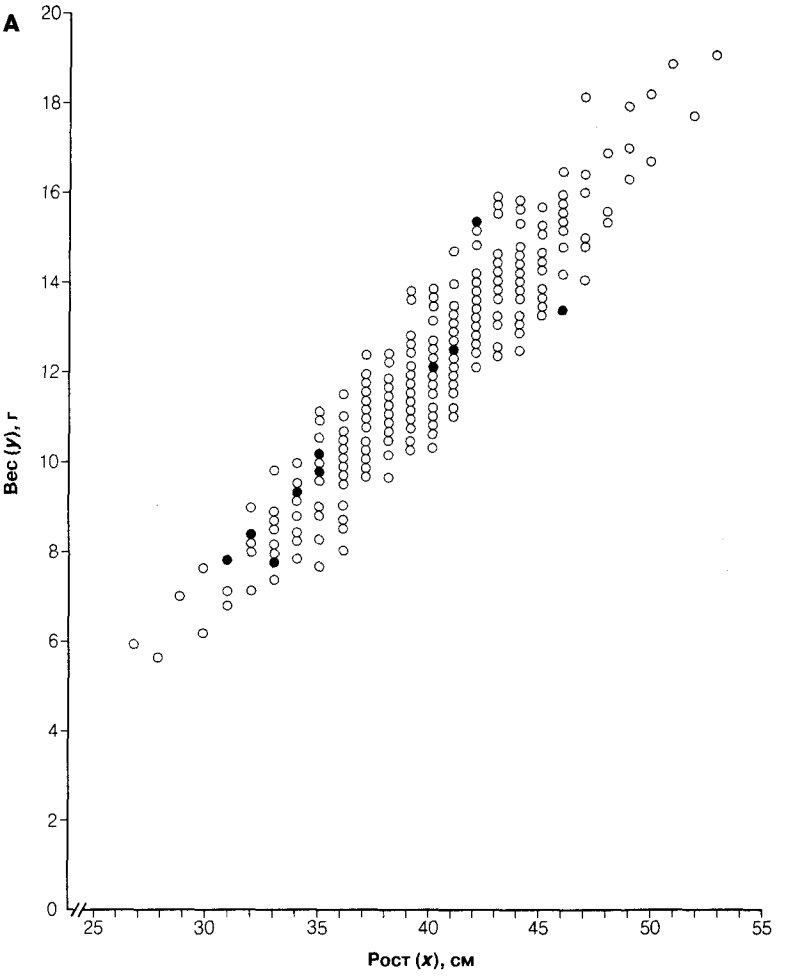


Рис. 8.3. А. Случайная выборка объемом 10 из совокупности марсиан.

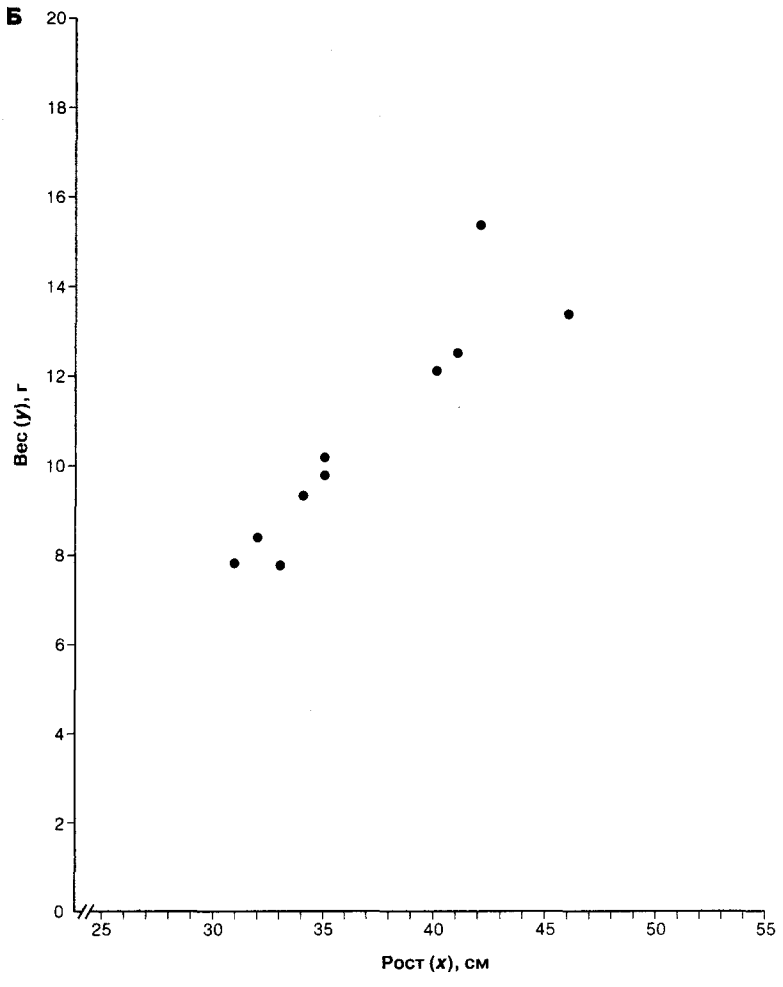


Рис. 8.3. Б. Такой эта выборка представляется исследователю, который не может наблюдать всю совокупность.

чем ближе она ко всем точкам выборки. Иными словами, лучше та прямая, относительно которой разброс точек минимален.

С оценкой разброса мы уже сталкивались в гл. 2. Там мы использовали средний квадрат отклонения от среднего. Поступим аналогичным образом. Определим расстояние по вертикали от каждой точки до прямой (рис. 8.5). Возведем полученные величины в квадрат и сложим. Возведение в квадрат потребовалось, чтобы отклонения, равные по абсолютной величине, но разные по знаку, вносили один и тот же вклад.

Сумма квадратов отклонений от прямой IV меньше, чем от прямой III. Следовательно, прямая IV лучше представляет зависимость y от x. Более того, можно доказать, что для прямой IV сумма квадратов отклонений выборочных значений зависимой переменной минимальна. Способ нахождения линии, сумма квадратов расстояний от которой до всех точек выборки минимальна, называется *методом наименьших квадратов*, саму линию мы будем называть *прямой регрессии*. Здесь мы не будем останавливаться на выводе формул* и сообщим сразу результат.

Напомним, что мы ищем параметры уравнения регрессии:

$$\hat{y} = a + bx.$$

Тогда коэффициент сдвига

$$a = \frac{(\Sigma Y)(\Sigma X^2) - (\Sigma X)(\Sigma XY)}{n(\Sigma X^2) - (\Sigma X)^2}$$

и коэффициент наклона

$$b = \frac{n(\Sigma XY) - (\Sigma X)(\Sigma Y)}{n(\Sigma X^2) - (\Sigma X)^2},$$

где X и Y — значения независимой и зависимой переменных у n членов выборки**.

---

* Интересующихся выводом этих формул отсылаем к книге: S. A. Glantz. Mathematics for biomedical applications. University of California Press, Berkely, 1979, pp. 322—325.

** Вычисления можно упростить, если сначала вычислить b, а уже потом найти a по формуле $a = \bar{Y} - b\bar{X}$, где $\bar{Y}$ и $\bar{X}$ — выборочные средние для переменных y и x.

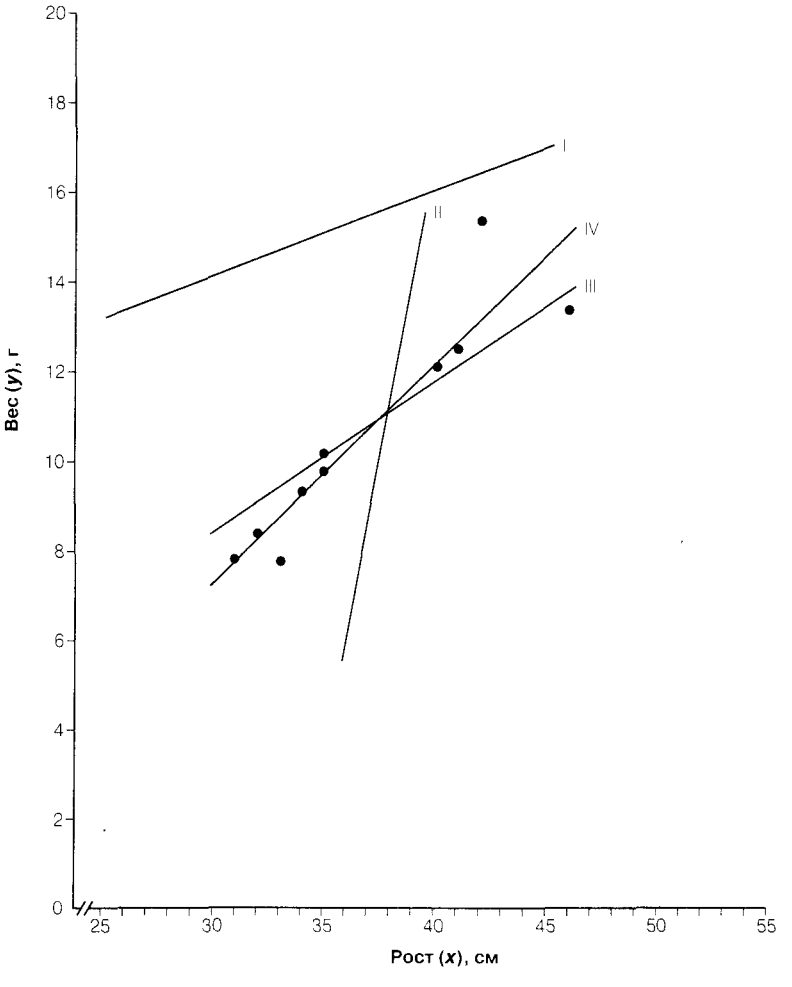


Рис. 8.4. Провести прямую через десять точек можно по-разному. Прямые I и II явно не годятся, прямые III и IV выглядят лучше.

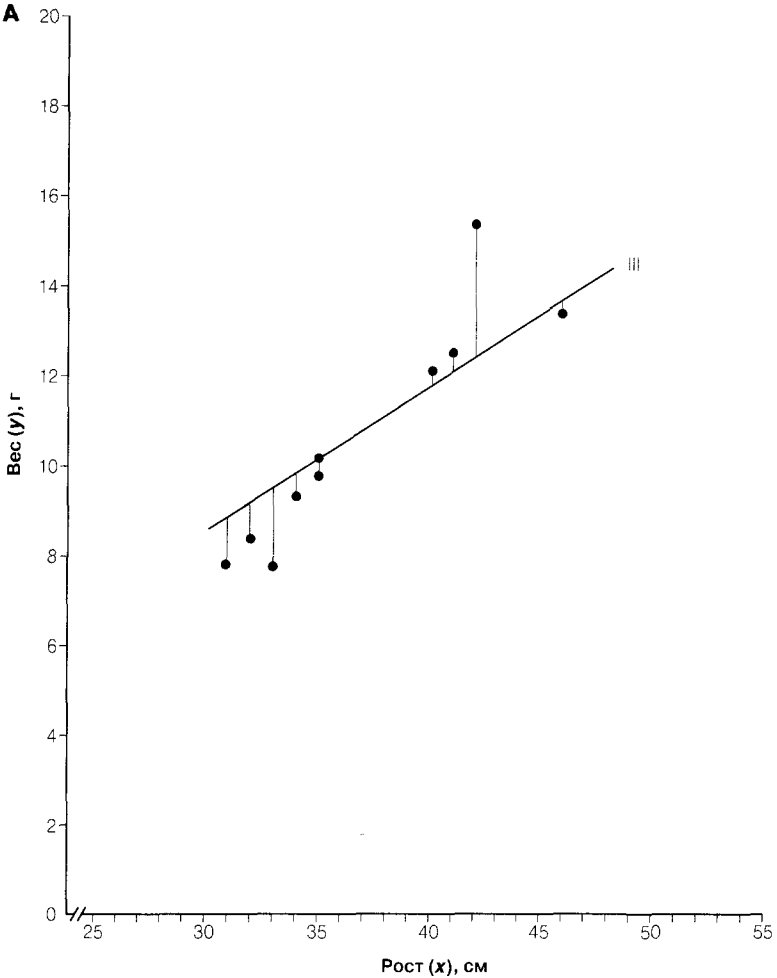


Рис. 8.5. Найдем расстояние по вертикали от каждой точки до прямой III (**А**) и IV (**Б**). Сумма квадратов расстояний до прямой IV меньше, чем до прямой III. Рядом с прямой IV серым цветом показана линия средних с рис. 8.2. Как видим, прямые достаточно близки.

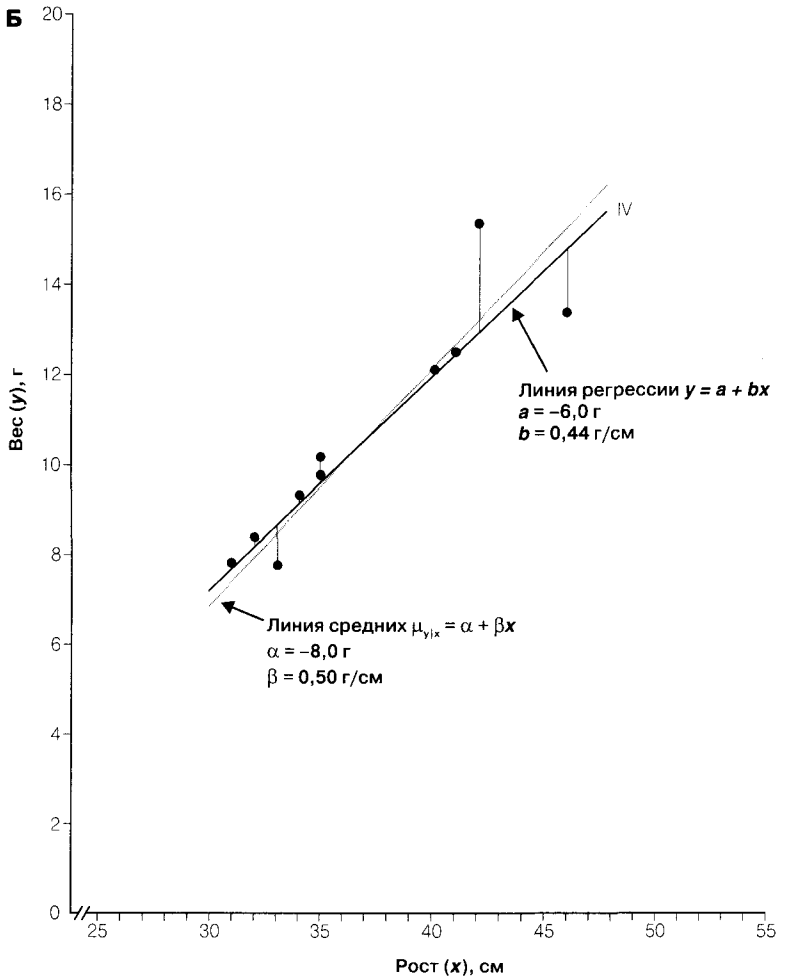


Рис. 8.5. Окончание

Таблица. 8.1. Расчет параметров уравнения регрессии

X	Y	X^2	XY
31	7,8	961	241,8
32	8,3	1024	265,6
33	7,6	1089	250,8
34	9,1	1156	309,4
35	9,6	1225	336,0
35	9,8	1225	343,0
40	11,8	1600	472,0
41	12,1	1681	496,1
42	14,7	1764	617,4
46	13,0	2116	598,0
369	103,8	13841	3930,1

Рассчитаем параметры уравнения регрессии для нашей выборки из 10 марсиан. Вспомогательные величины для вычислений приведены в табл. 8.1. Объем выборки $n = 10$, $\Sigma X = 369$, $\Sigma Y = 103{,}8$, $\Sigma X^2 = 13841$ и $\Sigma XY = 3930{,}1$. Подставим эти числа в формулы для коэффициентов регрессии:

$$a = \frac{103{,}8 \times 13841 - 369 \times 3930{,}1}{10 \times 13841 - 369^2} = -6{,}0$$

и

$$b = \frac{10 \times 3930{,}1 - 369 \times 103{,}8}{10 \times 13841 - 369^2} = 0{,}44.$$

Таким образом, прямая регрессии имеет вид:

$$\hat{y} = -6{,}0 + 0{,}44x.$$

Именно это уравнение задает прямую IV.

Разброс значений вокруг прямой регрессии

Мы получили a и b — оценки коэффициентов регрессии α и β. Хорошо бы получить также оценку разброса значений вокруг прямой регрессии. При каждом значении X стандартное отклонение постоянно и равно $\sigma_{y|x}$. Выборочной оценкой $\sigma_{y|x}$ служит

$$s_{y|x} = \sqrt{\frac{\Sigma[Y-(a+bX)]^2}{n-2}},$$

где $a+bX$ — значение уравнения регрессии в точке X, $Y-(a+bX)$ — расстояние от точки до прямой регрессии, Σ обозначает суммирование квадратов этих расстояний. Не будем объяснять, почему сумма квадратов отклонений должна быть поделена на $n-2$, а не на n или $n-1$. Скажем только, что причина аналогична той, по которой в оценке стандартного отклонения делитель равен $n-1$.

Величина $s_{y|x}$ называется *остаточным стандартным отклонением* (соответственно $s^2_{y|x}$ называется *остаточной дисперсией*). Связь $s_{y|x}$ со стандартными отклонениями s_Y и s_X зависимой и независимой переменных определяется формулой

$$s_{y|x} = \sqrt{\frac{n-1}{n-2}(s_Y^2 - b^2 s_X^2)}.$$

Для рассмотренной нами выборки $s_X = 5{,}0$, $s_Y = 2{,}4$. Тогда

$$s_{y|x} = \sqrt{\frac{9}{8}(2{,}4^2 - 0{,}44^2 \times 5{,}0^2)} = 1{,}02.$$

Как видим, оценка $s_{y|x}$ оказалась близкой к истинному значению $\sigma_{y|x}$, равному 1,0 г.

Стандартные ошибки коэффициентов регрессии

Подобно тому как выборочное среднее — это оценка истинного среднего (среднего по совокупности), так и выборочные параметры уравнения регрессии a и b — не более чем оценки истинных коэффициентов регрессии α и β. Разные выборки дают разные оценки среднего — точно так же разные выборки будут давать разные оценки коэффициентов регрессии. Для выборки с рис. 8.3 мы получили значения $a = -6{,}0$ и $b = 0{,}44$. Рассмотрим другую выборку из той же совокупности (рис. 8.6А). На рис. 8.6Б эта выборка показана такой, какой ее видит исследователь. Общая закономерность осталась прежней — высокие марсиане ве-

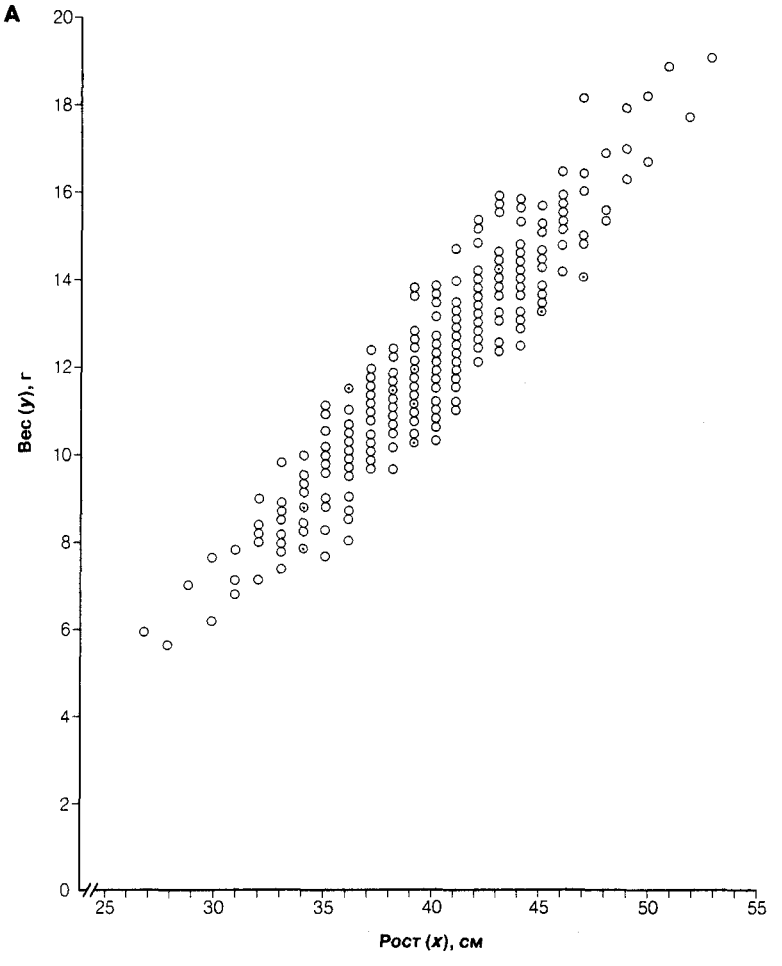


Рис. 8.6. А. Еще одна случайная выборка объемом 10 из совокупности марсиан. Марсиане, попавшие в выборку, помечены точками.

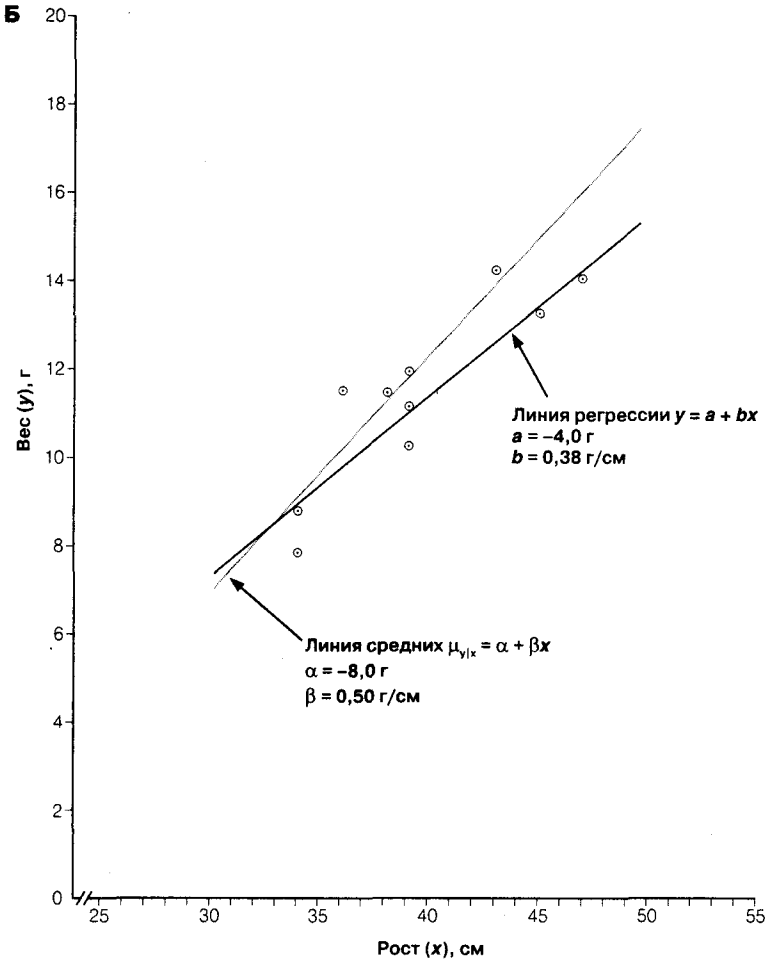


Рис. 8.6. Б. Линия регрессии, рассчитанная по этой выборке, несколько отличается от полученной ранее (см. рис. 8.5Б). Серым показана линия средних с рис. 8.2.

сят больше низкорослых. Однако, рассчитав коэффициенты регрессии, получим $a = -4{,}0$ г и $b = 0{,}38$ г/см.

Если построить все возможные выборки по 10 марсиан в каждой, получится совокупность всех значений a и b. Их средние равны α и β, а стандартные отклонения — σ_α и σ_β. Эти стандартные отклонения называются *стандартными ошибками коэффициентов регрессии*. Стандартные ошибки коэффициентов регрессии, подобно стандартной ошибке среднего или доли, используются при проверке гипотез и вычислении доверительных интервалов. Выборочные оценки для σ_α и σ_β обозначаются соответственно s_a и s_b и вычисляются по следующим формулам*:

$$s_a = s_{y|x}\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_X^2}}$$

и

$$s_b = \frac{1}{\sqrt{n-1}}\frac{s_{y|x}}{s_X}.$$

Для выборки с рис. 8.3Б имеем:

$$s_a = 1{,}02\sqrt{\frac{1}{10} + \frac{36{,}9^2}{(10-1)5{,}0^2}} = 2{,}53$$

и

$$s_b = \frac{1}{\sqrt{10-1}}\frac{1{,}02}{5{,}0} = 0{,}068.$$

Стандартные ошибки коэффициентов регрессии используются аналогично стандартной ошибке среднего – для нахождения доверительных интервалов и проверки гипотез.

---

* Вывод формул для стандартных ошибок коэффициентов регрессии можно найти в большинстве учебников статистики. См., например, J. Neter and W. Wasserman. Applied statistical models. Irwin, Homewood, III., 1974, chap. 3, «Inferences in regression analysis».

Есть ли зависимость?

Помня о досадном недоразумении с «диуретиком» из гл. 1 (см. рис. 1.2), исследователь вправе спросить: как убедиться, что зависимость действительно существует? Иными словами, как по выборочным данным определить вероятность P нулевой гипотезы о том, что коэффициент наклона $\beta = 0$*?

Совокупность всех выборочных значений коэффициента наклона b приближенно подчиняется нормальному распределению. Поэтому можно воспользоваться критерием Стьюдента, аналогично тому, как мы пользовались им в гл. 4 для проверки гипотезы относительно среднего. В общем виде критерий Стьюдента можно определить как:

$$t = \frac{\text{Выборочная оценка} - \text{Истинная величина}}{\text{Стандартная ошибка выборочной оценки}}.$$

Для оценки коэффициента наклона:

$$t = \frac{b - \beta}{s_b}.$$

Оценить вероятность гипотезы о равенстве $\beta = 0$ можно двумя способами.

Приравняв β к нулю, имеем

$$t = \frac{b}{s_b}.$$

Теперь по табл. 4.1 найдем t_α — критическое значение t для выбранного уровня значимости α и числа степеней свободы $\nu = n - 2$. Если полученное значение t по абсолютной величине превосходит t_α, то $P < \alpha$, то есть зависимость статистически значима.

Потренируемся на марсианах. Для выборки с рис. 8.3Б мы нашли $b = 0{,}44$ и $s_b = 0{,}068$. Тогда $t = 0{,}44/0{,}068 = 6{,}47$. Объем выборки равен 10. Положим уровень значимости равным 0,001. В табл. 4.1 для этого уровня значимости и числа степеней свободы

---

* Речь идет исключительно о линейной зависимости. Как мы вскоре увидим, зависимость может быть и нелинейной; в таком случае излагаемый способ даст неправильный результат.

$\nu = 10 - 2 = 8$ находим критическое значение $t_\alpha = 5{,}041$. Поскольку $t > t_\alpha$, гипотезу об отсутствии зависимости веса от роста следует отвергнуть.

Конечно, как и всегда при проверке гипотез, это заключение может оказаться ложным (опять-таки вспоминается злополучный диуретик из гл. 1). Но вероятность совершить эту ошибку не превышает 0,001.

Второй способ основан на использовании доверительных интервалов. $100(1-\alpha)$-процентный доверительный интервал для β имеет вид

$$b - t_\alpha s_b < \beta < b + t_\alpha s_b.$$

Рассчитаем 95% доверительный интервал. Число степеней свободы $\nu = 10 - 2 = 8$. По таблице 4.1 находим $t_{0,05} = 2{,}306$. Выборочные значения $b = 0{,}44$ и $s_b = 0{,}068$. Следовательно, доверительный интервал для β:

$$0{,}44 - 2{,}306 \times 0{,}068 < \beta < 0{,}44 + 2{,}306 \times 0{,}068,$$

$$0{,}28 < \beta < 0{,}60.$$

Поскольку ноль в этот интервал не попадает, вероятность того, что $\beta = 0$, меньше 5%.

Если рассчитать 99,9% доверительный интервал, можно убедиться, что и он не содержит нуля. Вывод, полученный выше при использовании критерия Стьюдента, как и следовало ожидать, совпадает с полученным с помощью доверительного интервала. Заметим, что истинное значение $\beta = 0{,}5$ попадает в доверительный интервал.

Можно вычислить доверительный интервал и для коэффициента α. Например, 95% доверительный интервал имеет вид:

$$a - t_{0,05} s_a < \alpha < a + t_{0,05} s_a,$$

то есть

$$-6{,}0 - 2{,}306 \times 2{,}53 < \alpha < -6{,}0 + 2{,}306 \times 2{,}53,$$

$$-11{,}8 < \alpha < -0{,}17.$$

Интервал покрывает истинное значение $\alpha = -8$ г.

Следующим этапом будет построение доверительной области для линии регрессии и значений зависимой переменной.

Доверительная область для линии регрессии

Обычно мы не знаем истинных величин коэффициентов регрессии α и β. Нам известны только их оценки a и b. Иначе говоря, истинная прямая регрессии может пройти выше или ниже, быть более крутой или пологой, чем построенная по выборочным данным. Мы вычислили доверительные интервалы для коэффициентов регрессии. Можно вычислить доверительную область и для самой линии регрессии. На рис. 8.7А показана 95% доверительная область для выборки с рис. 8.3. Как видим, это довольно узкая полоса, которая несколько расширяется при крайних значениях x.

Мы знаем, что при любом значении независимой переменной x соответствующие значения зависимой переменной y распределены нормально. Средним является значение уравнения регрессии $\hat{y}$. Неопределенность его оценки характеризуется стандартной ошибкой регрессии:

$$s_{\hat{y}} = s_{y|x}\sqrt{\frac{1}{n}+\frac{(x-\bar{X})^2}{(n-1)s_X^2}}.$$

В отличие от стандартных ошибок, с которыми мы имели дело до сих пор, $s_{\hat{y}}$ при разных x принимает разные значения: чем дальше x от выборочного среднего $\bar{X}$, тем она больше.

Теперь можно вычислить $100(1-\alpha)$-процентный доверительный интервал для значения уравнения регрессии в точке x:

$$\hat{y} - t_\alpha s_{\hat{y}} < y < \hat{y} + t_\alpha s_{\hat{y}},$$

где t_α — критическое значение с $\nu = n-2$ степенями свободы, а $\hat{y}$ — значение уравнения регрессии в точке x:

$$\hat{y} = a + bx.$$

Итак, мы получили уравнение для кривых, ограничивающих доверительную область линии регрессии (см. рис. 8.3). С заданной вероятностью, обычно 95%, можно утверждать, что истин-

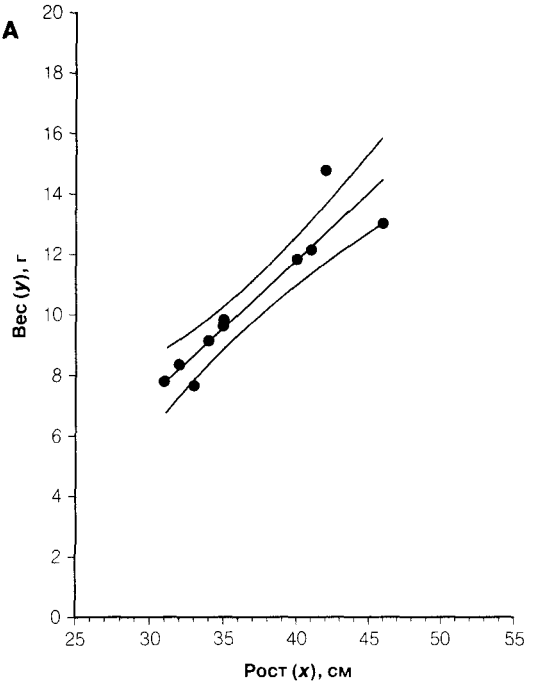


Рис. 8.7. А. 95% доверительная область для *линии регрессии* (по выборке с *рис. 8.3*).

ная линия находится где-то внутри этой области. Обратите внимание, что три точки из десяти оказались вне доверительной области. Это совершенно естественно, поскольку речь идет о доверительной области линии регресии, а не самих значений (доверительная область для значений гораздо шире).

Авторы медицинских публикаций нередко приводят доверительную область линии регрессии и говорят о ней так, как будто это — доверительная область значений. Это примерно то же самое, что выдавать стандартную ошибку среднего за характеристику разброса значений, путая ее со стандартным отклонением. Например, из рис. 8.7А видно, что *средний* вес марсиан ростом 40 см с вероятностью 95% окажется между 11,0 и 12,5 г — из этого

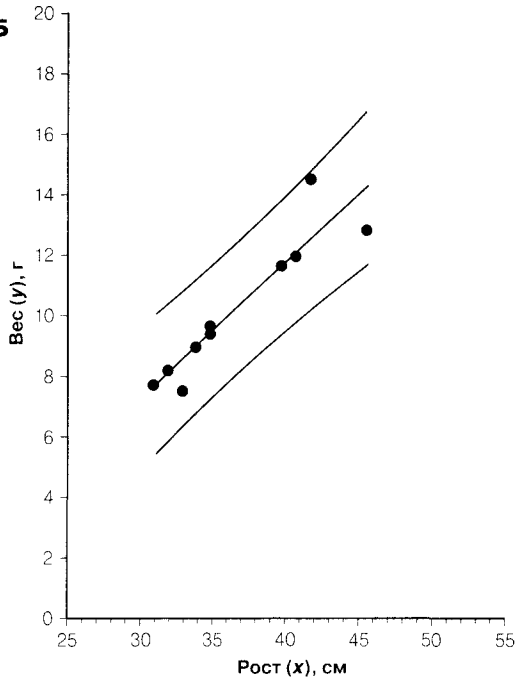


Рис. 8.7. Б. 95% доверительная область для *значений*. Если мы хотим определить вес марсианина по его росту, нам следует воспользоваться именно этой доверительной областью.

вовсе не следует, что в этих пределах окажется вес 95% марсиан такого роста.

Теперь займемся доверительной областью для *значений* зависимой переменной.

Доверительная область для значений

Разброс значений складывается из разброса значений вокруг линии регресии и неопределенности положения самой этой линии. Характеристикой разброса значений вокруг линии регрессии является остаточное стандартное отклонение $s_{y|x}$, а неопределен-

ности положения линии регрессии — стандартная ошибка регрессии $s_{\hat{y}}$. Дисперсия суммы двух величин равна сумме дисперсий, поэтому

$$s_Y = \sqrt{s_{y|x}^2 + s_{\hat{y}}^2}.$$

Подставив в эту формулу выражение для $s_{\hat{y}}$ из предыдущего раздела, получим:

$$s_Y = s_{y|x}\sqrt{1 + \frac{1}{n} + \frac{(x - \bar{X})^2}{(n-1)s_X^2}}.$$

Тогда $100(1-\alpha)$-процентный доверительный интервал для зависимой переменной

$$\hat{y} - t_\alpha s_Y < y < \hat{y} + t_\alpha s_Y.$$

Заметьте, что входящие в это неравенство величины $\hat{y}$ и s_Y зависят от x.

На рис. 8.7Б изображена полученная по этой формуле 95% доверительная область для значений зависимой переменной. В эту область попадет 95% всех возможных значений веса марсиан любого роста. Например, с вероятностью 95% можно утверждать, что любой 40-сантиметровый марсианин весит от 9,5 до 14,0 г.

СРАВНЕНИЕ ДВУХ ЛИНИЙ РЕГРЕССИИ

Часто требуется сравнить линии регрессии, рассчитанные по двум выборкам. Это можно сделать тремя способами.

- Сравнить коэффициенты наклона b.
- Сравнить коэффициенты сдвига a.
- Сравнить линии в целом.

В первых двух случаях следует воспользоваться критерием Стьюдента. Если нужно проверить, значимо ли различие в наклоне двух прямых регрессии, критерий Стьюдента t вычисляется по формуле:

$$t = \frac{b_1 - b_2}{s_{b_1 - b_2}},$$

где $b_1 - b_2$ — разность коэффициентов наклона, а $s_{b_1 - b_2}$ — ее стандартная ошибка. Затем вычисленное t сравним, как обычно, с критическим значением t_α, имеющим $(n_1 - 2) + (n_2 - 2) = n_1 + n_2 - 4$ степени свободы.

Если обе регрессии оценены по одинаковому числу наблюдений, то стандартная ошибка разности

$$s_{b_1 - b_2} = \sqrt{s_{b_1}^2 + s_{b_2}^2}.$$

Если же объемы выборок различны, следует воспользоваться объединенной оценкой остаточной дисперсии (она аналогична объединенной оценке дисперсии, приведенной в гл. 4):

$$s_{y|x_{\text{общ}}}^2 = \frac{(n_1 - 2)s_{y|x_1}^2 + (n_2 - 2)s_{y|x_2}^2}{n_1 + n_2 - 4}.$$

Тогда формула для $s_{b_1 - b_2}$ принимает вид

$$s_{b_1 - b_2} = \sqrt{\frac{s_{y|x_{\text{общ}}}^2}{(n_1 - 1)s_{x_1}^2} + \frac{s_{y|x_{\text{общ}}}^2}{(n_2 - 1)s_{x_2}^2}}.$$

Можно сравнить и коэффициенты сдвига a_1 и a_2. В этом случае

$$t = \frac{a_1 - a_2}{s_{a_1 - a_2}}.$$

Здесь

$$s_{a_1 - a_2} = \sqrt{s_{a_1}^2 + s_{a_2}^2},$$

когда обе регрессии вычислены по одинаковому числу точек. При неодинаковом числе точек следует воспользоваться объединенной оценкой дисперсии так же, как это было сделано выше.

Перейдем к сравнению двух линий регресии в целом. Сравнить две линии регрессии — значит оценить вероятность нуле-

вой гипотезы о совпадении линий*. Напомним, что коэффициенты регрессии вычисляются так, чтобы разброс точек вокруг линии регрессии был минимален. Разброс этот характеризуется остаточной дисперсией $s^2_{y|x}$: чем меньше остаточная дисперсия, тем лучше прямая регрессии соответствует имеющимся точкам. Воспользуемся этим показателем для оценки результатов такого мысленного эксперимента. Объединим обе выборки в одну и построим для нее линию регрессии. Если линии регрессии для двух выборок близки, остаточная дисперсия при этом существенно не изменится. И наоборот, если они различаются, то совпадение точек и линии ухудшится и остаточная дисперсия возрастет.

Порядок действий таков.

- Построить прямую регресии для каждой из выборок.
- По остаточным дисперсиям $s^2_{y|x_1}$ и $s^2_{y|x_2}$ каждой из регрессий вычислить объединенную оценку остаточной дисперсии $s^2_{y|x_{общ}}$.
- Объединить обе выборки. Построить прямую регрессии для получившейся выборки и вычислить остаточную дисперсию $s^2_{y|x_{един}}$.
- Вычислить «выигрыш» от использования двух раздельных регрессий. Мерой выигрыша служит величина:

$$s^2_{y|x_в} = \frac{(n_1+n_2-2)s^2_{y|x_{един}} - (n_1+n_2-4)s^2_{y|x_{общ}}}{2}.$$

- По $s^2_{y|x_в}$ и $s^2_{y|x_{общ}}$ вычислить критерий F:

$$F = \frac{s^2_{y|x_в}}{s^2_{y|x_{общ}}}.$$

- Сравнить вычисленное значение с критическим значением F для числа степеней свободы $\nu_{меж} = 2$ и $\nu_{вну} = n_1 + n_2 - 4$. Если полученное значение больше критического, то гипотеза о совпадении линий регрессии должна быть отклонена.

---

* Методы, предназначенные для сравнения более чем двух линий регрессии, описаны в книге: J. H. Zar. Biostatistical analysis. 2nd ed. Prentice-Hall, Englewood Cliffs, N. J., 1984.

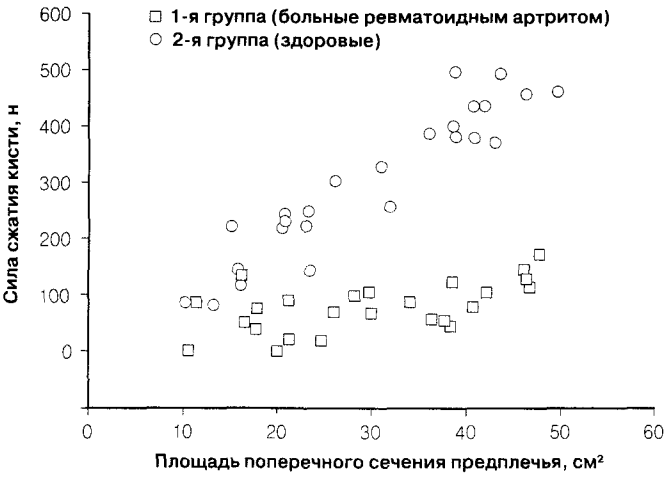


Рис. 8.8. Зависимость мышечной силы от мышечной массы. Здоровые обозначены кружками, больные ревматоидным артритом — квадратиками. Одинакова ли зависимость у больных и здоровых?

Мышечная сила при ревматоидном артрите

Причины ограниченной подвижности при ревматоидном артрите разнообразны: болезненность суставов, их тугоподвижность, атрофия мышц. Каков вклад каждого из этих факторов? Пытаясь ответить на этот вопрос, П. С. Хелливелл и С. Джексон* исследовали, в частности, связь между мышечной массой и силой. В исследовании приняли участие 25 больных ревматоидным артритом (1-я группа) и 25 здоровых (2-я группа). Рассчитывали площадь поперечного сечения предплечья и ручным динамометром определяли силу сжатия кисти. Результат показан на рис. 8.8. Кружки — результаты здоровых, квадратики — больных ревматоидным артритом.

На рис. 8.9А представлены те же наблюдения, что и на рис. 8.8, и кроме того, две построенные по ним линии регрессии. Проверим, есть ли значимое различие между линиями регрес-

* P. S. Helliwell, S. Jackson. Relationship between weakness and muscle wasting in rheumatoid arthritis. *Ann. Rheum. Dis.*, 53:726—728, 1994.

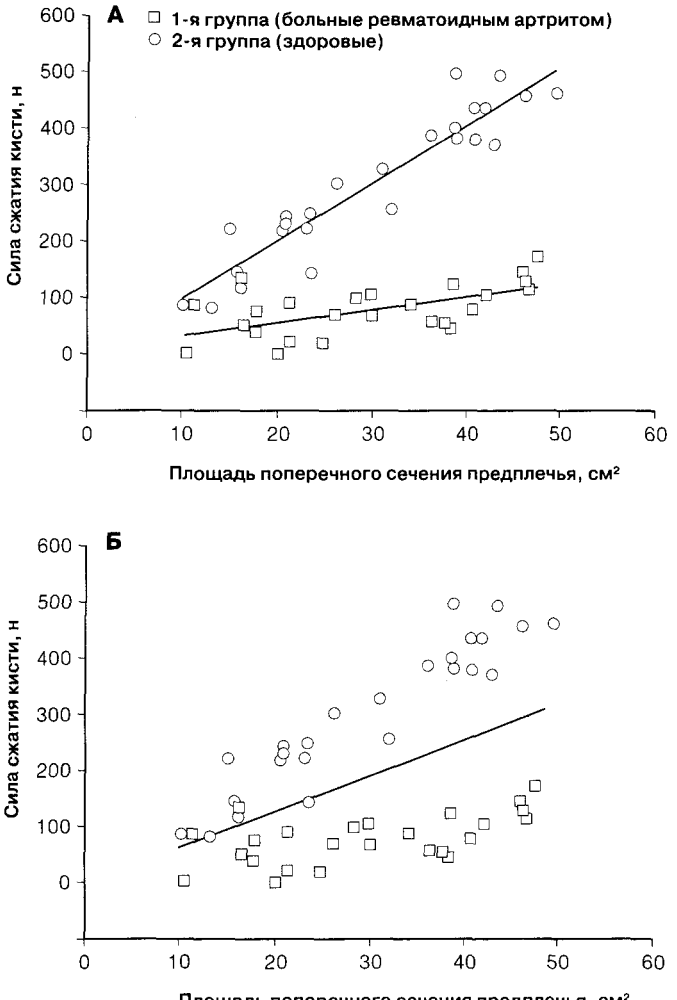


Рис. 8.9. А. Построим линии регрессии для каждой из групп и оценим разброс точек относительно этих линий. **Б.** Объединим группы и найдем линию регрессии для получившейся группы. Если разброс точек относительно этой линии значительно превышает разброс относительно двух отдельных линий, то *различия линий следует считать значимыми.*

Таблица 8.2. Зависимость силы сжатия кисти от мышечной массы

	1-я группа	2-я группа	Объединенная группа
Численность группы	25	25	50
Коэффициенты регрессии			
сдвиг a (s_a)	3,3 (22,4)	−7,3 (25,3)	−23,1 (50,5)
наклон b (s_b)	2,41 (0,702)	10,19 (0,789)	6,39 (1,579)
Остаточное стандартное отклонение $s_{x\|y}$	40,5	45,7	129,1

сии. Параметры уравнений регрессии и остаточные стандартные отклонения указаны в табл. 8.2. Вычислим объединенную оценку остаточной дисперсии

$$s^2_{y|x_{\text{общ}}} = \frac{(n_1 - 2)s^2_{y|x_1} + (n_2 - 2)s^2_{y|x_2}}{n_1 + n_2 - 4},$$

где n_1 и n_2 — численность 1-й и 2-й групп, $s^2_{y|x_1}$ и $s^2_{y|x_2}$ — соответствующие остаточные дисперсии. Тогда

$$s^2_{y|x_{\text{общ}}} = \frac{(25-2)40{,}5^2 + (25-2)45{,}7^2}{25+25-4} = 1864.$$

Теперь объединим группы и найдем уравнение регрессии для получившейся группы. Опустим вычисления, результат приведен в табл. 8.2. Линия регрессии изображена на рис. 8.9Б. Остаточная дисперсия единой регрессии $s^2_{y|x_{\text{един}}} = 129{,}1^2 = 16667$. Выигрыш от использования раздельных регрессий:

$$s^2_{y|x_{\text{в}}} = \frac{(n_1 + n_2 - 2)s^2_{y|x_{\text{един}}} - (n_1 + n_2 - 4)s^2_{y|x_{\text{общ}}}}{2} =$$

$$= \frac{(25+25-2)16667 - (25+25-4)1864}{2} = 357136.$$

Значение F:

$$F = \frac{s^2_{y|x_{\text{в}}}}{s^2_{y|x_{\text{общ}}}} = \frac{357136}{1864} = 191{,}596.$$

Критическое значение F при уровне значимости $\alpha = 0{,}01$ и числе степеней свободы $\nu_{меж} = 2$ и $\nu_{вну} = 25+25-4$ равно 5,10, то есть гораздо меньше полученного нами. Таким образом, у здоровых людей сила сжатия зависит от размера предплечья иначе, чем у больных артритом.

В чем заключается отличие? Сравним коэффициенты регрессий. Начнем с коэффициента сдвига a.

$$s_{a_1-a_2} = \sqrt{s_{a_1}^2 + s_{a_2}^2} = \sqrt{22{,}4^2 + 25{,}3^2} = 33{,}8.$$

Тогда

$$t = \frac{a_1 - a_2}{s_{a_1-a_2}} = \frac{3{,}3-(-7{,}3)}{33{,}8} = 0{,}314.$$

При уровне значимости $\alpha = 0{,}05$ при числе степеней свободы $\nu = n_1 + n_2 - 4 = 46$ критическое значение t равно 2,013. Поскольку полученное нами значение t меньше критического, заключаем, что между a_1 и a_2 нет значимого различия.

При сравнении коэффициентов наклона получим $t = 7{,}367$, что больше критического. Итак, линии регрессии различаются наклоном, который круче в группе здоровых.

КОРРЕЛЯЦИЯ

Регрессионный анализ позволяет оценить, как одна переменная зависит от другой и каков разброс значений зависимой переменной вокруг прямой, определяющей зависимость. Эти оценки и соответствующие доверительные интервалы позволяют предсказать значение зависимой переменной и определить точность этого предсказания. Результаты регрессионного анализа можно представить только в достаточно сложной цифровой или графической форме. Однако нас часто интересует не предсказание значения одной переменной по значению другой, а просто характеристика тесноты (силы) связи между ними, при этом выраженная одним числом.

Эта характеристика называется *коэффициентом корреляции*, обычно ее обозначают буквой r. Коэффициент корреляции мо-

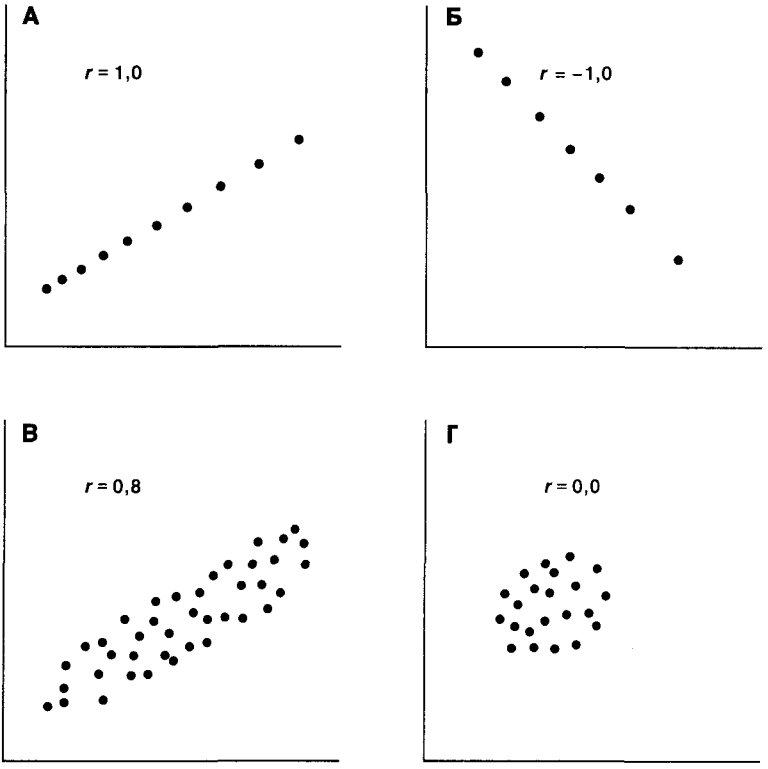


Рис. 8.10. Чем теснее связь между переменными, тем ближе абсолютная величина коэффициента корреляции к 1.

жет принимать значения от –1 до +1. Знак коэффициента корреляции показывает направление связи (прямая или обратная), а абсолютная величина — тесноту связи. Коэффициент, равный –1, определяет столь же жесткую связь, что и равный 1. В отсутствие связи коэффициент корреляции равен нулю.

На рис. 8.10 приведены примеры зависимостей и соответствующие им значения *r*. Мы рассмотрим два коэффициента корреляции.

Коэффициент корреляции Пирсона предназначен для описания линейной связи количественных признаков; как и регресси-

онный анализ, он требует нормальности распределения. Когда говорят просто о «коэффициенте корреляции», почти всегда имеют в виду коэффициент корреляции Пирсона, именно так мы и будем поступать.

Коэффициент ранговой корреляции Спирмена можно использовать, когда связь нелинейна — и не только для количественных, но и для порядковых признаков. Это непараметрический метод, он не требует какого-либо определенного типа распределения.

О количественных, качественных и порядковых признаках мы уже говорили в гл. 5. Количественные признаки — это обычные числовые данные, такие, как рост, вес, температура. Значения количественного признака можно сравнить между собой и сказать, какое из них больше, на сколько и во сколько раз. Например, если один марсианин весит 15 г, а другой 10, то первый тяжелее второго и в полтора раза и на 5 г. Значения порядкового признака тоже можно сравнить, сказав, какое из них больше, но нельзя сказать, ни на сколько, ни во сколько раз. В медицине порядковые признаки встречаются довольно часто. Например, результаты исследования влагалищного мазка по Папаниколау оценивают по такой шкале: 1) норма, 2) легкая дисплазия, 3) умеренная дисплазия, 4) тяжелая дисплазия, 5) рак *in situ*. И количественные, и порядковые признаки можно расположить по порядку — на этом общем свойстве основана большая группа непараметрических критериев, к которым относится и коэффициент ранговой корреляции Спирмена. С другими непараметрическими критериями мы познакомимся в гл. 10.

Коэффициент корреляции Пирсона

И все же, почему для описания тесноты связи нельзя воспользоваться регрессионным анализом? В качестве меры тесноты связи можно было бы использовать остаточное стандартное отклонение. Однако если поменять местами зависимую и независимую переменные, то остаточное стандартное отклонение, как и другие показатели регрессионного анализа, будет иным. Взглянем на рис. 8.11. По известной нам выборке из 10 марсиан построены две линии регрессии. В одном случае вес — зависимая переменная, во втором — независимая. Линии регрессии заметно раз-

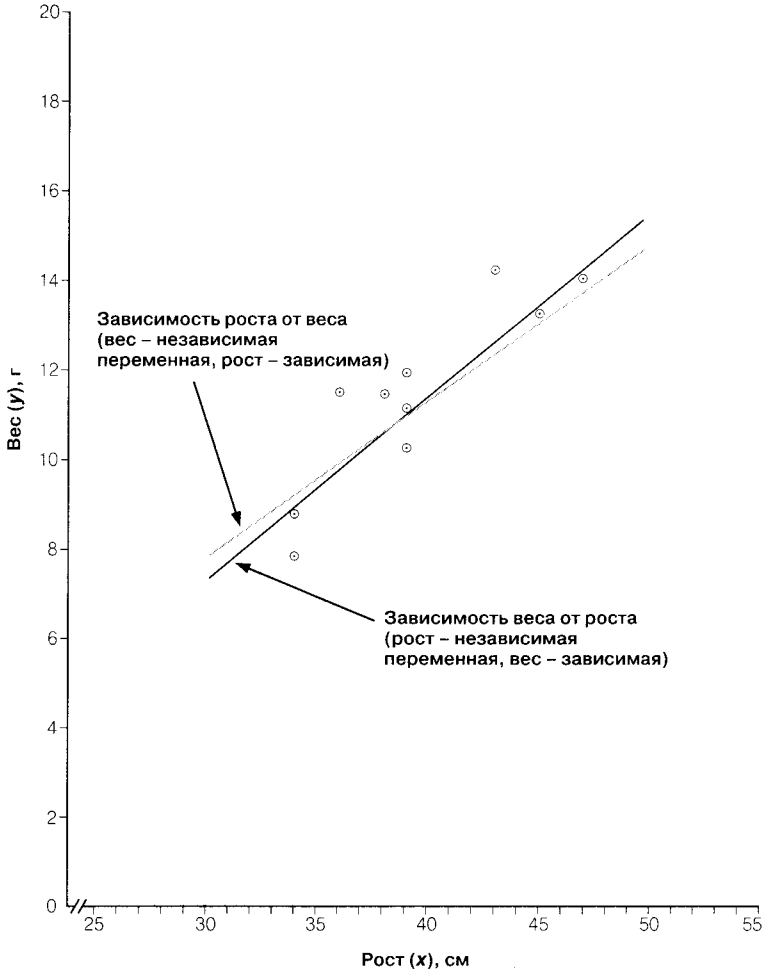


Рис. 8.11. Если поменять местами *x* и *y*, уравнение регрессии получится другим, а коэффициент корреляции останется прежним.

личаются. Получается, что связь роста с весом одна, а веса с ростом — другая. Асимметричность регрессионного анализа — вот что мешает непосредственно использовать его для характеристики силы связи. Коэффициент корреляции, хотя его идея вытекает из регрессионного анализа, свободен от этого недостатка. Приводим формулу.

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2 \, \Sigma(Y - \bar{Y})^2}},$$

где $\bar{X}$ и $\bar{Y}$ — средние значения переменных X и Y. Выражение для r «симметрично» —поменяв местами X и Y, мы получим ту же величину. Коэффициент корреляции принимает значения от –1 до +1. Чем теснее связь, тем больше абсолютная величина коэффициента корреляции. Знак показывает направление связи. При $r > 0$ говорят о прямой корреляции (с увеличением одной переменной другая также возрастает), при $r < 0$ — об обратной (с увеличением одной переменной другая уменьшается). Вернемся к рис. 8.10. На рис. 8.10А изображена максимально сильная прямая корреляция: $r = +1$. На рис. 8.10Б — максимально сильная обратная корреляция: $r = -1$. На рис. 8.10В корреляция прямая, тоже достаточно сильная: $r = 0{,}8$. Наконец, на рис. 8.10Г какая-либо связь между признаками отсутствует: $r = 0$.

Возьмем пример с 10 марсианами, который мы уже рассматривали с точки зрения регрессионного анализа. Вычислим коэффициент корреляции. Исходные данные и промежуточные результаты вычислений приведены в табл. 8.3. Объем выборки $n = 10$, средний рост $\bar{X} = \Sigma X/n = 369/10 = 36{,}9$ и вес $\bar{Y} = \Sigma Y/n = 103{,}8/10 = 10{,}38$. Находим $\Sigma(X - \bar{X})(Y - \bar{Y}) = 99{,}9$, $\Sigma(X - \bar{X})^2 = 224{,}8$, $\Sigma(Y - \bar{Y})^2 = 51{,}9$.

Подставим полученные значения в формулу для коэффициента корреляции:

$$r = \frac{99{,}9}{\sqrt{224{,}8 \times 51{,}9}} = 0{,}925.$$

Величина r близка к 1, что говорит о тесной связи роста и веса. Чтобы лучше представить себе, какой коэффициент корреляции следует считать большим, а какой незначительным, взгля-

Таблица 8.3. Вычисление коэффициента корреляции

X	Y	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})(Y-\bar{Y})$	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$
31	7,8	−5,9	−2,6	15,3	34,8	6,8
32	8,3	−4,9	−2,1	10,3	24,0	4,4
33	7,6	−3,9	−2,8	10,9	15,2	7,8
34	9,1	−2,9	−1,3	3,8	8,4	1,7
35	9,6	−1,9	−0,8	1,5	3,6	0,6
35	9,8	−1,9	−0,6	1,1	3,6	0,4
40	11,8	3,1	1,4	4,3	9,6	2,0
41	12,1	4,1	1,7	7,0	16,8	2,9
42	14,7	5,1	4,3	22,0	26,0	18,5
46	13,0	9,1	2,6	23,7	82,8	6,8
369	103,8	0,0	0,2	99,9	224,8	51,9

ните на табл. 8.4 — в ней приведены коэффициенты корреляции для примеров, которые мы разбирали ранее.

Связь регрессии и корреляции

Все примеры коэффициентов корреляции (табл. 8.4) мы первоначально использовали для построения линий регрессии. Действительно, между коэффициентом корреляции и параметрами регрессионного анализа существует тесная связь, которую мы сейчас продемонстрируем. Разные способы представления коэффициента корреляции, которые мы при этом получим, позволят лучше понять смысл этого показателя.

Вспомним, что уравнение регрессии строится так, чтобы минимизировать сумму квадратов отклонений от линии регрессии.

Таблица 8.4. Примеры корреляций

Пример	Коэффициент корреляции r	Объем выборки n
Рост и вес марсиан (рис. 8.7)	0,925	10
Сила сжатия кисти и мышечная масса у здоровых (рис. 8.9А)	0,938	25
Сила сжатия кисти и мышечная масса, объединенная группа (рис. 8.9Б)	0,581	50

Обозначим эту минимальную сумму квадратов $S_{ост}$ (эту величину называют остаточной суммой квадратов). Сумму квадратов отклонений значений зависимой переменной Y от ее среднего $\overline{Y}$ обозначим $S_{общ}$. Тогда:

$$r^2 = 1 - \frac{S_{ост}}{S_{общ}}.$$

Величина r^2 называется *коэффициентом детерминации* — это просто квадрат коэффициента корреляции. Коэффициент детерминации показывает силу связи, но не ее направленность.

Из приведенной формулы видно, что если значения зависимой переменной лежат на прямой регрессии, то $S_{ост} = 0$, и тем самым $r = +1$ или $r = -1$, то есть существует линейная связь зависимой и независимой переменной. По любому значению независимой переменной можно совершенно точно предсказать значение зависимой переменной. Напротив, если переменные вообще не связаны между собой, то $S_{ост} = S_{общ}$. Тогда $r = 0$.

Видно также, что коэффициент детерминации равен той доле общей дисперсии $S_{общ}$, которая обусловлена или, как говорят, объясняется линейной регрессией*.

Остаточная сумма квадратов $S_{ост}$ связана с остаточной дисперсией $s^2_{y|x}$ соотношением $S_{ост} = (n-2)s^2_{y|x}$, а общая сумма квадратов $S_{общ}$ с дисперсией s^2_y соотношением $S_{общ} = (n-1)s^2_y$. В таком случае

$$r^2 = 1 - \frac{n-2}{n-1}\frac{s^2_{y|x}}{s^2_y}.$$

Эта формула позволяет судить о зависимости коэффициента корреляции от доли остаточной дисперсии в полной дисперсии $s^2_{y|x}/s^2_y$. Чем эта доля меньше, тем больше (по абсолютной величине) коэффициент корреляции, и наоборот.

Мы убедились, что коэффициент корреляции отражает тесноту линейной связи переменных. Однако если речь идет о предсказании значения одной переменной по значению другой, на

---

* Следует помнить, что в статистике слова «обусловлена» и «объясняется» не обязательно означают причинную связь.

коэффициент корреляции не следует слишком полагаться. Например, данным на рис. 8.7 соответствует весьма высокий коэффициент корреляции ($r = 0{,}92$), однако ширина доверительной области значений показывает, что неопределенность предсказания довольно значительна. Поэтому даже при большом коэффициенте корреляции обязательно вычислите доверительную область значений.

И под конец приведем соотношение коэффициента корреляции и коэффициента наклона прямой регрессии b:

$$r = b\frac{s_X}{s_Y},$$

где b — коэффициент наклона прямой регрессии, s_X и s_Y — стандартные отклонения переменных.

Если не брать во внимание случай $s_X = 0$, то коэффициент корреляции равен нулю тогда и только тогда, когда $b = 0$. Этим фактом мы сейчас и воспользуемся для оценки статистической значимости корреляции.

Статистическая значимость корреляции

Поскольку из $b = 0$ следует $r = 0$, гипотеза об отсутствии корреляции равнозначна гипотезе о нулевом наклоне прямой регрессии. Поэтому для оценки статистической значимости корреляции можно воспользоваться уже известной нам формулой для оценки статистической значимости отличия b от нуля:

$$t = \frac{b}{s_b}.$$

Здесь число степеней свободы $\nu = n - 2$.

Однако если коэффициент корреляции уже вычислен, удобнее воспользоваться формулой:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}.$$

Число степеней свободы здесь также $\nu = n - 2$.

При внешнем несходстве двух формул для t, они тождественны. Действительно, из того, что

$$r^2 = 1 - \frac{n-2}{n-1}\frac{s^2_{y|x}}{s^2_y},$$

следует

$$s^2_{y|x} = \frac{n-1}{n-2}(1-r^2)s^2_Y.$$

Подставив значение $s_{y|x}$ в формулу для стандартной ошибки

$$s_b = \frac{1}{\sqrt{n-1}}\frac{s_{y|x}}{s_x},$$

получим

$$s_b = \frac{s_Y}{s_X}\sqrt{\frac{1-r^2}{n-2}}.$$

С другой стороны, поскольку

$$r = b\frac{s_X}{s_Y},$$

имеем

$$b = r\frac{s_Y}{s_X}.$$

Теперь подставим выражения для s_b и b в формулу

$$t = b/s_b.$$

Получим:

$$t = \frac{r\dfrac{s_Y}{s_X}}{\dfrac{s_Y}{s_X}\sqrt{\dfrac{1-r^2}{n-2}}} = \frac{r}{\sqrt{\dfrac{1-r^2}{n-2}}}.$$

Животный жир и рак молочной железы

В опытах на лабораторных животных показано, что высокое содержание животного жира в рационе повышает риск рака молочной железы. Наблюдается ли эта зависимость у людей? К. Кэррол* собрал данные о потреблении животных жиров и смертности от рака молочной железы по 39 странам. Результат представлен на рис. 8.12А. Коэффициент корреляции между потреблением животных жиров и смертностью от рака молочной железы оказался равен 0,90. Оценим статистическую значимость корреляции.

$$t = \frac{0{,}90}{\sqrt{\frac{1-0{,}90^2}{39-2}}} = 12{,}56.$$

Критическое значение $t_{0.001}$ при числе степеней свободы $\nu = 39-2 = 37$ равно 3,574, то есть меньше полученного нами. Таким образом, при уровне значимости 0,001 можно утверждать, что существует корреляция между потреблением животных жиров и смертностью от рака молочной железы.

Теперь проверим, связана ли смертность с потреблением растительных жиров? Соответствующие данные приведены на рис. 8.12Б. Коэффициент корреляции равен 0,15. Тогда

$$t = \frac{0{,}15}{\sqrt{\frac{1-0{,}15^2}{39-2}}} = 0{,}92.$$

Даже при уровне значимости 0,10 вычисленное значение t меньше критического. Корреляция статистически не значима.

Таким образом, риск рака молочной железы статистически значимо связан с потреблением животных, но не растительных жиров. Значит ли это, что животный жир способствует развитию рака молочной железы? Пока нет. Ведь обе рассматриваемые переменные могут зависеть от какой-то третьей. В обсервацион-

* K. K. Carroll. Experimental evidence of dietary factors and hormone-dependent cancers. *Cancer Res.*, 35:3375—3383, 1975.

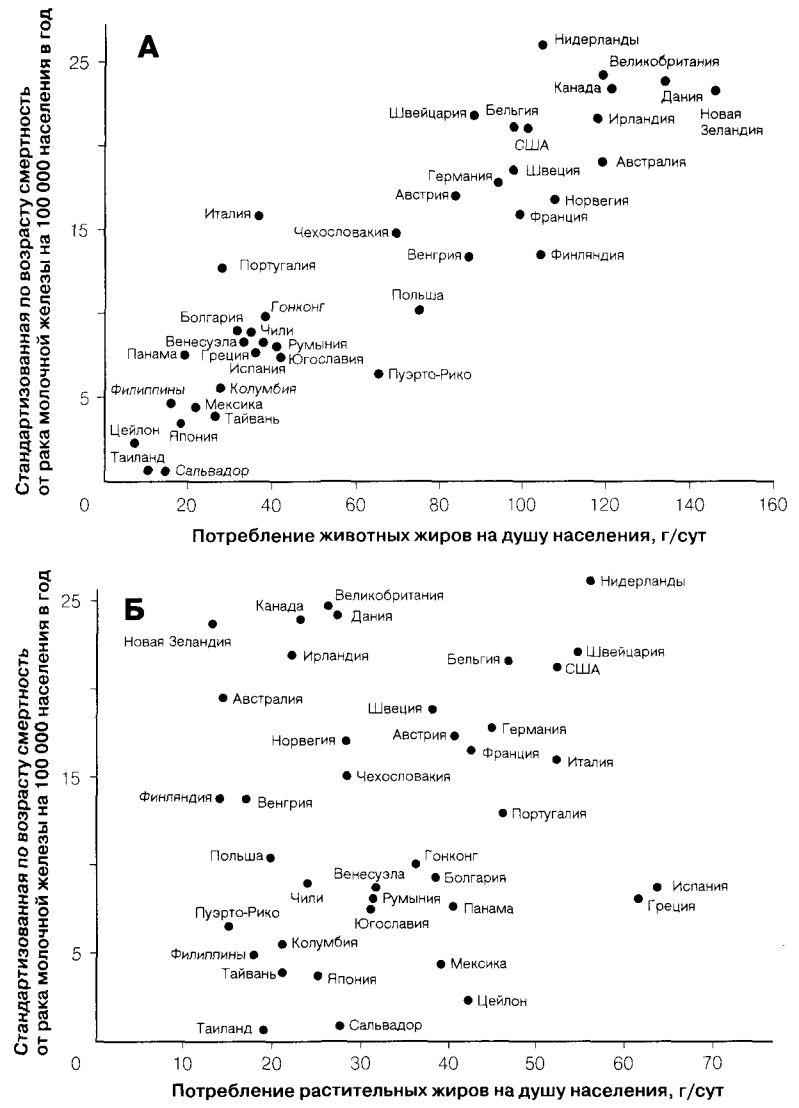


Рис. 8.12. Смертность от рака молочной железы и потребление жиров на душу населения в разных странах. **А.** Потребление животных жиров. **Б.** Потребление растительных жиров. Связь смертности с потреблением животных жиров достаточно отчетлива, чего не скажешь о связи с потреблением растительных жиров.

ном исследовании, каковым является работа Кэррола, такую возможность отвергнуть нельзя*. Однако экспериментальные данные, о которых мы упомянули выше, — сильный аргумент в пользу именно причинно-следственной связи.

Вообще истолкование результатов регрессионного и корреляционного анализа зависит от того, в каком исследовании были получены данные — обсервационном или экспериментальном. Если мы обнаружили связь переменных в обсервационном исследовании, то это не значит, что одна из них *влияет* на другую. Возможно, их согласованные изменения — результат действия какого-то неизвестного нам фактора. В экспериментальном исследовании, произвольно меняя одну из переменных, мы можем быть уверены, что связь, если она будет выявлена, является причинной. Впрочем, осторожность не помешает и в этом случае. В самом деле, трудно менять *только одну* переменную. Увеличивая содержание жира в рационе, мы либо увеличиваем общую калорийность, либо снижаем содержание белков и углеводов. Кто поручится, что канцерогенное действие оказывает именно жир, а не дисбаланс питательных веществ?

КОЭФФИЦИЕНТ РАНГОВОЙ КОРРЕЛЯЦИИ СПИРМЕНА

Расчет коэффициента корреляции возможен при тех же условиях, что и регрессионный анализ. Это прежде всего линейность связи переменных и нормальность распределения. Эти условия выполняются далеко не всегда. Кроме того, в клинических исследованиях мы часто имеем дело с порядковыми признаками, а к ним ни регрессионный анализ, ни расчет коэффициента кор-

---

* Например, исследования показывают, что заболеваемость раком молочной железы связана с уровнем доходов, числом автомобилей и телевизоров в семье. (B. S. Drasar, D. Irving. Environmental factors and cancer of the colon and breast. *Br. J. Cancer*, 27:167—172, 1973.) Но значит ли это, что, покупая новый автомобиль, домашняя хозяйка увеличивает риск заболеть раком молочной железы? На основании таких данных мы вправе только предположить, что какой-то фактор, связанный с уровнем жизни, влияет на риск рака молочной железы, но не можем точно указать этот фактор.

реляции, разумеется, неприменим. В подобных случаях следует воспользоваться коэффициентом ранговой корреляции Спирмена*. Это непараметрический метод — он не требует нормальности распределения; не требует он и линейной зависимости, его можно применять как к количественным, так и к порядковым признакам**.

Идея коэффициента ранговой корреляции Спирмена (его обозначают r_s) проста. Нужно упорядочить данные по возрастанию и заменить реальные значения их рангами. *Рангом* значения называется его номер в упорядоченном ряду. Например, в ряду 1, 4, 8, 8, 12 ранг числа 4 равен 2. Затем, беря вместо самих значений их ранги, рассчитывают обычный коэффициент корреляции Пирсона. Это и будет коэффициент ранговой корреляции Спирмена. Его можно рассчитать и проще:

$$r_s = 1 - \frac{6\Sigma d^2}{n^3 - n},$$

где d — разность рангов для каждого члена выборки.

Как быть, если в ряду встретятся одинаковые значения? Скажем, в приведенном примере это две восьмерки. Им следует

---

* Упомянем также коэффициент ранговой корреляции *Кендалла*, обозначаемый τ. В отличие от коэффициента ранговой корреляции Спирмена он может быть обобщен для случая нескольких независимых переменных. Заключения, основанные на использовании обоих коэффициентов, одинаковы, хотя числовые значения коэффициентов не совпадают. О коэффициенте ранговой корреляции Кендалла можно прочесть в книге: S. Siegel, N. J. Castellar. Nonparametric statistics for the behavioral sciences (2d ed.). McGraw-Hill, New York, 1988.

** Если параметрические методы, требующие нормального распределения, применить к данным с иным типом рапределения, это приведет к ошибочному заключению. Напротив, непараметрические методы можно смело применять и в случае нормального распределения. Однако тогда чувствительность их будет несколько ниже чувствительности параметрических методов. Что касается коэффициента ранговой корреляции Спирмена, то он и в этом случае проигрывает коэффициенту корреляции Пирсона весьма незначительно.

Таблица 8.5. Вычисление коэффициента ранговой корреляции Спирмена

Рост		Вес		
Значение, см	Ранг	Значение, г	Ранг	Разность рангов
31	1	7,7	2	−1
32	2	8,3	3	−1
33	3	7,6	1	2
34	4	9,1	4	0
35	5,5	9,6	5	0,5
35	5,5	9,9	6	−0,5
40	7	11,8	7	0
41	8	12,2	8	0
42	9	14,8	9	0
46	10	15,0	10	0

присвоить один и тот же ранг, равный среднему занимаемых ими мест: $(3+4)/2 = 3{,}5$. Рангом стоящего за ними числа 12 будет 5.

Посмотрим, как вычислить r_s для знакомой нам выборки из 10 марсиан (табл. 8.5). Вначале упорядочим по возрастанию значения каждой из переменных. Ранг 1 присваивается меньшему значению, 10 — большему. Упорядочим марсиан по росту. На 5-м и 6-м месте в нем стоят одинаковые значения. Присвоим им общий ранг $(5+6)/2 = 5{,}5$. Затем упорядочим марсиан по весу и для каждого марсианина вычислим разность рангов роста и веса.

Наконец, вычислим коэффициент ранговой корреляции Спирмена:

$$r_s = 1 - \frac{6[(-1)^2 + (-1)^2 + 2^2 + 0^2 + 0{,}5^2 + (-0{,}5)^2 + 0^2 + 0^2 + 0^2]}{10^3 - 10} =$$

$$= 0{,}96.$$

Обратимся к таблице 8.6, где приведены критические значения коэффициента ранговой корреляции Спирмена для разных уровней значимости и объемов выборки. Критическое значение для уровня значимости 0,001 и объема выборки $n = 10$ равно 0,903, что меньше полученного нами. Тем самым, корреляция статистически значима ($P < 0{,}001$).

Таблица 8.6. Критические значения коэффициента ранговой корреляции Спирмена

	Уровень значимости α								
n	0,50	0,20	0,10	0,05	0,02	0,01	0,005	0,002	0,001
4	0,600	1,000	1,000						
5	0,500	0,800	0,900	1,000	1,000				
6	0,371	0,657	0,829	0,886	0,943	1,000	1,000		
7	0,321	0,571	0,714	0,786	0,893	0,929	0,964	1,000	1,000
8	0,310	0,524	0,643	0,738	0,833	0,881	0,905	0,952	0,976
9	0,267	0,483	0,600	0,700	0,783	0,833	0,867	0,917	0,933
10	0,248	0,455	0,564	0,648	0,745	0,794	0,830	0,879	0,903
11	0,236	0,427	0,536	0,618	0,709	0,755	0,800	0,845	0,873
12	0,217	0,406	0,503	0,587	0,678	0,727	0,769	0,818	0,846
13	0,209	0,385	0,484	0,560	0,648	0,703	0,747	0,791	0,824
14	0,200	0,367	0,464	0,538	0,626	0,679	0,723	0,771	0,802
15	0,189	0,354	0,446	0,521	0,604	0,654	0,700	0,750	0,779
16	0,182	0,341	0,429	0,503	0,582	0,635	0,679	0,729	0,762
17	0,176	0,328	0,414	0,485	0,566	0,615	0,662	0,713	0,748
18	0,170	0,317	0,401	0,472	0,550	0,600	0,643	0,695	0,728
19	0,165	0,309	0,391	0,460	0,535	0,584	0,628	0,677	0,712
20	0,161	0,299	0,380	0,447	0,520	0,570	0,612	0,662	0,696
21	0,156	0,292	0,370	0,435	0,508	0,556	0,599	0,648	0,681
22	0,152	0,284	0,361	0,425	0,496	0,544	0,586	0,634	0,667
23	0,148	0,278	0,353	0,415	0,486	0,532	0,573	0,622	0,654
24	0,144	0,271	0,344	0,406	0,476	0,521	0,562	0,610	0,642
25	0,142	0,265	0,337	0,398	0,466	0,511	0,551	0,598	0,630
26	0,138	0,259	0,331	0,390	0,457	0,501	0,541	0,587	0,619
27	0,136	0,255	0,324	0,382	0,448	0,491	0,531	0,577	0,608
28	0,133	0,250	0,317	0,375	0,440	0,483	0,522	0,567	0,598
29	0,130	0,245	0,312	0,368	0,433	0,475	0,513	0,558	0,589
30	0,128	0,240	0,306	0,362	0,425	0,467	0,504	0,549	0,580
31	0,126	0,236	0,301	0,356	0,418	0,459	0,496	0,541	0,571
32	0,124	0,232	0,296	0,350	0,412	0,452	0,489	0,533	0,563
33	0,121	0,229	0,291	0,345	0,405	0,446	0,482	0,525	0,554
34	0,120	0,225	0,287	0,340	0,399	0,439	0,475	0,517	0,547
35	0,118	0,222	0,283	0,335	0,394	0,433	0,468	0,510	0,539
36	0,116	0,219	0,279	0,330	0,388	0,427	0,462	0,504	0,533
37	0,114	0,216	0,275	0,325	0,383	0,421	0,456	0,497	0,526
38	0,113	0,212	0,271	0,321	0,378	0,415	0,450	0,491	0,519
39	0,111	0,210	0,267	0,317	0,373	0,410	0,444	0,485	0,513
40	0,110	0,207	0,264	0,313	0,368	0,405	0,439	0,479	0,507

Таблица 8.6. Окончание

	Уровень значимости α								
n	0,50	0,20	0,10	0,05	0,02	0,01	0,005	0,002	0,001
41	0,108	0,204	0,261	0,309	0,364	0,400	0,433	0,473	0,501
42	0,107	0,202	0,257	0,305	0,359	0,395	0,428	0,468	0,495
43	0,105	0,199	0,254	0,301	0,355	0,391	0,423	0,463	0,490
44	0,104	0,197	0,251	0,298	0,351	0,386	0,419	0,458	0,484
45	0,103	0,194	0,248	0,294	0,347	0,382	0,414	0,453	0,479
46	0,102	0,192	0,246	0,291	0,343	0,378	0,410	0,448	0,474
47	0,101	0,190	0,243	0,288	0,340	0,374	0,405	0,443	0,469
48	0,100	0,188	0,240	0,285	0,336	0,370	0,401	0,439	0,465
49	0,098	0,186	0,238	0,282	0,333	0,366	0,397	0,434	0,460
50	0,097	0,184	0,235	0,279	0,329	0,363	0,393	0,430	0,456

Если объем выборки больше 50, нужно применить критерий Стьюдента:

$$t = \frac{r_s}{\sqrt{\frac{1-r_s^2}{n-2}}}$$

с числом степеней свободы $\nu = n-2$.

В данном случае связь веса и роста можно было установить и без помощи коэффициента ранговой корреляции Спирмена. Применение обычного коэффициента корреляции, как мы видели, приводит к тем же результатам.

Сколько лабораторных анализов нужно врачу?

В первые дни пребывания в больнице больному обычно делают множество дорогостоящих анализов. Все ли из них необходимы? Шредер с коллегами* попытались, анализируя работу 21 врача, выяснить, существует ли связь между квалификацией врача и стоимостью необходимых ему анализов. Прежде всего, специальная комиссия оценила квалификацию каждого врача. Каждому из врачей присвоили ранг от 1 (лучшая квалификация) до

* S. A. Schroeder, A. Schliftman, T. E. Piemine. Variation among physicians in use of laboratory tests: relation to quality of care. *Med. Care*, 12: 709—713, 1974.

21 (худшая квалификация). Затем была подсчитана средняя стоимость анализов, которые потребовались каждому из врачей за первые 3 суток пребывания больного в клинике. Эти данные упорядочили по возрастанию; наименьшей стоимости присвоили ранг 1, наибольшей — 21.

В результате каждому врачу была присвоена пара рангов — ранг по шкале квалификации и ранг по шкале расходов. Эти пары представлены на рис. 8.13. Остается выяснить связь между квалификацией врача и величиной расходов на необходимые ему анализы. Вычислив коэффициент Спирмена, получим всего лишь $r_s = -0{,}13$. Абсолютная величина r_s оказалась меньше критического значения даже при уровне значимости $\alpha = 0{,}05$ (критическое значение $r_{0.05} = 0{,}435$).

Однако значит ли это, что не существует связи между квалификацией врача и затратами на анализы? Нет. Связь существует, но она не линейная. Присмотревшись к рис. 8.13, можно заметить, что самыми дешевыми анализы были у лучших и... худших врачей. И тем и другим, чтобы уверенно судить о болезни, не требуется много анализов. Причем, похоже, большей уверенностью отличаются именно худшие специалисты.

Но почему эта связь не была уловлена коэффициентом корреляции? Исключительно из-за ее нелинейной формы. Ни один из коэффициентов корреляции не сможет уловить зависимость, график которой — перевернутая U-образная кривая с рис. 8.13.

Этот пример показывает, что, прежде чем применять какие-либо методы анализа связей, следует примерно определить, какой может быть форма зависимости. Лучший способ для этого — просто нарисовать график, подобный изображенному на рис. 8.13.

ЧУВСТВИТЕЛЬНОСТЬ КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ

Как уже говорилось, из статистической значимости коэффициента корреляции вытекает статистическая значимость коэффициента наклона. Ограничимся поэтому вычислением чувствительности коэффициента корреляции.

Можно показать, что величина

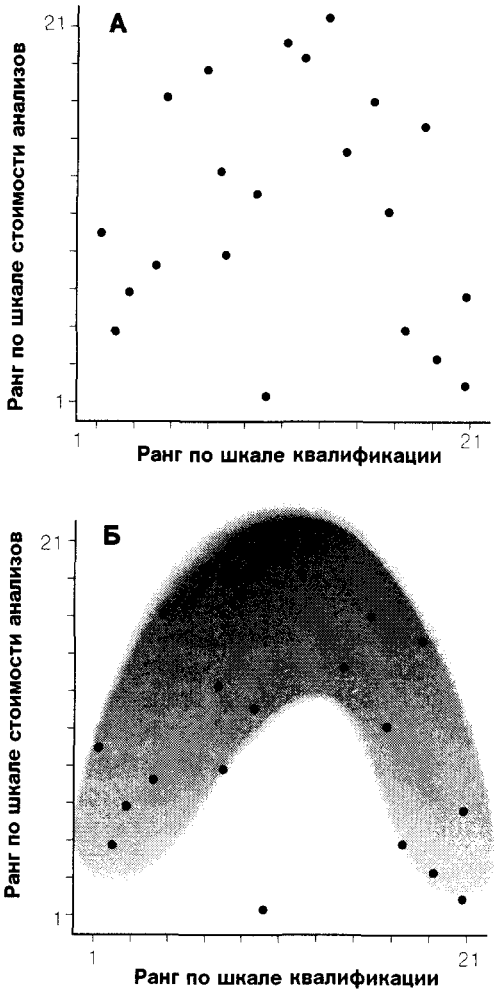


Рис. 8.13. А. Квалификация врача и стоимость анализов, которые он назначает больному в первые 3 дня госпитализации. Коэффициент ранговой корреляции Спирмена — всего лишь –0,13. Можно было бы заключить, что стоимость анализов от квалификации никак не зависит. **Б.** Приглядевшись к данным повнимательнее, можно заметить, что зависимость на самом деле есть, только не линейная, а похожая на перевернутую букву U. Расходы на анализы выше у врачей средней квалификации, у наиболее и наименее квалифицированных врачей расходы ниже.

$$Z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

имеет нормальное распределение со стандартным отклонением

$$\sigma_Z = \sqrt{\frac{1}{n-3}}.$$

Тогда величина

$$z = \frac{Z}{\sigma_Z}$$

в отсутствие корреляции имеет стандартное нормальное распределение со средним, равным нулю. Обозначим истинное значение коэффициента корреляции ρ (греческая «ро»). Тогда средним значением z будет Z_ρ/σ_Z, где

$$Z_\rho = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right).$$

Найдем, какой должна быть чувствительность, чтобы по выборке объемом 10 при уровне значимости 0,05 обнаружить корреляцию ρ, не меньшую 0,9. На рис. 8.14 приведены два распределения z — для нулевого коэффициента корреляции и истинного, равного ρ. (Заметьте, насколько этот этот рисунок похож на рис. 6.7.) Чувствительность равна площади под истинной кривой распределения z справа от критического значения z_α.

Вычислим

$$Z_\rho = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) = \frac{1}{2}\ln\left(\frac{1+0{,}9}{1-0{,}9}\right) = 1{,}472$$

и

$$\sigma_Z = \sqrt{\frac{1}{n-3}} = 0{,}378.$$

Уровню значимости $\alpha = 0{,}05$ соответствует критическое значение $z_\alpha = 1{,}960$. Центром распределения z является $Z_\rho/\sigma_Z = 1{,}472/0{,}378 = 3{,}894$. От этого центра критическое значение z_α от-

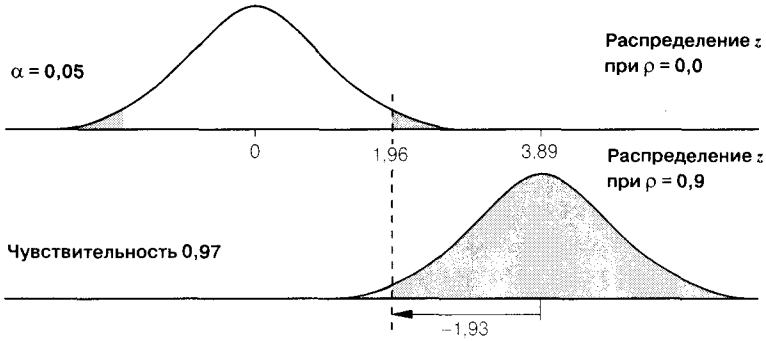


Рис. 8.14. Чувствительность выявления корреляции ρ = 0,9 при объеме выборки n = 10 и уровне значимости α = 0,05.

стоит на 1,960 – 3,894 = –1,934 стандартных отклонения. Из табл. 6.4 находим, что площадь части стандартного нормального распределения, расположенной правее–1,934 стандартного отклонения от центра, составляет примерно 0,97. То есть искомая чувствительность равна 97%.

Итак, чувствительность 1–β, необходимая для обнаружения корреляции, не меньшей ρ, при уровне значимости α и при объеме выборки n равна площади под кривой стандартного нормального распределения правее точки

$$z_{1-\beta} = z_{\alpha} - \frac{Z_{\rho}}{\sqrt{\frac{1}{n-3}}}.$$

Эта формула для нахождения чувствительности по известному объему выборки. Если нужно найти объем выборки, при котором достигалась бы чувствительность 1–β, то, разрешив это уравнение относительно n, получим:

$$n = \left(\frac{z_{\alpha} - z_{1-\beta}}{Z_{\rho}} \right)^2 + 3.$$

СРАВНЕНИЕ ДВУХ СПОСОБОВ ИЗМЕРЕНИЯ: МЕТОД БЛЭНДА—АЛТМАНА

Нередко требуется сравнить результаты измерений, выполненных двумя методами, ни один из которых не является абсолютно надежным. Например, некий гемодинамический показатель определяли непрямым, неинвазивным, методом. Допустим, изобретен новый метод, также непрямой. Естественно выяснить, согласуются ли результаты измерений, выполненных старым и новым методами. Или похожий вопрос — насколько согласованы результаты повторных измерений, выполненных одним и тем же методом.

Итак, с помощью двух методов получены две серии измерений. Казалось бы, ничто не мешает применить регрессионный анализ или рассчитать коэффициент корреляции. Увы, эти, на первый взгляд, очевидные действия могут привести к ложным выводами.

Регрессионный анализ неприменим уже потому, что его результаты зависят от того, какую переменную считать независимой, а какую зависимой. Тут следует подчеркнуть отличие задачи сравнения двух методов измерения от задачи *калибровки*, в которой приближенные измерения сравниваются с некоторым эталоном. Типичный пример калибровки: приготовив ряд растворов известной концентрации, измерить ее исследуемым методом. Здесь регрессионный анализ вполне применим, поскольку эталон – достоверно известная концентрация — очевидным образом и является независимой переменной. Напротив, при сравнении результатов двух приближенных методов никакого эталона нет.

Что может дать коэффициент корреляции? Положим, он статистически значимо отличается от нуля. Но ценен ли этот факт? Нет, ведь проверялась корреляция измерений *одной и той же* величины. В этом случае удивления было бы достойно как раз отсутствие значимой корреляции, говорящее о том, что результаты, как минимум, одного из методов нимало не схожи с истинными значениями измеряемого признака. Это практически исключено. Кроме того, как мы видели, даже весьма высоким ко-

эффициентам корреляции соответствует довольно значительная неопределенность предсказания зависимой переменной.

Д. Блэнд и Дж. Алтман предложили описательный метод оценки согласованности измерений, выполненных двумя способами*. Идея метода очень проста. Для каждой — выполненной одним и другим способами — пары измерений вычислим их разность. Найдем среднюю величину и стандартное отклонение разности. Средняя разность характеризует *систематическое расхождение,* а стандартное отклонение — степень разброса результатов. Далее, если в качестве оценки измеряемого признака взять среднее значение пары измерений, то можно определить, зависит ли расхождение от величины признака. Последнее станет понятнее после того, как мы разберем пример применения метода Блэнда—Алтмана.

Два способа оценки митральной регургитации

Вспомним схему кровообращения. Из правого желудочка кровь поступает в легкие, где насыщается кислородом. Из легких кровь попадает в левое предсердие, затем — в левый желудочек. Отсюда кровь перекачивается по всему телу, снабжая органы кислородом, после чего попадает в правое предсердие и вновь в правый желудочек. Митральный клапан, расположенный между левым предсердием и левым желудочком, при сокращении желудочка закрывается и преграждает крови путь обратно в предсердие. При митральной недостаточности возникает так называемая *митральная регургитация*: часть крови при сокращении левого желудочка выбрасывается в предсердие. В результате легкие переполняются кровью, что затрудняет их работу. Если митральная регургитация слишком велика, клапан необходимо заменять искусственным, — вот почему ее количественная оценка чрезвычайно важна. Такой оценкой служит *фракция регургитации* — доля крови, которая при каждом сокращении выбрасы-

* Более подробное изложение этой процедуры можно найти в статьях: D. G. Altman and J. M. Bland. Measurement in medicine: the analysis of method comparison studies. *Statistician*, 32:307—317, 1983 и J. M. Bland and D. G. Altman. Statistical methods for assessing agreement between two measures of clinical measurement. *Lancet,* 1(8476):307—310, 1986.

Таблица 8.7. Фракция митральной регургитации по данным катетеризации сердца и допплеровского исследования

Допплеровское исследование	Катетеризация	Разность	Среднее значение
0,49	0,62	–0,13	0,56
0,83	0,72	0,11	0,78
0,71	0,63	0,08	0,67
0,38	0,61	–0,23	0,50
0,57	0,49	0,08	0,53
0,68	0,79	–0,11	0,74
0,69	0,72	–0,03	0,71
0,07	0,11	–0,04	0,09
0,75	0,66	0,09	0,71
0,52	0,74	–0,22	0,63
0,78	0,83	–0,05	0,81
0,71	0,66	0,05	0,69
0,16	0,34	0,18	0,25
0,33	0,50	–0,17	0,42
0,57	0,62	–0,05	0,60
0,11	0,00	0,11	0,06
0,43	0,45	–0,02	0,44
0,11	0,06	0,05	0,85
0,31	0,46	–0,15	0,39
0,20	0,03	0,17	0,12
0,47	0,50	–0,03	0,49

вается из левого желудочка в левое предсердие. В норме фракция регургитации равна нулю; чем тяжелее митральная недостаточность, тем более фракция регургитации приближается к единице.

Фракцию регургитации можно определить с помощью катетеризации сердца. В левый желудочек вводят катетер, а через него — рентгеноконтрастный препарат. Наблюдая за его распространением, можно определить, какая доля крови выбрасывается в левое предсердие. Описанный способ трудно назвать приятным, дешевым и безопасным.

Э. Мак-Исаак с соавт. предложили определять фракцию ре-

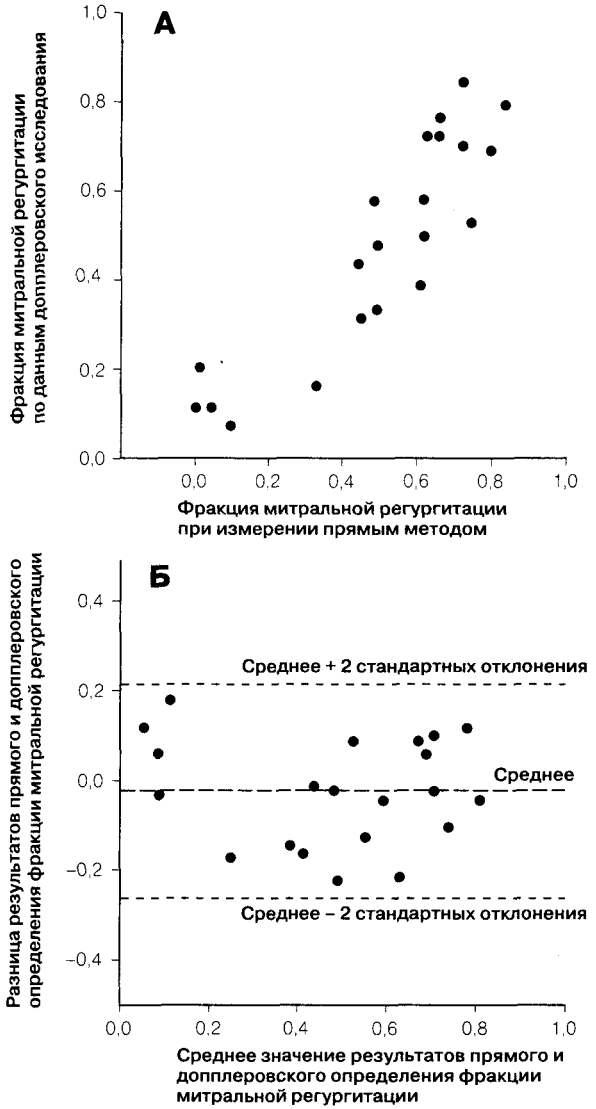


Рис. 8.15. **А.** Фракция митральной регургитации при измерении прямым методом и по данным допплеровского исследования. **Б.** Сравнение результатов по методу Блэнда—Алтмана.

гургитации с помощью допплеровского исследования*. Этот способ значительно проще и вполне безопасен. Насколько согласуются оценки, полученные двумя способами? Фракцию регургитации обоими способами определили у 21 человека. Результаты приведены на рис. 8.15А и в табл. 8.7. Коэффициент корреляции между измерениями, выполненными обоими способами, составил 0,89. Высокое значение коэффициента корреляции говорит о тесной линейной связи, однако для оценки согласованности этого недостаточно.

Помимо самих измерений в табл. 8.7 приведены усредненные по каждому больному значения фракции регургитации и разности этих долей. На рис. 8.15Б изображены разности долей для каждого усредненного значения. Такое представление позволяет сделать ряд выводов. Во-первых, средняя разность между измерениями равна всего лишь −0,03, что говорит об отсутствии систематического расхождения. Во-вторых, стандартное отклонение разностей составило 0,12, что невелико по сравнению с самими значениями. В-третьих, отсутствует зависимость разности измерений от величины фракции регургитации. Таким образом, измерения, полученные обоими способами, хорошо согласуются друг с другом.

ЗАКЛЮЧЕНИЕ

Мы рассмотрели методы, предназначенные для оценки связи между двумя признаками. Успех применения этих методов определяется тем, насколько математическая модель, лежащая в их основе, соответствует действительности. Особенно важна форма зависимости — она должна быть линейной. Поэтому, перед тем как приступить к расчетам, нанесите данные на график — это поможет вам правильно выбрать статистический метод (или отказаться от применения любого из них).

---

* A. I. MacIsaac, I. G. McDonald, R. L. G. Kirsner, S. A. Graham, R. W. Gill. Quantification of mitral regurgitation by integrated Doppler backscatter power. *J. Am. Coll. Cardiol.*, 24:690–695, 1994.

ЗАДАЧИ

8.1. Постройте графики для приведенных наборов данных. Найдите для линии регрессии и коэффициенты корреляции.

X	Y
30	37
30	47
40	50
40	60

X	Y
30	37
30	47
40	50
40	60
20	25
20	35
50	62
50	72

X	Y
30	37
30	47
40	50
40	60
20	25
20	35
50	62
50	72
10	13
10	23
60	74
60	84

Нанесите данные и прямые регрессии на графики. Что в этих трех случаях общего, в чем различия?

8.2. Постройте графики для двух наборов данных. Найдите для каждого линию регрессии и коэффициент корреляции.

X	Y
15	19
15	29
20	25
20	35
25	31
25	41
30	37
30	47
60	40

X	Y
20	21
20	31
30	18
30	28
40	15
40	25
40	75
40	85
50	65
50	75
60	55
60	65

Нанесите полученные прямые регрессии на графики с исходными данными. Обсудите результаты.

8.3. На рис. 8.16 и в таблице под ним представлены результаты четырех экспериментов. Вычислите для каждого эксперимента коэффициенты линейной регрессии и коэффициент корреляции. В чем сходство и различие результатов экспериментов? Проверьте, выполняются ли условия применимости регрессионного анализа.

8.4. Исследуя проницаемость сосудов сетчатки, Дж. Фишман и соавт. (G. A. Fishman et al. Blood-retinal barrier function in patients with cone or cone-rod dystrophy. *Arch. Ophthalmol.*, 104:545—548, 1986) решили выяснить, связан ли этот показатель с электрической активностью сетчатки. Позволяют ли полученные данные говорить о существовании связи?

Проницаемость сосудов сетчатки	Электрическая активность сетчатки
19,5	0,0
15,0	38,5
13,5	59,0
23,3	97,4
6,3	119,2
2,5	129,5
13,0	198,7
1,8	248,7
6,5	318,0
1,8	438,5

8.5. Наиболее точную оценку объема левого желудочка дает рентгеноконтрастная вентрикулография — метод, требующий катетеризации сердца, а потому дорогой и небезопасный. Продолжается поиск методов, не требующих катетеризации. Р. Слуцкий* и соавт. (R. Slutsky et al. Left ventricular volumes by gated equilibrium

* Роберт Слуцкий был обвинен в подтасовке данных, и ряд его работ объявлен фальсификацией. Принадлежит ли цитируемая статья к их числу, мне неизвестно. Как бы то ни было, мы рассматриваем данные исключительно в учебных целях. Интересующиеся судьбой работ Слуцкого могут обратиться в Калифорнийский университет в Сан-Диего.

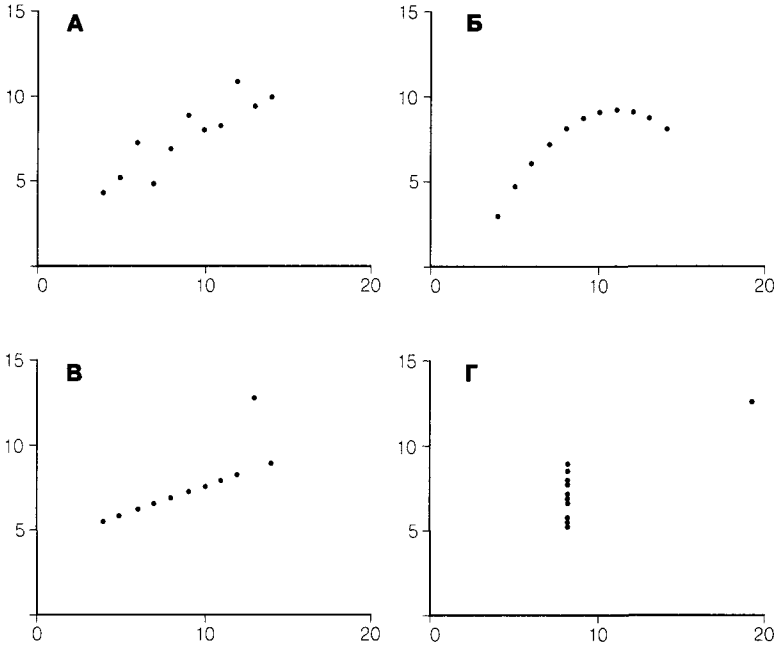


Рис. 8.16. К задаче 8.3.

Эксперимент А		Эксперимент Б		Эксперимент В		Эксперимент Г	
X	*Y*	*X*	*Y*	*X*	*Y*	*X*	*Y*
10	8,04	10	9,14	10	7,46	8	6,58
8	6,95	8	8,14	8	6,77	8	5,76
13	7,58	13	8,74	13	12,74	8	7,71
9	8,81	9	8,77	9	7,11	8	8,84
11	8,33	11	9,26	11	7,81	8	8,47
14	9,96	14	8,10	14	8,84	8	7,04
6	7,24	6	6,13	6	6,08	8	5,25
4	4,26	4	3,10	4	5,39	19	12,50
12	10,84	12	9,13	12	8,15	8	5,56
7	4,82	7	7,26	7	6,42	8	7,91
5	5,68	5	4,74	5	5,73	8	6,89

radionuclide angiography method. *Circulation*, 60:556— 564, 1979) исследовали метод оценки объема левого желудочка по данным изотопной вентрикулографии с внутривенным введением изотопа.

Конечно-диастолический объем		Конечно-систолический объем	
Изотопная вентрикулография	Рентгеноконтрастная вентрикулография	Изотопная вентрикулография	Рентгеноконтрастная вентрикулография
75	101	35	47
48	75	30	35
126	126	52	49
93	106	23	23
201	195	103	88
260	265	182	173
40	60	14	12
293	288	166	163
95	94	27	29
58	67	24	25
91	81	50	25
182	168	139	131
91	89	50	49
88	102	40	44
161	150	57	60
118	94	41	18
120	129	48	40

Хорошо ли согласуются результаты?

8.6. Азотистый баланс — разность между количеством азота, который попадает в организм с пищей, и количеством азота, выводимого из организма, — важный показатель полноценности питания. Отрицательный азотистый баланс свидетельствует о том, что организм не получает достаточно белка. Нормы суточного потребления белка, рекомендуемые Всемирной организацией здравоохранения и Японским комитетом питания, рассчитаны главным образом на мужчин. Целью исследования К. Канеко и Г. Койке (K. Kaneko, G. Koike. Utilization and requirement of egg protein in Japanese women. *J. Nutr. Sci. Vitaminol. (Tokyo)*, 31:43—52, 1985) было определить количество белка в

рационе, необходимое для поддержания нулевого азотистого баланса у японских женщин. Связь суточного потребления азота и азотистого баланса определили при калорийности суточного рациона 37 и 33 ккал/кг. Были получены следующие данные.

Калорийность суточного рациона			
37 ккал/кг		33 ккал/кг	
Потребление азота, мг/кг	Азотистый баланс, мг/кг	Потребление азота, мг/кг	Азотистый баланс, мг/кг
49	−30	32	−32
47	−22	32	−20
50	−29	32	−17
76	−22	51	−10
77	−15	53	−20
99	−10	51	−18
98	−11	52	−21
103	−10	74	4
118	−1	72	−16
105	−4	74	−14
100	−13	98	6
98	−14	97	−7

Найдите уравнения регрессии для обеих групп. Изобразите на одном рисунке результаты наблюдений и линии регрессии. Является ли различие между линиями регрессии статистически значимым? Для группы 37 ккал/кг найдите величину потребления азота, обеспечивающую нулевой азотистый баланс.

8.7. В. Ернайчик (W. Jernajczyk. Latency of eye movement and other REM sleep parameters in bipolar depression. *Biol. Psychiatry*, 21:465—472, 1986), изучая физиологию сна при депрессии, столкнулся с необходимостью оценки тяжести этого заболевания. Шкала депрессии Бека основана на опроснике, заполняемом самим больным. Она проста в применении, однако специфичность ее недостаточна. Применение шкалы депрессии Гамильтона более сложно, поскольку требует участия врача, но именно эта шкала дает наиболее точные результаты. Тем не менее автор был склонен использовать шкалу Бека. В самом деле, если ее специфичность недостаточна для диагностики, то это еще не

говорит о том, что ее нельзя использовать для оценки тяжести депрессии у больных с уже установленным диагнозом. Сравнив оценки по обеим шкалам у 10 больных, В. Ернайчик получил следующие результаты.

Номер больного	Оценка по шкале депрессии Бека	Оценка по шкале депрессии Гамильтона
1	20	22
2	11	14
3	13	10
4	22	17
5	37	31
6	27	22
7	14	12
8	20	19
9	37	29
10	20	15

Насколько согласованы оценки?

8.8. Полоскание с хлоргексидином предотвращает образование зубного налета, но имеет вкус, который трудно назвать приятным, кроме того, оно окрашивает зубы. Полоскание на основе хлорида аммония приятнее на вкус, не окрашивает зубы; считается, однако, что оно менее эффективно. Ф. Эшли и соавт. (F. P. Ashley et al. Effect of a 0,1% cetylpyridinium chloride mouthrinse on the accumulation and biochemical composition of dental plaque in young adults. *Caries Res.*, 18:465—471, 1984) сравнили эффективность двух видов полоскания. Участники исследования полоскали рот одним из растворов, после чего зубной налет отделяли и взвешивали. Опыт проводился 48 часов: за меньший срок налет не успевал накопиться в количестве, достаточном для точного взвешивания. Исследователей больше интересовало образование налета за 24 часа, поэтому в середине опыта налет оценивали визуально по специально разработанной шкале. Чтобы оценить точность визуальных оценок, их проводили и на 48-м часу и сравнивали с результатами взвешивания. Результаты

сопоставления двух способов оценки зубного налета представлены в таблице.

Визуальная оценка зубного налета, баллы	Сухой вес зубного налета, мг
25	2,7
32	1,2
45	2,7
60	2,1
60	3,5
65	2,8
68	3,7
78	8,9
80	5,8
83	4,0
100	5,1
110	5,1
120	4,8
125	5,8
140	11,7
143	8,5
143	11,1
145	7,1
148	14,2
153	12,2

Насколько, судя по этим данным, можно полагаться на визуальный способ оценки?

8.9. Нормальный эритроцит легко меняет форму и проходит через мельчайшие сосуды. При генетическом дефекте β-цепи гемоглобин полимеризуется, в результате форма части эритроцитов меняется, они становятся ригидными, закупоривают сосуды и разрушаются. Такова в сильно упрощенном виде сущность серповидноклеточной анемии — тяжелого заболевания с многообразными проявлениями. Наиболее мучительны болевые кризы. Они развиваются, когда под влиянием гипоксии происходит массовая полимеризация гемоглобина, деформация эритроцитов и закупорка сосудов. Существует предположение,

что дело не только в деформации и ригидности эритроцитов — определенную роль играет также повышенная склонность эритроцитов к адгезии — прилипанию к эндотелию (внутренней выстилке сосудов). Р. Хебелл и соавт. (R. Hebbel et al. Erythrocyte adherence to endothelium in sickle-cell anemia: a possible determinant of disease severity. *N. Engl. J. Med.*, 302:992—995, 1980) решили выяснить, есть ли связь между тяжестью заболевания и адгезивностью эритроцитов. Прежде всего необходимо было разработать способы оценки этих признаков.

Для оценки тяжести серповидноклеточной анемии была построена специальная шкала.

Показатель	Число баллов
Ежегодное число болевых кризов, требующих госпитализации или применения наркотических анальгетиков	
1—5	1
6—10	2
более 10	3
Язвы на коже	2
Поражение сетчатки	1
Поражение ЦНС (судороги, инсульт)	2
Поражение костей (инфаркты, асептический некроз)	2

Баллы за отдельные признаки суммировали, таким образом каждый больной получал от 0 до 13 баллов и более (по 2 балла дается за каждый инфаркт или асептический некроз кости).

Для оценки адгезивности эритроцитов известное их количество наносили на культуру эндотелия, инкубировали и затем смывали. Подсчитав число смытых эритроцитов, определяли число прилипших. Одновременно такой же опыт делали с нормальными эритроцитами. Результат выражали в виде коэффициента адгезии: отношения числа прилипших эритроцитов больного к числу прилипших эритроцитов здорового.

Было обследовано 20 больных. У каждого оценили тяжесть заболевания и коэффициент адгезии. Подтверждают ли эти дан-

ные гипотезу о связи между адгезивностью эритроцитов и тяжестью серповидноклеточной анемии?

Тяжесть заболевания, баллы	Коэффициент адгезии
0	1,0
0	1,4
1	1,0
1	1,0
1	1,9
1	2,0
1	2,5
1	3,0
2	2,0
2	3,2
3	3,0
3	3,2
3	6,3
4	2,7
5	3,0
5	5,0
5	17,0
6	5,2
9	19,8
11	25,0

8.10. Какова вероятность выявить коэффициент корреляции не меньше 0,6 при объеме выборки 39 и уровне значимости 5%?

8.11. Каков должен быть объем выборки, чтобы с вероятностью 80% выявить коэффициент корреляции не меньше 0,6 при уровне значимости 5%.

8.12. Ожирение предрасполагает к развитию инсулинонезависимого сахарного диабета. При этом играет роль тип ожирения: наиболее опасным считается так называемое ожирение по мужскому типу, когда жир откладывается преимущественно на туловище (при ожирении по женскому типу жир откладывается главным образом на бедрах и ягодицах). Однако далеко не у всех людей с ожирением по мужскому типу развивается инсулино-

независимый сахарный диабет. Необходимо действие дополнительного фактора, предположительно генетического. Т. Эндр и соавт. (T. Endre et al. Insulin resistance is coupled to low physical fitness in normotensive men with a family history of hypertension. *J. Hypertension*, 12:81—88, 1994) исследовали связь чувствительности к инсулину (ее снижение лежит в основе инсулинонезависимого сахарного диабета) и отношения объема талии к объему бедра (показатель типа ожирения). Индекс чувствительности к инсулину рассчитывали как логарифм снижения уровня глюкозы плазмы после введения инсулина. В исследование вошло 15 мужчин, у которых не было родственников первой степени с артериальной гипертонией (1-я группа) и 15 мужчин, у которых такие родственники были (2-я группа).

1-я группа		2-я группа	
Отношение объема талии к объему бедра	Индекс чувствительности к инсулину	Отношение объема талии к объему бедра	Индекс чувствительности к инсулину
0,775	1,322	0,800	1,000
0,800	1,301	0,810	0,699
0,810	1,130	0,850	0,978
0,800	0,929	0,875	0,398
0,850	1,021	0,850	0,602
0,860	1,000	0,870	0,760
0,925	1,106	0,910	0,989
0,900	0,954	0,925	0,903
0,925	0,813	0,925	0,778
0,945	1,041	0,940	0,628
0,945	1,021	0,945	0,929
0,950	0,978	0,960	0,954
0,975	0,740	1,100	0,929
1,050	0,778	1,100	0,653
1,075	0,574	0,990	0,352

Одинакова ли связь показателей в обеих группах?

Глава 9

Анализ повторных измерений

В гл. 3—5 мы рассмотрели методы сравнения данных, полученных на нескольких группах. В типичном случае мы сравнивали группу получавших препарат с группой получавших плацебо. Об эффективности препарата судили по статистической значимости различий между этими группами. Если разброс в группах велик, эффект лечения «тонет» в нем, и мы не выявляем реально существующих различий. Существует другой подход. В нем вместо двух групп нужна одна, а сравнению подлежит состояние каждого больного до и после лечения. Методически такой подход достаточно труден — ведь нужно быть уверенным, что изменение состояния не обусловлено естественным течением болезни. Тем не менее учет изменения состояния у каждого больного в отдельности, нивелируя влияние разброса данных, значительно повышает чувствительность статистических критериев.

Выявить изменение, располагая *парами* наблюдений, позволяет *парный критерий Стьюдента*. С него мы и начнем, после чего перейдем к сравнению *более чем двух* состояний больного.

Для сравнения нескольких измерений, выполненных у каждого из больных, предназначен *дисперсионный анализ повторных измерений*. В нем разброс результатов измерений разлагается на три составляющие: разброс значений между больными, в реакциях одного и того же больного и, наконец, между методами лечения. Как обычно, рассматриваемые процедуры основаны на предположении о нормальном распределении измеряемого признака. (В гл. 10 излагаются не требующие этого ранговые методы.) И, завершая рассмотрение методов анализа повторных измерений, мы разберем *критерий Мак-Нимара*. Он позволяет выявить изменения не числовых, а качественных признаков, представленные таблицами сопряженности.

ПАРНЫЙ КРИТЕРИЙ СТЬЮДЕНТА

Раньше, чтобы оценить эффективность лечения, мы выбирали две группы. Одна проходила лечение, другая нет. Затем мы вычисляли среднее по каждой группе и определяли статистическую значимость различия этих средних. Теперь мы набираем *одну* группу, измеряем у каждого больного значение признака *до* и *после* лечения и вычисляем *изменение* признака. Затем находим среднее изменение и проверяем статистическую значимость его отличия от нуля.

Такой подход более точно улавливает различия, вызванные лечением, нежели сравнение двух независимых групп, «зашумленное» разбросом значений у разных больных.

Почему такой подход повышает чувствительность критерия, легко понять из следующего примера. На рис. 9.1А и 9.1Б представлены *одни и те же* данные. Различие в том, как они получены. Данные на рис. 9.1А получены в результате наблюдения за двумя независимыми группами: левый столбец образуют данные о суточном диурезе больных, получавших плацебо, правый — получавших препарат. Напротив, оба столбца на рис. 9.1Б относятся к *одним и тем же* больным, левый содержит данные о величине диуреза до приема препарата, правый — после приема. Отрезками соединены пары точек, относящиеся к одному больному.

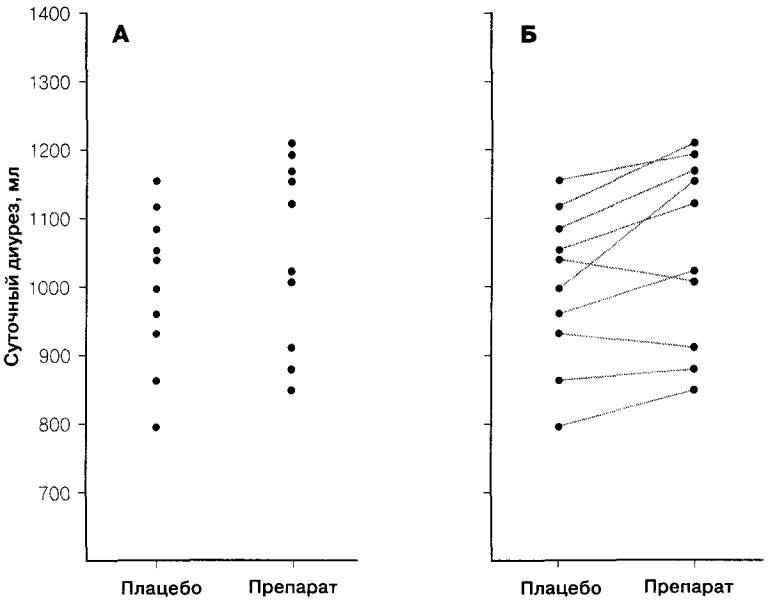


Рис. 9.1. А. Суточный диурез у 10 человек после приема плацебо и *у других* 10 человек после приема препарата (предполагаемого диуретика). На основании таких данных нельзя сделать вывод о наличии диуретического эффекта. **Б.** Суточный диурез у 10 человек после приема плацебо и *у них же* — после приема препарата. Диуретический эффект налицо. Обратите внимание, что положение точек на обоих графиках одинаково. Учет изменения диуреза у каждого обследованного в отдельности позволил выявить эффект, который был скрыт, пока мы рассматривали группы в целом.

Глядя на рис. 9.1А, никак не скажешь, что препарат оказывает диуретический эффект. Разброс данных слишком велик по сравнению со скромной тенденцией к увеличению диуреза. Вычислив критерий Стьюдента, получим $t = 1{,}33$. Это меньше $t_{0,05} = 2{,}101$ — критического значения при уровне значимости 0,05 и числе степеней свободы $\nu = 2(n-1) = 2(10-1) = 18$. Тем самым, статистически значимых различий не выявлено.

Казалось бы, результат в случае повторных измерений (рис. 9.1Б) будет таким же. Ведь положение точек на рисунках совпадает. Однако теперь мы располагаем дополнительной информацией: мы знаем, как изменился диурез у каждого больно-

го. Судя по наклону отрезков, препарат увеличил диурез у 8 из 10 больных. А это достаточно веский довод в пользу того, что препарат — диуретик.

Перейдем к количественной оценке этого впечатления. Оценить статистическую значимость изменения позволяет *парный критерий Стьюдента*. Нулевая гипотеза будет состоять в том, что среднее изменение равно нулю.

В общем случае критерий Стьюдента можно представить в таком виде:

$$t = \frac{\text{Оценка параметра} - \text{Истинное значение параметра}}{\text{Стандартная ошибка оценки параметра}}.$$

Интересующий нас параметр — истинное среднее изменение диуреза — обозначим δ. Его оценкой является наблюдаемое (выборочное) среднее изменение диуреза $\bar{d}$. Выборочное стандартное отклонение изменения диуреза составляет

$$s_d = \sqrt{\frac{\sum (d - \bar{d})^2}{n-1}},$$

а стандартная ошибка

$$s_{\bar{d}} = \frac{s_d}{\sqrt{n}}.$$

Таким образом, критерий Стьюдента принимает вид:

$$t = \frac{\bar{d} - \delta}{s_{\bar{d}}}.$$

При условии справедливости нулевой гипотезы $\delta = 0$. Подставив это значение в формулу, получим:

$$t = \frac{\bar{d}}{s_{\bar{d}}}.$$

Осталось сравнить полученное значение с критическим для выбранного уровня значимости и числа степеней свободы $\nu = n - 1$.

Подытожим. Когда имеются данные об изменении интере-

сующего признака у каждого больного, для оценки статистической значимости этих изменений нужно сделать следующее.

- Вычислить величину изменения для каждого больного d.
- Вычислить среднее этих изменений $\bar{d}$ и его стандартную ошибку $s_{\bar{d}}$.
- Вычислить значение критерия Стьюдента $t = \bar{d}/s_{\bar{d}}$.
- Сравнить полученное значение t с критическим для числа степеней свободы $\nu = n - 1$.

Если обычный критерий Стьюдента требует нормального распределения самих данных, то парный критерий Стьюдента требует нормального распределения их *изменений*.

Курение и функция тромбоцитов

Известно, что курение способствует развитию ишемический болезни сердца. Известно также, что определенную роль в патогенезе этого заболевания играют тромбоциты. Связан ли эффект курения с влиянием на тромбоциты? В поисках ответа на этот вопрос П. Левин исследовал влияние курения на функцию тромбоцитов*. Одним из показателей, который интересовал исследователя, была *агрегация тромбоцитов* – доля тромбоцитов, слипшихся под воздействием аденозиндифосфата – вещества, стимулирующего агрегацию.

Одиннадцати добровольцам было предложено выкурить по сигарете. Перед курением и сразу после него были взяты пробы крови и определена агрегация тромбоцитов.

Результаты представлены на рис. 9.2. Левый столбик образовали наблюдения до выкуривания сигареты, правый — после. Отрезками соединены наблюдения, относящиеся к одному добровольцу. Когда из одной точки на рисунке выходит два отрезка, это значит, что данный результат наблюдался у двух больных. Агрегация тромбоцитов до курения составила в среднем 43,1%, после курения – 53,5%. Стандартные отклонения равны 15,9 и 18,7% соответственно. Уже при взгляде на эти цифры ясно, что о статистической значимости различий вряд ли может идти речь.

* P. H. Levine. An acute effect of cigarette smoking on platelet function: a possible link between smoking and arterial thrombosis. *Circulation*, 48: 619—623, 1973.

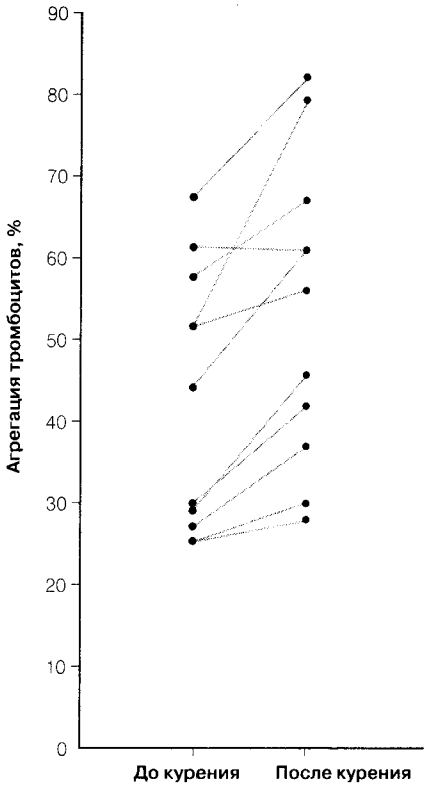


Рис. 9.2. Агрегация тромбоцитов до и после выкуривания сигареты. Агрегация тромбоцитов повысилась, но значит ли это, что она повысилась от табака?

Действительно, расчет критерия Стьюдента в том виде, в каком он был изложен в гл. 4, дает $t = 1{,}405$, что меньше критического значения для 5% уровня значимости и 20 степеней свободы. При сравнении двух независимых групп следовало бы признать влияние курения статистически не значимым. Однако в данном случае наблюдалась одна группа, причем данные позволяют вычислить изменения для каждого ее члена.

Сделав это, мы обнаружим, что у всех обследованных, за ис-

ключением одного, агрегация тромбоцитов после курения повысилась. Выпишем изменения у каждого из обследованных. Получим 2, 4, 10, 12, 16, 15, 4, 27, 9, –1 и 15%. Средняя величина изменения $\bar{d} = 10{,}3\%$. Стандартное отклонение величины изменения $s_d = 8\%$ и стандартная ошибка $s_{\bar{d}} = 8{,}0/\sqrt{11} = 2{,}41\%$. Тогда:

$$t = \frac{\bar{d}}{s_{\bar{d}}} = \frac{10{,}3}{2{,}41} = 4{,}27.$$

В табл. 4.1 находим критическое значение $t_{0{,}01}$ для уровня значимости 0,01 и $\nu = n - 1$ степеней свободы. Оно равно 3,169, то есть меньше полученного нами. Таким образом, повышение агрегации тромбоцитов после курения статистически значимо.

На этом выводе Левин не остановился. Если курение повышает агрегацию тромбоцитов, то значит ли это, что повышение вызвано курением *табака*? Нет, не значит. С тем же успехом можно признать причиной вдыхание окиси углерода, выделяющейся при горении сигареты. Не менее веской причиной будет и волнение, испытываемое участниками эксперимента. Имеющиеся данные не позволяют отвергнуть такие объяснения. Значит, нужно провести эксперименты, совпадающие с исходным во всем, кроме интересующего нас фактора — в данном случае курения сигарет с табаком. Именно это и сделал Левин. Добровольцам пришлось выкуривать не только обычные, но и безникотиновые сигареты из салатных листьев. Кроме того, им предлагали подержать в зубах незажженную сигарету, изображая курение. Результаты приведены на рис. 9.3 вместе с данными с рис. 9.2. Оказалось, что в отличие от обычной сигареты незажженная или безникотиновая сигарета не вызывает повышения агрегации тромбоцитов.

Разобранное исследование служит иллюстрацией следующего правила.

Единственным различием между контрольной и экспериментальной группой должно быть воздействие исследуемого, и никакого другого, фактора.

Чем лучше удается вычленить действие изучаемого фактора, тем достовернее выводы эксперимента. Так, рассмотренный экс-

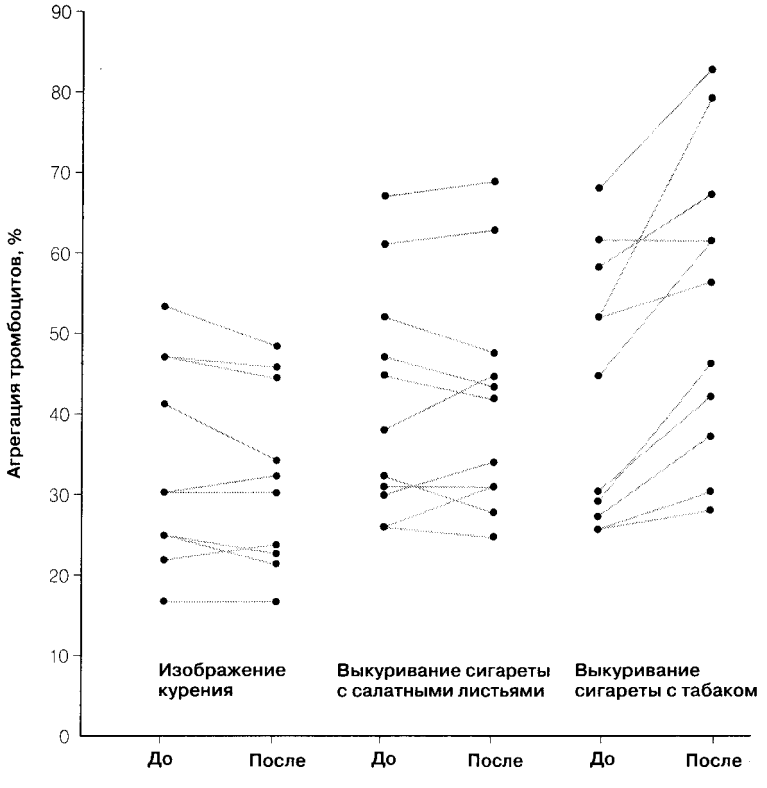


Рис. 9.3. Агрегация тромбоцитов до и после изображения курения с незажженной сигаретой, выкуривания сигареты с салатными листьями, выкуривания сигареты с табаком. Похоже, что именно табак, а не сам факт курения и не дым вызывает повышение агрегации тромбоцитов.

перимент доказал, что повышение агрегации тромбоцитов вызвано не просто курением, а именно курением табака.

Затронув вопрос о планировании эксперимента, стоит упомянуть еще об одной важной проблеме. Кроме необходимости выделить исследуемый фактор и тем самым исключить неоднозначное толкование результатов эксперимента, нужно избежать искажений, привносимых участниками эксперимента. В меди-

цинских экспериментах человек не только оказывает воздействие и наблюдает его результат – он присутствует и как объект наблюдений. Но люди пристрастны и внушаемы. Пристрастность экспериментатора может повлечь неосознанную подтасовку. А лаборантке, поборнице некурения, не составит труда чуть-чуть завысить долю склеившихся тромбоцитов в крови курильщика и чуть-чуть занизить ее для некурящего.

При проведении клинических испытаний на первый план выходит роль больного. Особенно велика она, если критерием эффективности служат его собственные оценки (боль уменьшилась – усилилась, стал спать лучше – хуже). Вера больного в новый метод лечения – могучий (и благотворный) фактор, однако объективной оценке он мешает. Вернемся к исследованиям агрегации тромбоцитов. Как в данном случае на результат эксперимента может повлиять испытуемый? Человек не может усилием воли изменять состояние своих тромбоцитов, однако, обратившись еще раз к рис. 9.3, можно заметить, что у добровольцев, которым *только еще предстояло* выкурить (возможно, безвредную салатную) сигарету, агрегация тромбоцитов была заметно выше, чем у тех, которым было известно, что им придется лишь подержать сигарету в зубах. Следовательно, не только субъективные оценки, но и объективные показатели могут изменяться под влиянием отношения испытуемого к экспериментальному воздействию.

Чтобы исключить влияние субъективного фактора, Левин применил *двойной слепой метод*. Суть метода в том, что экспериментальное воздействие не известно ни испытуемым, ни наблюдателям, оценивающим его результаты. В эксперименте Левина ни исследователям, ни добровольцам не было известно содержимое сигарет, а производившим анализ крови лаборантам — курил ли доброволец, и если да, то что именно.

В действительности исследование Левина не было полностью двойным слепым (о чем свидетельствуют различия исходной агрегации тромбоцитов). Действительно, даже если о содержимом сигареты добровольцам не сообщали, они могли легко определить его на вкус.

Предвидя подобные трудности, исследование часто заранее планируют как *простое слепое*. В этом случае одна из сторон

(обычно наблюдатель) осведомлена о характере экспериментального воздействия, а другая (обычно испытуемый) – нет. Наконец, характер исследования может быть таков, что ни одну из сторон нельзя держать в полном неведении и обе располагают частью информации – в таких случаях говорят о *частично слепом исследовании.*

Завершая обсуждение парного критерия Стьюдента, повторим, что он используется для проверки эффективности *одного метода* лечения в случае, когда имеются данные о состоянии каждого участника *до и после* лечения. Когда же требуется сравнить эффективность *нескольких методов* лечения, испытанных на *одних и тех же* больных, применяют *дисперсионный анализ повторных наблюдений*. Для его изложения нам потребуется пересмотреть тот вариант дисперсионного анализа, который был изложен в гл. 3, то есть вариант на случай использования *разных методов* для лечения *разных больных*. Затем перейдем к варианту дисперсионного анализа на случай повторных наблюдений за одними и теми же больными, подвергаемыми разным методам лечения.

НОВЫЙ ПОДХОД К ДИСПЕРСИОННОМУ АНАЛИЗУ*

Напомним вкратце схему дисперсионного анализа, изложенную в гл. 3. В качестве нулевой гипотезы мы брали предположение о том, что несколько (обычно более двух) методов лечения обладают равной эффективностью, то есть экспериментальные группы — это просто выборки из одной нормально распределенной совокупности и различия между ними обусловлены случайностью. Для проверки нулевой гипотезы мы сравнивали разброс

---

* Если этот раздел, посвященный дисперсионному анализу повторных измерений, покажется вам слишком утомительным из-за обилия выкладок, пропустите его при первом чтении. Только не забудьте вернуться, когда возникнет необходимость. А она обязательно возникнет. Эксперименты, для обработки которых предназначен этот вариант дисперсионного анализа, типичны для медицины. Сам же анализ, увы, не очень. Чаще приходится сталкиваться с многократным использованием критерия Стьюдента, совершенно ошибочным (см. гл. 4).

Таблица 9.1. Сердечный выброс, л/мин

	Группа			
	Контрольная	Макароны	Мясо	Фрукты
	4,6	4,6	4,3	4,3
	4,7	5,0	4,4	4,4
	4,7	5,2	4,9	4,5
	4,9	5,2	4,9	4,9
	5,1	5,5	5,1	4,9
	5,3	5,5	5,3	5,0
	5,4	5,6	5,6	5,6
Среднее	4,96	5,23	4,93	4,80
Вариация	0,597	0,734	1,294	1,200
Среднее по всем группам = 4,98				
Общая вариация = 4,51				

значений относительно групповых средних с разбросом самих групповых средних. Если разброс средних значительно превышал разброс значений, мы отвергали нулевую гипотезу. В качестве показателя разброса мы использовали дисперсию. Дисперсию можно определить как сумма квадратов отклонений, деленную на число степеней свободы. Теперь показателем разброса будет служить сама сумму квадратов отклонений*, которую мы будем называть *вариацией*. Основываясь на вариации, мы повторим построение дисперсионного анализа. Перспектива второй раз разбирать уже знакомый метод не слишком вдохновляет, однако мы будем вознаграждены: новый взгляд позволит нам перейти к дисперсионному анализу повторных измерений.

В гл. 3 мы рассмотрели такой пример. Чтобы выяснить, влияет ли питание на сердечный выброс, из 200 обитателей городка были случайным образом выбраны четыре группы по семь человек в каждой. Члены первой (контрольной) группы продолжали питаться как обычно, членам второй группы пришлось есть одни макароны, третьей — мясо, а четвертой — фрукты. Эксперимент длился ровно месяц, после чего у каждого участника был изме-

* Такой подход мы уже использовали в гл. 8 при рассмотрении регрессионного анализа.

рен сердечный выброс. Как видно из рис. 3.1, диета не влияет на величину сердечного выброса. Экспериментальные группы — это просто четыре случайные выборки из нормально распределенной совокупности. Однако рис. 3.1 недоступен исследователю, в распоряжении которого есть только данные об участниках эксперимента. Эти данные представлены на рис. 3.2 и в табл. 9.1. Как видим, группы все же различаются по средней величине сердечного выброса. Можно ли объяснить эти различия случайностью?

Новые обозначения

Прежде чем двигаться дальше, введем новые обозначения (табл. 9.2). Отвлечемся от фруктов и макарон и вообще специфики рассматриваемого эксперимента. Перенумеруем группы от 1 до 4. Участников исследования также перенумеруем и впредь будем называть больными (хотя применительно к данному случаю это не совсем удачно). Значения признака (в данном случае это сердечный выброс) обозначим $X_{\text{гб}}$, например X_{25} — значение у 5-го больного 2-й группы. Средние по группам обозначим $\overline{X}_{\text{г}}$, например $\overline{X}_3$ — среднее по 3-й группе. Под средними в таблице мы видим групповые вариации $S_{\text{г}}$ — суммы квадратов отклонений от среднего по группе:

$$S_{\text{г}} = \sum_{\text{б}} (X_{\text{гб}} - \overline{X}_{\text{г}})^2.$$

Значок «б» под символом суммы означает, что мы суммируем значения для всех больных данной группы. Для примера рассчитаем вариацию для 1-й группы:

$$S_1 = \sum_{\text{б}} (X_{1\text{б}} - \overline{X}_1)^2 =$$

$$= (4{,}6 - 4{,}96)^2 + (4{,}7 - 4{,}96)^2 + (4{,}7 - 4{,}96)^2 + (4{,}9 - 4{,}96)^2 +$$

$$+ (5{,}1 - 4{,}96)^2 + (5{,}3 - 4{,}96)^2 + (5{,}4 - 4{,}96)^2 = 0{,}597.$$

Вспомним определение выборочной дисперсии:

$$s^2 = \frac{\sum (X - \overline{X})^2}{n - 1},$$

Таблица 9.2. Обозначения однофакторного дисперсионного анализа

	Группа			
	1	2	3	4
	X_{11}	X_{21}	X_{31}	X_{41}
	X_{12}	X_{22}	X_{32}	X_{42}
	X_{13}	X_{23}	X_{33}	X_{43}
	X_{14}	X_{24}	X_{34}	X_{44}
	X_{15}	X_{25}	X_{35}	X_{45}
	X_{16}	X_{26}	X_{36}	X_{46}
	X_{17}	X_{27}	X_{37}	X_{47}
Среднее $\bar{X}_г$	$\bar{X}_1$	$\bar{X}_2$	$\bar{X}_3$	$\bar{X}_4$
Вариация $S_г$	$\sum_б (X_{1б} - \bar{X}_1)^2$	$\sum_б (X_{2б} - \bar{X}_2)^2$	$\sum_б (X_{3б} - \bar{X}_3)^2$	$\sum_б (X_{4б} - \bar{X}_4)^2$
Среднее по всем группам $\bar{X}$				
Общая вариация $\sum_г \sum_б (X_{гб} - \bar{X})^2$				

где n — объем выборки. В числителе стоит сумма квадратов отклонений от выборочного среднего, то есть вариация. Тем самым

$$s^2 = \frac{S}{n-1}.$$

Следовательно, выборочную дисперсию для группы можно записать как

$$s_г^2 = \frac{S_г}{n-1},$$

где n — численность группы. Если все выборки извлечены из одной совокупности, оценкой ее дисперсии можно взять среднее выборочных дисперсий. Такая оценка называется *внутригрупповой дисперсией*:

$$s_{вну}^2 = \frac{1}{m}(s_1^2 + s_2^2 + s_3^2 + s_4^2),$$

где m — число групп, в данном случае равное 4. Заменим теперь

каждую выборочную дисперсию ее выражением через вариацию:

$$s^2_{\text{вну}} = \frac{1}{m}\left(\frac{S_1}{n-1} + \frac{S_2}{n-1} + \frac{S_3}{n-1} + \frac{S_4}{n-1}\right),$$

где n — численность каждой из групп. Перенесем $n-1$ под дробную черту:

$$s^2_{\text{вну}} = \frac{1}{m}\frac{S_1 + S_2 + S_3 + S_4}{n-1}.$$

В числителе — сумма вариаций по всем группам. Назовем ее *внутригрупповой вариацией* и обозначим $S_{\text{вну}}$. Обратите внимание, что внутригрупповая вариация — это сумма квадратов отклонений от групповых средних, поэтому она не зависит от того, различаются эти средние или нет.

В примере с диетой и сердечным выбросом

$$S_{\text{вну}} = 0{,}597 + 0{,}734 + 1{,}294 + 1{,}200 = 3{,}825.$$

Перепишем еще раз формулу для внутригрупповой дисперсии:

$$s^2_{\text{вну}} = \frac{S_{\text{вну}}}{m(n-1)}.$$

В знаменателе теперь стоит выражение, знакомое нам по гл. 3. Это внутригрупповое число степеней свободы: $\nu_{\text{вну}} = m(n-1)$. В рассматриваемом примере $\nu_{\text{вну}} = 4(7-1) = 24$. Таким образом, внутригрупповую дисперсию можно выразить через внутригрупповую вариацию и внутригрупповое число степеней свободы:

$$s^2_{\text{вну}} = \frac{S_{\text{вну}}}{\nu_{\text{вну}}}.$$

По данным из табл. 9.1 находим

$$s^2_{\text{вну}} = \frac{3{,}825}{24} = 0{,}159.$$

Как нам известно из гл. 3, чтобы вычислить F, помимо внут-

ригрупповой нужна межгрупповая дисперсия. Внутригрупповую дисперсию нам удалось выразить через вариацию и число степеней свободы. Проделаем те же действия с межгрупповой дисперсией.

Межгрупповая дисперсия $s^2_{\text{меж}}$ отражает разброс групповых средних. Мы вычисляли ее по формуле

$$s^2_{\text{меж}} = ns^2_{\bar{X}}.$$

Здесь $s^2_{\bar{X}}$ равно

$$s^2_{\bar{X}} = \frac{(\bar{X}_1 - \bar{X})^2 + (\bar{X}_2 - \bar{X})^2 + (\bar{X}_3 - \bar{X})^2 + \ldots + (\bar{X}_m - \bar{X})^2}{m-1}.$$

В более общем виде:

$$s^2_{\bar{X}} = \frac{\sum_{\text{г}} (\bar{X}_{\text{г}} - \bar{X})^2}{m-1},$$

где m — число групп. Под символом суммы стоит значок «г», это означает, что теперь мы суммируем по группам, а не по больным.

Подставив это выражение в формулу межгрупповой дисперсии, получим:

$$s^2_{\text{меж}} = \frac{n\sum_{\text{г}} (\bar{X}_{\text{г}} - \bar{X})^2}{m-1}.$$

Величину в числителе назовем межгрупповой вариацией и обозначим $S_{\text{меж}}$:

$$S_{\text{меж}} = n\sum_{\text{г}} (\bar{X}_{\text{г}} - \bar{X})^2.$$

Тогда

$$s^2_{\text{меж}} = \frac{S_{\text{меж}}}{m-1}.$$

В этой формуле мы снова обнаруживаем число степеней свободы из гл. 3, на этот раз это межгрупповое число степеней свободы: $\nu_{\text{меж}} = m-1$. Тем самым

$$s^2_{\text{меж}} = \frac{S_{\text{меж}}}{\nu_{\text{меж}}}.$$

В нашем примере (табл. 9.1) $\nu_{\text{меж}} = m - 1 = 4 - 1 = 3$. Тогда

$$s^2_{\text{меж}} = 0{,}685/3 = 0{,}228.$$

Формула для критерия F в новых обозначениях принимает вид:

$$F = \frac{S_{\text{меж}}/\nu_{\text{меж}}}{S_{\text{вну}}/\nu_{\text{вну}}}.$$

Соответственно, в рассматриваемом примере

$$F = \frac{0{,}228}{0{,}159} = 1{,}4.$$

Новая формула для F получена непосредственно из приведенной в гл. 3 и отличается от нее только обозначениями. Поэтому, конечно, значение $F = 1{,}4$ совпадает с найденным в гл. 3.

Естественно спросить, зачем же потребовались столь пространные рассуждения и многочисленные тождественные замены? Неужели для одного только повторения ранее полученных результатов? Ответ состоит в том, что переход к использованию вариации дает возможность понять, из каких компонентов она складывается, и в дальнейшем перейти к дисперсионному анализу повторных измерений.

Разложение общей вариации

Внутригрупповая вариация $S_{\text{вну}}$ служит мерой разброса значений внутри групп. В свою очередь, межгрупповая вариация $S_{\text{меж}}$ — это мера разброса групповых средних, то есть различий между группами. Но существует и мера общего разброса значений. Это общая сумма квадратов отклонений всех наблюдаемых значений от их общего среднего. Она называется *общей вариацией* и обозначается $S_{\text{общ}}$:

$$S_{\text{общ}} = \sum_{\text{г}} \sum_{\text{б}} (X_{\text{гб}} - \bar{X})^2.$$

Два символа суммы означают, что суммирование производится по всем группам и всем больным внутри каждой группы.

Число степеней свободы общей вариации обозначается $\nu_{общ}$ и равно $mn-1$, то есть оно на единицу меньше общего числа больных (m — число групп, n — число больных в каждой группе).

В рассматриваемом примере $S_{общ} = 4{,}51$ и $\nu_{общ} = 4 \times 7 - 1 = 27$

Обратите внимание, что общая дисперсия, вычисленная по всем наблюдениям, равна

$$s^2_{общ} = \frac{\sum_г \sum_б (X_{гб} - \bar{X})^2}{mn-1} = \frac{S_{общ}}{mn-1} = \frac{S_{общ}}{\nu_{общ}}.$$

Существует ли связь между рассмотренными видами вариации: общей, внутригрупповой и межгрупповой? Оказывается, существует, и очень простая. *Общая вариация равна сумме внутригрупповой и межгрупповой вариаций:*

$$S_{общ} = S_{вну} + S_{меж}.$$

Докажем справедливость этого разложения (это доказательство можно пропустить). Тождественно верно

$$(X_{гб} - \bar{X}) = (\bar{X}_{гб} - \bar{X}_г) + (\bar{X}_г - \bar{X}).$$

Возведем левую и правую части тождества в квадрат:

$$(X_{гб} - \bar{X})^2 = [(\bar{X}_{гб} - \bar{X}_г) + (\bar{X}_г - \bar{X})]^2.$$

Просуммируем левую часть по всем наблюдениям:

$$\sum_г \sum_б (X_{гб} - \bar{X})^2.$$

Это не что иное, как общая вариация $S_{общ}$.

Правая часть преобразуется в

$$(\bar{X}_{гб} - \bar{X}_г)^2 + 2(\bar{X}_{гб} - \bar{X}_г)(\bar{X}_г - \bar{X}) + (\bar{X}_г - \bar{X})^2.$$

Суммируя по всем наблюдениям, получим

$$\sum_г \sum_б (X_{гб} - \bar{X}_г)^2 + 2\sum_г \sum_б (X_{гб} - \bar{X}_г)(\bar{X}_г - \bar{X}) + \sum_г \sum_б (\bar{X}_г - \bar{X})^2.$$

Первый член этого выражения, $\sum_{г}\sum_{б}(X_{гб}-\bar{X}_{г})^2$, представляет собой значение $S_{вну}$.

Покажем, что второй член, $2\sum_{г}\sum_{б}(X_{гб}-\bar{X}_{г})(\bar{X}_{г}-\bar{X})$, тождественно равен нулю.

В самом деле, разность $(\bar{X}_{г}-\bar{X})$ в каждой из групп постоянна, и поэтому ее можно вынести за знак суммирования по больным:

$$2\sum_{г}\sum_{б}(X_{гб}-\bar{X}_{г})(\bar{X}_{г}-\bar{X})=2\sum_{г}(\bar{X}_{г}-\bar{X})\sum_{б}(X_{гб}-\bar{X}_{г}).$$

Но $\bar{X}_{г}$ — это среднее по группе, то есть

$$\bar{X}_{г}=\frac{\sum_{б}X_{гб}}{n}.$$

В таком случае

$$\sum_{б}(X_{гб}-\bar{X}_{г})=\sum_{б}X_{гб}-\sum_{б}\bar{X}_{г}=\sum_{б}X_{гб}-n\bar{X}_{г}=$$

$$=n\left(\frac{\sum_{б}X_{гб}}{n}-\bar{X}_{г}\right)=n\left(\bar{X}_{г}-\bar{X}_{г}\right)=0.$$

Рассмотрим третий член. Поскольку $\bar{X}_{г}-\bar{X}$ для всех больных в группе одинаково,

$$\sum_{г}\sum_{б}(\bar{X}_{г}-\bar{X})^2=n\sum_{г}(\bar{X}_{г}-\bar{X})^2,$$

а это величина $S_{меж}$.

Итак, имеем:

$$S_{общ}=S_{вну}+0+S_{меж}=S_{вну}+S_{меж},$$

что и требовалось доказать.

Как общая вариация разлагается на две составляющие — внутригрупповую и межгрупповую, так и общее число степеней свободы разлагается на внутригрупповое и межгрупповое. Действительно, поскольку

$\nu_{общ}=mn-1$, $\nu_{меж}=m-1$ и $\nu_{вну}=m(n-1)$, то

$$\nu_{меж}+\nu_{вну}=m-1+m(n-1)=m(1+n-1)-1=mn-1=\nu_{общ}.$$

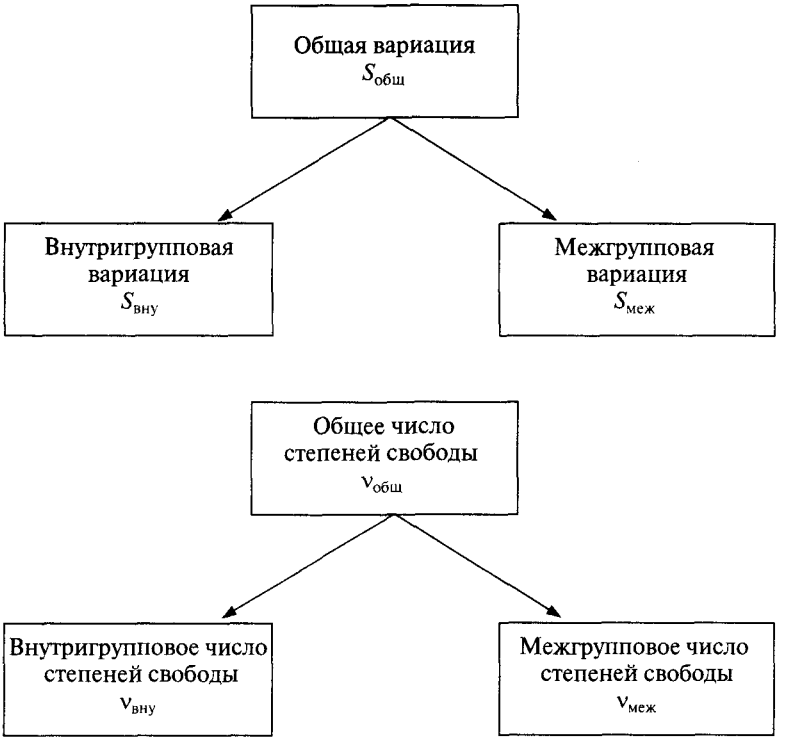


Рис. 9.4. Разложение вариации и числа степеней свободы при дисперсионном анализе.

Таблица 9.3. Таблица дисперсионного анализа для эксперимента с 4 диетами

	Вариация	Число степеней свободы	Дисперсия
Межгрупповая	0,685	3	0,228
Внутригрупповая	3,825	24	0,159
Общая	4,51	27	

$$F = \frac{S_{меж}/\nu_{меж}}{S_{вну}/\nu_{вну}} = \frac{0,228}{0,159} = 1,4$$

Рис. 9.5. Разложение вариации и числа степеней свободы при дисперсионном анализе повторных измерений.

Оба разложения изображены на рис. 9.4. Перечисленные величины обычно включают в *таблицы дисперсионного анализа* наподобие табл. 9.3.

Теперь, наконец, мы располагаем средствами, необходимыми в дисперсионном анализе повторных измерений.

ДИСПЕРСИОННЫЙ АНАЛИЗ ПОВТОРНЫХ ИЗМЕРЕНИЙ

До сих пор мы имели дело с несколькими группами больных, которые подвергались различным методам лечения. В дисперсионном анализе повторных измерений ситуация иная: *одни и те же* больные последовательно подвергаются *нескольким* методам лечения или просто наблюдаются в несколько последовательных моментов времени. По-другому распределяется и общая вариация $S_{\text{общ}}$ (рис. 9.5). Прежде всего можно выделить межиндивидуальную ($S_{\text{МИ}}$) и внутрииндивидуальную ($S_{\text{ВИ}}$) вариацию, последняя, в свою очередь, распадается на обусловленную методом лечения ($S_{\text{ле}}$) и остаточную ($S_{\text{ост}}$), обусловленную случайными колебаниями, ошибкой измерения и т. п.

Обозначения, которые мы будем использовать в дисперсионном анализе повторных измерений, приведены в табл. 9.4. Представлены 4 больных, каждого из которых последовательно лечили 3 методами. Значения интересующего нас признака обоз-

Таблица 9.4. Обозначения, используемые в дисперсионном анализе повторных измерений

	Метод лечения			Среднее	Вариация
Больной	1	2	3	$\bar{X}_{\text{б}}$	$S_{\text{ВИб}}$
1	X_{11}	X_{21}	X_{31}	$\bar{X}_1$	$\sum_{\text{м}}(X_{\text{м}1}-\bar{X}_1)^2$
2	X_{12}	X_{22}	X_{32}	$\bar{X}_2$	$\sum_{\text{м}}(X_{\text{м}2}-\bar{X}_2)^2$
3	X_{13}	X_{23}	X_{33}	$\bar{X}_3$	$\sum_{\text{м}}(X_{\text{м}3}-\bar{X}_3)^2$
4	X_{14}	X_{24}	X_{34}	$\bar{X}_4$	$\sum_{\text{м}}(X_{\text{м}4}-\bar{X}_4)^2$
Среднее	$\bar{T}_1$	$\bar{T}_2$	$\bar{T}_3$		

начены $X_{мб}$, например, X_{12} — значение у 2-го больного при 1-м методе лечения, X_{31} — значение у 1-го больного при 3-м методе лечения и так далее. Величины $\bar{X}_б$ ($\bar{X}_1$, $\bar{X}_2$, $\bar{X}_3$, и $\bar{X}_4$) — это «индивидуальные» средние (средние значения признака при всех методах лечения у 1-го, 2-го и т. д. больного):

$$\bar{X}_б = \frac{\sum_м X_{мб}}{m},$$

где m — число методов лечения. $\bar{T}_м$ ($\bar{T}_1$, $\bar{T}_2$, $\bar{T}_3$, и $\bar{T}_4$) — средние значения признака у всех больных при 1-м, 2-м и т. д. методе лечения:

$$\bar{T}_м = \frac{\sum_б X_{мб}}{n},$$

где n — число больных.

Общая вариация — это сумма квадратов отклонений всех значений (у всех больных при всех методах лечения) от общего среднего, которое составляет

$$\bar{X} = \frac{\sum_м \sum_б X_{мб}}{mn};$$

таким образом,

$$S_{общ} = \sum_м \sum_б (X_{мб} - \bar{X})^2.$$

Соответствующее число степеней свободы $\nu_{общ} = mn - 1$.

Общая вариация складывается из межиндивидуальной и внутрииндивидуальной вариации. Рассчитаем внутрииндивидуальную вариацию $S_{ВИ}$. У первого больного сумма квадратов отклонений от индивидуального среднего $\bar{X}_1$ равна

$$S_{ВИ_1} = \sum_м (X_{м1} - \bar{X}_1)^2.$$

У второго больного

$$S_{ВИ_2} = \sum_м (X_{м2} - \bar{X}_2)^2$$

и так далее. Чтобы рассчитать внутрииндивидуальную вариацию, просуммируем $S_{ВИ_б}$ по всем больным:

$$S_{ВИ} = S_{ВИ_1} + S_{ВИ_2} + S_{ВИ_3} + S_{ВИ_4} = \sum_{б}\sum_{м}(X_{мб} - \bar{X}_б)^2 .$$

Соответствующее число степеней свободы составляет $\nu_{ВИ} = n(m-1)$.

Перейдем к межиндивидуальной вариации. Она складывается из квадратов отклонений индивидуальных средних $\bar{X}_б$ от общего среднего $\bar{X}$:

$$S_{МИ} = m\sum(\bar{X}_б - \bar{X})^2 .$$

Множитель m появляется из-за того, что каждое $\bar{X}_б$ — это среднее по m методам лечения. Число степеней свободы $\nu_{МИ} = n-1$.

Можно показать*, что общая вариация равна сумме внутри- и межиндивидуальной вариаций:

$$S_{общ} = S_{ВИ} + S_{МИ} .$$

Теперь из внутрииндивидуальной вариации нам предстоит выделить вариацию, связанную с лечением $S_{ле}$, и остаточную вариацию $S_{ост}$, связанную со случайными отклонениями и ошибками измерения. Вариация, связанная с лечением, складывается из квадратов отклонений средних по методам лечения $\bar{T}_м$ от общего среднего $\bar{X}$:

$$S_{ле} = n\sum(\bar{T}_м - \bar{X})^2 .$$

Наличие коэффициента n связано с тем, что каждое $T_м$ — это среднее по n больным.

Соответствующее число степеней свободы $\nu_{ле} = m-1$.

Остаточная вариация — вторая составляющая внутрииндивидуальной вариации — получается вычитанием:

$$S_{ост} = S_{ВИ} - S_{ле} .$$

---

* Вывод этого равенства см. в: B. J. Winer, D. R. Brown, K. M. Michels. Statistical principles in experimental design, 3d ed. McGraw-Hill, New York, 1991.

Аналогично вычисляется и остаточное число степеней свободы $\nu_{ост}$:

$$\nu_{ост} = \nu_{ВИ} - \nu_{ле} = n(m-1) - (m-1) = (n-1)(m-1).$$

Теперь мы можем получить две независимые оценки дисперсии: на основании вариации, связанной с лечением

$$s^2_{ле} = \frac{S_{ле}}{\nu_{ле}},$$

и на основании остаточной вариации:

$$s^2_{ост} = \frac{S_{ост}}{\nu_{ост}},$$

после чего можно применить знакомый нам критерий F:

$$F = \frac{s^2_{ле}}{s^2_{ост}}.$$

Далее следует поступить как при обычном дисперсионном анализе. Вычисленное значение F сравнивают с критическим для выбранного уровня значимости и числа степеней свободы. Чтобы воспользоваться табл. 3.1, нужно в качестве $\nu_{меж}$ взять $\nu_{ле}$, а в качестве $\nu_{вну}$ — соответственно $\nu_{ост}$.

Боюсь, читателя утомили сложные выкладки и громоздкие термины, которыми несколько перегружен этот раздел. Пора перейти к практическим применениям. Как мы уже говорили, дисперсионный анализ повторных наблюдений можно использовать не только когда к одним и тем же больным применяется несколько методов лечения, но и когда больные просто наблюдаются в несколько разных моментов времени. Именно на таком, очень простом примере мы и рассмотрим применение дисперсионного анализа повторных измерений.

Гидралазин при первичной легочной гипертензии

Первичная легочная гипертезия — редкое и чрезвычайно тяжелое заболевание, при котором вследствие неизвестных причин повышается давление в артериях легких. Стенки артерий утол-

щаются, что затрудняет газообмен в легких. Из-за повышенной нагрузки на правый желудочек страдает сердце. Без лечения больные живут не более нескольких лет. Гидралазин — препарат, расширяющий сосуды, — успешно используется при гипертонической болезни. Л. Рубин и Р. Питер* предположили, что его можно использовать и при первичной легочной гипертензии. В исследование вошли 4 больных. Измерения производили трижды: перед началом лечения, спустя 48 ч и 3—6 мес лечения. (В дальнейшем мы будем говорить просто о 1, 2 и 3-м измерениях.) Измеряли, в частности, легочное сосудистое сопротивление. Этот показатель отражает тяжесть легочной гипертензии: чем выше сопротивление, тем тяжелее гипертензия. Результаты представлены на рис. 9.6. Похоже, данные говорят в пользу препарата. С другой стороны, они получены на малочисленной выборке. Поэтому не будем доверяться впечатлениям, а воспользуемся дисперсионным анализом повторных измерений.

Обратимся к табл. 9.5. Здесь помимо первичных данных приведены средние значения легочного сосудистого сопротивления для каждого из 4 больных и для каждого из трех моментов измерения. Например, у второго больного среднее легочное сосудистое сопротивление составило

$$\bar{X}_2 = \frac{17{,}0 + 6{,}3 + 6{,}2}{3} = 9{,}83,$$

а среднее легочное сосудистое сопротивление при 1-м измерении:

$$\bar{T}_1 = \frac{22{,}2 + 17{,}0 + 14{,}1 + 17{,}0}{3} = 17{,}58.$$

Среднее сопротивление по всем измерениям $\bar{X} = 11{,}63$, а общая вариация $S_{\text{общ}} = 289{,}82$.

В табл. 9.5 приведены также суммы квадратов отклонений от индивидуального среднего. Например, для второго больного

$$S_{\text{ВИ}_2} = (17{,}0 - 9{,}83)^2 + (6{,}3 - 9{,}83)^2 + (6{,}2 - 9{,}83)^2 = 77{,}05.$$

---

* L. J. Rubin and R. H. Peter. Oral hydralazine therapy for primary pulmonary hypertension. *N. Engl. J. Med.*, 302:69—73, 1980.

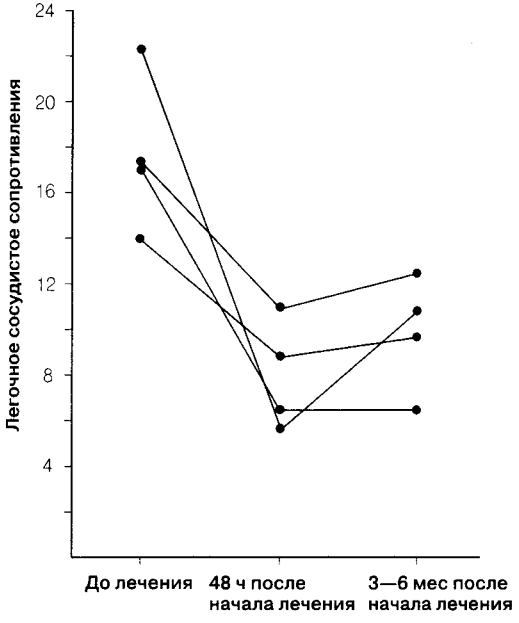


Рис. 9.6. Изменение легочного сосудистого сопротивления у 4 больных с легочной гипертензией при лечении гидралазином.

Внутрииндивидуальная вариация составляет

$$S_{\text{ВИ}} = 147{,}95 + 77{,}05 + 18{,}35 + 21{,}45 = 264{,}80.$$

Можно найти межиндивидуальную вариацию:

$$S_{\text{МИ}} = 3[(12{,}73 - 11{,}63)^2 + (9{,}83 - 11{,}63)^2 + (10{,}63 - 11{,}63)^2 + (13{,}33 - 11{,}63)^2] = 25{,}02.$$

Заметьте, что, как это и должно быть, выполняется равенство $S_{\text{общ}} = S_{\text{ВИ}} + S_{\text{МИ}}$.

Рассчитаем $S_{\text{ле}}$ (теперь эта вариация связана со временем, но мы оставим прежнее обозначение):

$$S_{\text{ле}} = 4[(17{,}58 - 11{,}63)^2 + (7{,}73 - 11{,}63)^2 + (9{,}60 - 11{,}63)^2] = 218{,}93.$$

Соответствующее число степеней свободы:

Таблица 9.5. Легочное сосудистое сопротивление у больных первичной легочной гипертензией на фоне лечения гидралазином

	Измерение				
Больной	1	2	3	Среднее	Вариация
1	22,2	5,4	10,6	12,73	147,95
2	17,0	6,3	6,2	9,83	77,05
3	14,1	8,5	9,3	10,63	18,35
4	17,0	10,7	12,3	13,33	21,45
Среднее	17,58	7,73	9,60		

Общее среднее $\bar{X} = 11{,}63$. *Общая вариация* $S_{\text{общ}} = 289{,}82$.

$$\nu_{\text{ле}} = m - 1 = 3 - 1 = 2.$$

Наконец, остаточная вариация определяется равенством

$$S_{\text{ост}} = S_{\text{ВИ}} - S_{\text{ле}} = 264{,}80 - 218{,}93 = 45{,}87$$

и имеет $\nu_{\text{ост}} = (n-1)(m-1) = (4-1)(3-1) = 6$ степеней свободы.

Все найденные величины сведены в табл. 9.6. Обратите внимание, что здесь общая вариация разложена на большее число составляющих, чем в табл. 9.3. Причина в том, что теперь рассматриваются результаты повторных измерений одной группы, а не однократных измерений нескольких групп.

Вычисляем оценку дисперсии на основании вариации, обусловленной лечением:

$$s^2_{\text{ле}} = \frac{S_{\text{ле}}}{\nu_{\text{ле}}} = \frac{218{,}93}{2} = 109{,}47$$

и на основании остаточной вариации:

$$s^2_{\text{ост}} = \frac{S_{\text{ост}}}{\nu_{\text{ост}}} = \frac{45{,}87}{6} = 7{,}65.$$

Теперь, наконец, можно вычислить F:

$$F = \frac{s^2_{\text{ле}}}{s^2_{\text{ост}}} = 14{,}31.$$

Критическое значение для числа степеней свободы $\nu_{\text{меж}} = 2$ и

Таблица 9.6. Таблица дисперсионного анализа (исследование гидралазина при первичной легочной гипертензии)

Вариация	Число степеней свободы	Оценка дисперсии
Межиндивидуальная $S_{МИ} = 25{,}02$	3	
Внутрииндивидуальная $S_{ВИ} = 264{,}80$	8	
обусловленная лечением $S_{ле} = 218{,}93$	2	109,47
остаточная $S_{ост} = 45{,}87$	6	7,65
Общая $S_{общ} = 289{,}82$	11	

$$F = \frac{s^2_{ле}}{s^2_{ост}} = 14{,}31$$

$\nu_{ВНУ} = 6$ составляет 10,92, то есть меньше полученного нами. Таким образом, легочное сосудистое сопротивление нельзя считать постоянным. По крайней мере в один из моментов легочное сосудистое сопротивление значимо отличается от наблюдаемого в остальные моменты. Ответить на вопрос, что это за момент и что это за отличия, дисперсионный анализ не может. Для этого следует воспользоваться методами множественных сравнений (гл. 4).

Как выявить различия в повторных измерениях

В гл. 4 мы познакомились с критерием Стьюдента с поправкой Бонферрони. Он вычисляется как обычный критерий Стьюдента:

$$t = \frac{\overline{X}_i - \overline{X}_j}{\sqrt{\frac{2s^2}{n}}}.$$

Однако уровень значимости в каждом из сравнений, согласно поправке Бонферрони, принимается равным $\alpha = \alpha'/k$, где α' — истинный уровень значимости (по всем сравнениям в целом), а k — число сравнений. Критерий Стьюдента с поправкой Бонферрони, как и другие методы множественного сравнения, применяется лишь после того, как дисперсионный анализ обнаружит сам факт существования различий.

При дисперсионном анализе повторных измерений схема использования критерия остается прежней. Отличие в том, что в формуле для t вместо s^2 следует взять остаточную дисперсию $s^2_{\text{ост}}$, а средние по группам заменить на средние по методам лечения (моментам наблюдения) $\overline{T}_{\text{м}}$. Тогда формула для t примет вид:

$$t = \frac{\overline{T}_i - \overline{T}_j}{\sqrt{\frac{2s^2_{\text{ост}}}{n}}}.$$

Полученное значение нужно сравнить с критическим значением для распределения Стьюдента при $\nu_{\text{ост}}$ степенях свободы.

Вернемся к эксперименту с гидралазином. Остаточная оценка дисперсии $s^2_{\text{ост}} = 7{,}65$. Число больных при каждом измерении $n = 4$.

Сравним 1-е и 2-е измерения:

$$t = \frac{17{,}58 - 7{,}73}{\sqrt{\frac{2 \times 7{,}65}{4}}} = 5{,}036.$$

Сравним 1-е и 3-е измерения:

$$t = \frac{17{,}58 - 9{,}60}{\sqrt{\frac{2 \times 7{,}65}{4}}} = 4{,}080.$$

И наконец, 2-е и 3-е измерения:

$$t = \frac{7{,}73 - 9{,}60}{\sqrt{\frac{2 \times 7{,}65}{4}}} = -0{,}9561.$$

Чтобы вероятность ошибочно обнаружить различие была в совокупности по всем трем сравнениям меньше 0,05, нужно в каждом отдельном сравнении использовать в три раза меньший уровень значимости $0{,}05/3 = 0{,}016$. Для этого уровня значимости

и при числе степеней свободы $\nu = 6$ находим по табл. 4.1 критическое значение, приближенно равное 3,37 (поскольку таблица не содержит значений для $\alpha = 0{,}016$, оно расчитывается приблизительно по соседним значениям $\alpha = 0{,}01$ и $\alpha = 0{,}02$).

Значения t для первых двух сравнений больше критического, а для третьего — меньше. Поэтому при уровне значимости 0,05 (но ни в коем случае не 0,016, используемом в каждом сравнении) различие в величине общего легочного сопротивления до и после приема гидралазина статистически значимо, а между измерениями на фоне приема гидралазина статистически незначимо.

Заканчивая обсуждение парных сравнений, скажем, что вместо поправки Бонферрони можно воспользоваться более точным критерием Ньюмена—Кейлса или критерием Тьюки. Кроме того, в рассматриваемом примере, где измерения, выполненные до начала лечения, играют роль «контрольной группы», пригоден и критерий Даннета для множественного сравнения с контрольной группой. Все эти критерии описаны в гл. 4. При их применении нужно, как и в случае критерия Стьюдента с поправкой Бонферрони, в качестве оценки дисперсии брать $s^2_{\text{ост}}$, а при нахождении критического значения использовать число степеней свободы остаточной вариации.

Чувствительность дисперсионного анализа повторных измерений

Чувствительность вычисляется так же, как в обычном дисперсионном анализе, с той разницей, что в качестве оценки для s используется $s_{\text{ост}}$, а вместо численности отдельных групп — численность единственной рассматриваемой группы.

КАЧЕСТВЕННЫЕ ПРИЗНАКИ: КРИТЕРИЙ МАК-НИМАРА

Парный критерий Стьюдента и дисперсионный анализ повторных измерений применимы, только если зависимый признак является числовым и, сверх того, подчиняется нормальному закону распределения. Как быть, если признак качественный, то есть имеет своими значениями не числа, а «названия» (с такими при-

знаками мы познакомились в гл. 5). Они часто встречаются в медицине. Например, диагноз — типичный качественный признак. Сейчас мы познакомимся с критерием Мак-Нимара. Он предназначен для анализа повторных измерений качественных признаков и в некотором смысле является аналогом парного критерия Стьюдента. Знакомство с новым критерием мы начнем с примера.

Проба с динитрохлорбензолом при онкологических заболеваниях

Ослабление иммунитета повышает риск онкологических заболеваний. Считается также, что при уже развившемся злокачественном новообразовании ослабление иммунитета — плохой прогностический признак и наоборот — сохранность иммунитета говорит о высокой вероятности успеха лечения. Для оценки состояния иммунитета применяется кожная проба с динитрохлорбензолом. Проба считается положительной, если через 48 часов после нанесения динитрохлорбензола на кожу развивается выраженная воспалительная реакция. Положительная проба говорит о сохранности иммунитета.

Ряд авторов оспаривают значение пробы, указывая, в частности, на то, что воспалительная реакции может быть вызвана местнораздражающим действием динитрохлорбензола и не отражает состояния иммунитета.

Чтобы выяснить этот вопрос, Рот и соавт.* проделали следующий опыт. На кожу больных наносили динитрохлорбензол и одновременно — на соседний участок кожи — кротоновое масло. Кротоновое масло оказывает местнораздражающее действие, которое не зависит от состояния иммунитета. Если оба раздражителя вызовут сходную реакцию, рассуждал автор, то в обоих случаях она не отражает состояния иммунитета.

В табл. 9.7 приведены результаты опыта. Знак «плюс» соответствует наличию реакции, знак «минус» — отсутствию. При виде такой таблицы хочется немедленно рассчитать χ^2. Посмот-

* J. A. Roth, F. R. Eilber, J. A. Nizle, D. L. Morton. Lack of correlation between skin reactivity to dinitrochlorobenzene and croton oil in patients with cancer. *N. Engl. J. Med.*, 293:388—389, 1975.

рим, что из этого получится. Вычисленное с поправкой Йейтса значение $\chi^2 = 1{,}107$. Это заметно меньше критического значения 3,841, соответствующего уровню значимости 0,05 при одной степени свободы. Напрашивается вывод вроде: «Статистически значимых различий между реакцией на динитрохлорбензол и кротоновое масло не выявлено».

В этой формулировке есть неточность, на первый взгляд незначительная. При построении критерия χ^2 в гл. 5 мы проверяли нулевую гипотезу об отсутствии *связи* между признаками. Например, мы предполагали, что аспирин не влияет на частоту тромбоза. Если нулевая гипотеза отвергалась, мы признавали существование связи между признаками. Если строки таблицы представлены двумя методами лечения, это равнозначно признанию различий эффективности этих методов. В данном случае это не так, поэтому мы должны ограничиться констатацией отсутствия *связи* между реакцией на динитрохлорбензол и кротоновое масло. В отличие от поспешного вывода, который мы привели выше, это утверждение говорит в пользу самостоятельного значения пробы с динитрохлорбензолом: если бы она давала те же результаты, что и проба с кротоновым маслом, это как раз и говорило бы о том, что ее результат, скорее всего, обусловлен местнораздражающим действием.

Этого мало. С помощью критерия Мак-Нимара мы покажем, что динитрохлорбензол дает *меньше* положительных результатов пробы, чем кротоновое масло.

Реакция только на динитрохлорбензол наблюдалась у 23 больных, а только на кротоновое масло — у 48. Если действие динитрохлорбензола и кротонового масла примерно одинаково, то больные, у которых наблюдалась реакция только на один раздражитель, разделились бы примерно поровну — у одной половины реакцию вызвал бы динитрохлорбензол, у другой — кротоновое масло. Следовательно, ожидаемое число в обоих случаях $(23 + 48)/2 = 35{,}5$. Для сравнения наблюдаемых чисел с ожидаемыми воспользуемся критерием χ^2. (Поскольку число степеней свободы равно 1, применим также поправку Йейтса.) Имеем:

Таблица 9.7. Кожная реакция на ДНХБ и кротоновое масло

		Реакция на динитрохлорбензол	
		+	–
Реакция на кротоновое масло	+	81	48
	–	23	21

$$\chi^2 = \sum \frac{\left(|O-E|-\frac{1}{2}\right)^2}{E} =$$

$$= \frac{\left(|23-35{,}5|-\frac{1}{2}\right)^2}{35{,}5} + \frac{\left(|48-35{,}5|-\frac{1}{2}\right)^2}{35{,}5} = 8{,}817.$$

Для уровня значимости 0,01 табличное значение χ^2 с одной степенью свободы равно 6,635 (см. табл. 5.7), то есть меньше вычисленного. Таким образом, оказывается, что действие динитрохлорбензола отличается от действия кротонового масла.

Рассмотренный пример показывает, сколь далекими от истины могут оказаться выводы при необоснованном применении статистических методов.

Критерий Мак-Нимара, подобно парному критерию Стьюдента, часто используется для выявления изменений в наблюдениях типа «до—после», когда интересующий нас признак принимает одно из двух значений («есть—нет»). Другое, очень важное, применение критерия связано с анализом парных наблюдений. Что это такое, вы узнаете, решив задачи 9.9 и 9.10.

А теперь перечислим шаги критерия Мак-Нимара.

- Исключите из рассмотрения больных, реакция которых была неизменной, и подсчитайте число тех, чья реакция изменилась.
- Поделите это число пополам.
- Вычислите меру отклонения наблюдаемого числа меняющих реакцию больных от ожидаемого. Для этого воспользуйтесь критерием χ^2 с поправкой Йейтса.
- Сравните полученное значение χ^2 с критическим, имеющим одну степень свободы.

ЗАДАЧИ

9.1. В исследовании Ф. Эшли и соавт., о котором мы уже говорили в задаче 8.8, сравнивали два средства для предупреждения образования зубного налета: хлоргексидин и хлорид аммония. Каждый из участников исследования в течение 48 часов полоскал рот одним из средств, после чего налет оценивали визуально. Через некоторое время опыт повторяли с другим средством (очередность определялась случайным образом). Были получены следующие результаты.

Хлорид аммония	Хлоргексидин
32	14
60	39
25	24
45	13
65	9
60	3
68	10
83	14
120	1
110	36

Эффективно ли полоскание хлоридом аммония?

9.2. В раннем детстве антибактериальную защиту (в частности, от стрептококков) обеспечивают антитела, полученные от матери. Если антител к стрептококкам вырабатывается у матери недостаточно, ребенок оказывается беззащитным перед этим микробом. В таких случаях беременным предлагали вводить пневмококковую вакцину: считалось, что благодаря сходству антигенной структуры пневмококка и стрептококка это позволит усилить выработку антител к стрептококкам.

К. Бейкер с соавт. (C. Baker et al. Influence of preimmunization antibody levels on the specificity of the immune response to related polysaccharide antigens. *N. Engl. J. Med.*, 303:173—178, 1980) ввели пневмококковую вакцину 20 женщинам и определили уровень антител к пневмококкам и стрептококкам до и после вакцинации. Вот что обнаружили исследователи.

Концентрация антител до и после вакцинации

Антитела к пневмококкам, мкг/мл		Антитела к стрептококкам, мкг/мл	
До вакцинации	После вакцинации	До вакцинации	После вакцинации
79	163	0,4	0,4
100	127	0,4	0,5
133	288	0,4	0,5
141	1154	0,4	0,9
43	666	0,5	0,5
63	156	0,5	0,5
127	644	0,5	0,5
140	273	0,5	0,5
145	231	0,5	0,5
217	1097	0,6	12,2
551	227	0,6	0,6
170	310	0,7	1,1
1049	1189	0,7	1,2
986	1695	0,8	0,8
436	1180	0,9	1,2
1132	1194	0,9	1,9
129	1186	1,0	2,0
228	444	1,0	0,9
135	2690	1,6	8,1
110	95	2,0	3,7

Оцените статистическую значимость изменения уровня антител к пневмококкам и стрептококкам.

9.3. Чему равна вероятность обнаружить не менее чем двукратное увеличение концентрации антител к пневмококкам и стрептококкам при уровне значимости 0,05? Графики чувствительности критерия Стьюдента, изображенные на рис. 6.9, применимы к парному критерию Стьюдента, если используемое в них n приравнять к *удвоенному* объему выборки.

9.4. Решите задачу 9.2 с помощью дисперсионного анализа повторных измерений. Как связаны между собой значения F и парного критерия Стьюдента?

9.5. При ишемической болезни сердца курение может вызвать приступ стенокардии. Это связано с тем, что никотин увеличивает потребность миокарда в кислороде, а окись углерода связывается с гемоглобином, тем самым снижая поступление кислорода. Однако не способствуют ли развитию приступов и другие компоненты табачного дыма? Чтобы выяснить это, У. Аронов (W. Aronow. Effect of non-nicotine cigarettes and carbon monoxide on angina. *Circulation*, 61:262—265, 1979) определил у 12 больных ишемической болезнью сердца продолжительность физической нагрузки до развития приступа стенокардии. У каждого больного опыт проводили до и после выкуривания пяти безникотиновых сигарет, а затем до и после вдыхания эквивалентного количества окиси углерода. Были получены следующие результаты.

Длительность нагрузки до развития приступа стенокардии, секунды

	Курение безникотиновых сигарет		Вдыхание окиси углерода	
Больной	До	После	До	После
1	289	155	281	177
2	203	117	186	125
3	359	187	372	238
4	243	134	254	165
5	232	135	219	153
6	210	119	225	148
7	251	145	264	180
8	246	121	237	144
9	224	136	212	152
10	239	124	250	147
11	220	118	209	138
12	211	107	226	141

Какие выводы позволяют сделать эти данные?

9.6. Определяя эффективность гидралазина, Л. Рубин и Р. Питер измеряли не только легочное сосудистое сопротивление, но и сердечный выброс. Результаты приведены в таблице.

	Измерение		
Больной	1	2	3
1	3,5	8,6	5,1
2	3,3	5,4	8,6
3	4,9	8,8	6,7
4	3,6	5,6	5,0

Менялся ли сердечный выброс?

9.7. Существует операция ушивания желудка для похудания. Уменьшенный желудок наполняется быстрее и чувство насыщения возникает при меньшем объеме съеденной пищи. Нельзя ли обойтись без операции и ограничиться сдавливанием живота надувным поясом? При оценке эффективности последнего метода А. Гелибтер и соавт. (A. Geliebter et al. Extraabdominal pressure alters food intake, intragastric pressure, and gastric emptying rate. *Am. J. Physiol.*, 250:R549—R552, 1986) наблюдали, какой объем пищи съедают добровольцы. Однако истинная цель исследования была скрыта. Участникам опыта объясняли, что по давлению внутри поясов измеряется увеличение живота во время еды и что исследователям нужно подобрать такое исходное давление, при котором измерения были бы наиболее точны. От участников требовалось есть до появления сытости. Вот каких показателей они достигли.

	Исходное давление в поясе, мм рт. ст.		
Участник	0	10	20
1	448	470	292
2	472	424	390
3	631	538	508
4	634	496	560
5	734	547	602
6	820	578	508
7	643	711	724

Что позволяют заключить эти данные?

9.8. По данным предыдущей задачи определите вероятность выявить снижение объема съеденной пищи на 100 мл при уровне значимости 5%.

9.9. У плода легкие не функционируют. Артериальный проток — сосуд, соединяющий аорту и легочную артерию, — позволяет крови, минуя легкие, попадать в плаценту, где и происходит газообмен. После рождения артериальный проток закрывается; если этого не происходит, то кровь, по-прежнему минуя легкие, не насыщается кислородом и не очищается от двуокиси углерода. Закрытию артериального протока способствует индометацин. Однако на результаты лечения влияет множество обстоятельств — гестационный возраст, возраст начала лечения, сопутствующие заболевания и их лечение. В таких случаях для оценки лечения можно применить следующий метод: найти пары детей с совпадающими значениями всех факторов, которые могут повлиять на результат терапии, затем случайным образом одному ребенку из пары назначить индометацин, а другому — плацебо. Предположим, такое исследование было проведено и дало следующий результат:

		Индометацин	
		Эффект есть	Эффекта нет
Плацебо	Эффект есть	65	13
	Эффекта нет	27	40

Эффективен ли индометацин?

9.10. Представим результаты исследования по-другому.

	Эффект есть	Эффекта нет
Индометацин	92	53
Плацебо	78	67

Какой вывод можно сделать по этим данным? Почему изменилось заключение по результатам того же исследования? Какой способ представления результатов лучше?

9.11. Просмотрите все статьи, опубликованные в доступном вам медицинском журнале за последний год. В скольких из них можно было бы применить дисперсионный анализ повторных измерений? В скольких из них он действительно использован? Какие методы использованы в остальных статьях? Совпали бы, по-вашему, их выводы с выводами дисперсионного анализа повторных измерений?

Глава 10

Непараметрические критерии

Для определения эффективности одного или нескольких методов лечения используется дисперсионный анализ, в частности критерий Стьюдента. Эти критерии основаны на допущении, что наблюдаемый признак подчиняется нормальному распределению. Более того, для применимости этих методов требуется, чтобы сравниваемые совокупности имели одинаковые дисперсии. Различными могут быть только значения средних. По их различию и судят о различии совокупностей. Применяя тот или иной метод, нужно быть уверенным, что допущения, на которых он основан, выполняются хотя бы приближенно. Иначе велик риск, что, выполнив, казалось бы, правильную последовательность действий, мы придем к ошибочным выводам.

Условия применимости дисперсионного анализа и критерия Стьюдента выполняются часто, но не всегда. В одних случаях слишком велика разница дисперсий, в других распределение далеко от нормального. Наконец, измеряемый признак может оказаться нечисловым или «не вполне числовым». В такой ситуации

следует воспользоваться *непараметрическими методами*. Один из таких критериев знаком нам по гл. 5 — это критерий χ^2, другой пример — критерий Мак-Нимара (гл. 9). Теперь мы займемся непараметрическими критериями, основанными на рангах.

Ранее мы уже встречались с порядковыми признаками. Природа порядковых признаков такова, что о двух значениях можно сказать лишь, какое больше или меньше, но в принципе нельзя — на сколько или во сколько раз. (Любой количественный признак можно рассматривать как порядковый, но не наоборот.) Первое, что следует сделать при анализе таких признаков, это перейти к их рангам — номерам, под которыми будут стоять исходные данные, если выстроить их по возрастанию. Критерии, основанные на рангах, не нуждаются в предположениях о типе распределения. Единственное требование состоит в том, чтобы тип распределения в сравниваемых совокупностях был одинаковым. При этом не нужно знать, что это за распределение и каковы его параметры.

Мы начнем с аналогов критерия Стьюдента — критерия суммы рангов Манна—Уитни и критерия Уилкоксона. Затем будет изложен критерий Крускала—Уоллиса — аналог дисперсионного анализа и критерий Фридмана — аналог дисперсионного анализа повторных измерений.

ПАРАМЕТРИЧЕСКИЕ И НЕПАРАМЕТРИЧЕСКИЕ МЕТОДЫ. КАКОЙ ВЫБРАТЬ?

Математическая модель, которая используется при построении дисперсионного анализа, предполагает нормальное распределение. Вспомним жителей маленького городка, которых мучили диетами, якобы влияющими на сердечный выброс (гл. 3), и мужественных добровольцев, принимавших совершенно неэффективный диуретик (гл. 4), — все это были выборки из нормально распределенной совокупности. Поэтому критические значения F и t, которые мы нашли в этих главах, дадут правильное представление о статистической значимости различий только в случае, если выборки извлечены именно из такой совокупности.

Параметрические методы, как видно уже из их названия, опе-

рируют параметрами распределения. В частности, дисперсионный анализ и его частный случай, критерий Стьюдента, основаны на сравнении средних и дисперсий. Но эти параметры правильно описывают только нормально распределенную совокупность. Если распределение далеко от нормального, среднее и дисперсия дадут о нем неверное представление. Столь же неверными окажутся и критерии, основанные на этих параметрах.

В гл. 2 мы изучали рост юпитериан (см. рис. 2.3А). Средний рост составил 37,6 см, а стандартное отклонение 4,5 см. На рис. 2.3Б изображено, как выглядело бы нормальное распределение с такими параметрами. Оно мало похоже на распределение, наблюдаемое в действительности. Если бы распределение роста юпитериан было нормальным, рост большинства из них оказался бы в пределах 37—38 см и рост практически всех — в интервале от 26 до 49 см. Однако картина иная. Рост большинства юпитериан группируется вокруг 35 см, то есть ниже среднего. При этом интервал, охватывающий все значения роста (от 31 до 52 см), смещен вправо, то есть распределение асимметрично.

Непараметрические методы, которые мы рассмотрим в этой главе, заменяют реальные значения признака рангами. При этом мы сохраняем большую часть информации о распределении, но избавляемся от необходимости знать, что это за распределение. Нас не интересуют более параметры распределения, отпадает и необходимость равенства дисперсий. Остается в силе только предположение, что тип распределения во всех случаях одинаков*.

Если выполняется условие нормальности распределения, параметрические критерии обеспечивают наибольшую чувствительность. Если же это условие не выполняется хотя бы приблизительно, их чувствительность существенно снижается и непараметрические критерии дают больше шансов выявить реально существующие различия. Что будет, если применить непараметрический критерий при нормальном распределении? Чувствительность критериев, которые мы рассмотрим в этой главе, составляет в этом случае примерно 95% от чувствительности их па-

* Кроме того, теоретически распределение должно быть непрерывным. При практическом применении непараметрических критериев этим условием можно пренебречь.

раметрических аналогов (это обстоятельство можно использовать для оценки чувствительности непараметрических критериев и определения необходимого числа наблюдений).

Как выяснить, согласуются ли данные с предположением о нормальности распределения? Простейший способ состоит в том, чтобы нанести их на график, подобный тем, которые мы рисовали, изучая рост инопланетян в гл. 2. Нарисовав график, прикиньте, похож ли он на нормальное распределение. Та ли у него форма, достаточно ли он симметричен относительно среднего, покрывает ли интервал, равный плюс-минус двум стандартным отклонениям от среднего, практически все наблюдения? Сравните графики для разных групп. Близок ли разброс значений? Ответив на все вопросы утвердительно, воспользуйтесь параметрическим критерием. В противном случае следует использовать непараметрический критерий. Изложенный нехитрый прием почти наверняка поможет правильно выбрать тип критерия.

Для тех, кто не привык полагаться на зрительные впечатления, укажем еще два способа, иногда более точные и всегда более трудоемкие. Первый основан на использовании *нормальной вероятностной бумаги*. Вы легко поймете, о чем идет речь, если когда-нибудь видели логарифмическую бумагу. Вся разница в том, что на логарифмической бумаге вертикальная ось проградуирована так, чтобы графиком экспоненты была прямая, а на нормальной вероятностной бумаге прямой окажется функция нормального распределения. На такую бумагу определенным образом наносят имеющиеся значения. Если они расположатся почти на одной прямой, можно применять параметрические методы. Второй способ опирается на критерий χ^2. Он позволяет сравнить реальные данные с теми, которые дало бы нормальное распределение, имеющее то же среднее и дисперсию. Мы не будем останавливаться на этих процедурах*, поскольку их выводы наверняка совпадут с теми, что даст простая прикидка.

Как правило, основная трудность состоит не в том, какой из

---

* Желающие могут познакомиться с ними по книгам J. H. Zar. Biostatstical analisys. 2nd ed. Prentice-Hall, Englewood Cliffs, N. J., 1984 и W. J. Dixon, F. J. Massey, Jr. Introduction to statistical analisys. 4th ed., McGraw-Hill, New York, 1983.

перечисленных способов выбрать, а в том, что объем выборки слишком мал, чтобы применить любой из них. Убедительные свидетельства в пользу гипотезы нормальности или против нее встречаются редко. Гораздо чаще все решают интуиция, привычка и вкус исследователя. Существуют две точки зрения на то, как следует поступать в таких случаях. Согласно одной, в отсутствие очевидных противоречий между данными и гипотезой их нормального распределения следует применить параметрический метод. Согласно другой, если нет явного подтверждения гипотезы нормальности распределения, лучше воспользоваться непараметрическим методом. Сторонники первой точки зрения упирают на то, что параметрические методы более чувствительны и более известны. Приверженцы второй резонно замечают, что исследователь не должен исходить из предположений, которые нельзя проверить, и что, применяя непараметрические критерии, мы почти ничем не рискуем — ведь даже в случае нормального распределения их чувствительность не намного ниже чувствительности параметрических. Ни одна из сторон пока не одержала верх, и похоже, этого не произойдет никогда.

СРАВНЕНИЕ ДВУХ ВЫБОРОК: КРИТЕРИЙ МАННА—УИТНИ

Напомним схему, по которой строились все параметрические методы, будь то критерий Стьюдента, дисперсионный или корреляционный анализ. Из нормально распределенной совокупности мы извлекали все возможные выборки определенного объема и строили распределение значений соответствующего критерия. Теперь, упорядочив значения признака и перейдя от реальных значений к рангам, мы поступим несколько иначе. Мы просто перечислим все возможные варианты упорядочивания двух групп.

Как это сделать, мы покажем на простом примере. Чтобы вариантов упорядочивания было не слишком много, рассмотрим опыт с участием 7 добровольцев. Из них 3 принимают плацебо (контрольная группа), а 4 препарат, предположительно диуретик (экспериментальная группа). В табл. 10.1 приведены данные о суточном диурезе. Против каждого значения диуреза указан

Таблица 10.1. Эксперимент с диуретиком

Плацебо (контрольная группа)		Препарат (экспериментальная группа)	
Суточный диурез, мл	Ранг	Суточный диурез, мл	Ранг
1000	1	1400	6
1380	5	1600	7
1200	3	1180	2
		1220	4
	$T=9$		

его ранг — место в общем упорядоченном ряду. Рангом наименьшей величины будет 1; ранг наибольшей величины равен числу наблюдений, то есть 7. Если препарат увеличивает диурез, то ранги в экспериментальной группе должны быть больше, чем в контрольной. Мерой отличия изберем сумму рангов в меньшей из групп и обозначим ее T. В нашем примере меньшая группа — контрольная. Соответствующее значение T равно 9.

Достаточно ли мало значение T, чтобы отклонить гипотезу об отсутствии действия препарата?

Для ответа на этот вопрос рассмотрим совокупность всех возможных перестановок. Заметьте, после перехода к рангам нам уже не нужно рассматривать сами исходные величины и совокупность их возможных значений. Поэтому наши дальнейшие рассуждения полностью применимы к любым двум группам наблюдений по 3 и 4 наблюдения в каждой.

Итак, нулевая гипотеза — гипотеза об отсутствии влияния препарата на диурез. Если она справедлива, любой ранг может равновероятно оказаться в любой из групп. Чтобы узнать, велика ли вероятность случайно получить перестановку из табл. 10.1, рассмотрим все возможные перестановки. Понятно, что распределить ранги по двум группам — это то же самое, что набрать ранги для одной из групп (оставшиеся автоматически попадут во вторую). Тогда, перечислив все варианты выбора 3 рангов из 7, мы тем самым перечислим все варианты распределения семи рангов по двум группам. Число способов по-разному выбрать 3 ранга из 7 равно 35. Все 35 вариантов приведены в табл. 10.2. Крестиком помечены ранги, попадающие в контрольную группу. В правом

Таблица 10.2. Варианты разделения 7 рангов на две группы по 3 и 4 ранга

Ранги							Сумма рангов
1	2	3	4	5	6	7	
x	x	x					6
x	x		x				7
x	x			x			8
x	x				x		9
x	x					x	10
x		x	x				8
x		x		x			9
x		x			x		10
x		x				x	11
x			x	x			10
x			x		x		11
x			x			x	12
x				x	x		12
x				x		x	13
x					x	x	14
	x	x	x				9
	x	x		x			10
	x	x			x		11
	x	x				x	12
	x		x	x			11
	x		x		x		12
	x		x			x	13
	x			x	x		13
	x			x		x	14
	x				x	x	15
		x	x	x			12
		x	x		x		13
		x	x			x	14
		x		x	x		14
		x		x		x	15
		x			x	x	16
			x	x	x		15
			x	x		x	16
			x		x	x	17
				x	x	x	18

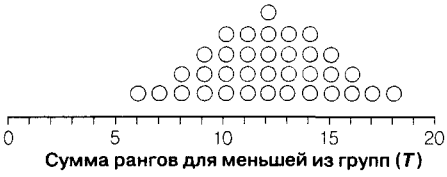


Рис. 10.1. 35 возможных сумм рангов для меньшей из групп (см. табл. 10.2).

столбце для каждого из вариантов указана величина T — сумма рангов меньшей (контрольной) группы. Если нанести значения T на график, получится распределение, показанное на рис. 10.1. Если справедлива нулевая гипотеза, то все сочетания рангов равновероятны. Это значит, что если, например, $T = 12$ в 5 вариантах из 35, то вероятность случайно получить значение $T = 12$ равна 5/35. Таким образом, на рис. 10.1 изображено распределение значений T в случае справедливости нулевой гипотезы об отсутствии действия препарата. По форме оно напоминает распределение t (рис. 4.5). Однако есть и отличия. Действительно, распределение t непрерывно. Оно построено по бесконечной совокупности значений, вычисленных для бесконечного числа выборок из бесконечной нормально распределенной совокупности. Напротив, распределение T конечно и дискретно, то есть имеет ступенчатый вид, принимая значения лишь в конечном числе целочисленных точек.

Глядя на рис. 10.1, легко определить вероятность получить то или иное значение T при условии справедливости нулевой гипотезы. Например, значения $T = 9$ и $T = 15$ наблюдаются в 3 вариантах, то есть вероятность появления каждой из этих сумм равна 3/15. Вероятность получить значение T, равное 8 или 16, составляет $2/35 = 0{,}057$. Будем считать эти значения T критическими. В нашем опыте $T = 9$, так что нулевую гипотезу отвергнуть мы не можем.

Уровень значимости обычно принимают равным 5% или 1%. Можно ли установить такой уровень в нашем примере? Оказывается, нет. У нас есть всего 13 разных значений T, поэтому уровень значимости может меняться только скачками. Назвав произвольный уровень значимости α, мы скорее всего обнаружим, что нет такого значения T, которому бы он соответствовал. В качестве критического берут то значение T, которому соответст-

вует уровень значимости, наиболее близкий к 1 или 5%. В нашем примере ближе всего к 5% находится уровень значимости 5,7%, соответствующий $T = 8$.

Критические значения критерия Манна—Уитни приведены в табл. 10.3. Столбец критических значений содержит пары чисел. Различия статистически значимы, если T не больше первого из них или не меньше второго. Например, когда в одной группе 3 человека, а в другой 6, различия статистически значимы, если $T \le 7$ или $T \ge 23$.

Изложенный вариант критерия известен как T-критерий Манна—Уитни*. Порядок его вычисления таков.

- Данные обеих групп объединяют и упорядочивают по возрастанию. Ранг 1 присваивают наименьшему из всех значений, ранг 2 — следующему и так далее. Наибольший ранг присваивают самому большому среди значений в обеих группах. Если значения совпадают, им присваивают один и тот же средний ранг (например, если два значения поделили 3-е и 4-е места, обоим присваивают ранг 3,5).
- Для меньшей группы вычисляют T — сумму рангов ее членов. Если численность групп одинакова, T можно вычислить для любой из них.
- Полученное значение T сравнивают с критическими значениями. Если T меньше или равно первому из них либо больше или равно второму, то нулевая гипотеза отвергается (различия статистически значимы).

Что делать, если нужной численности групп в таблице не оказалось? Можно самому построить распределение T. К сожалению, с ростом численности групп сделать это становится все труднее. Например, если объем каждой из групп равен 10, то

---

* Существует еще U-критерий Манна—Уитни, в котором вместо T вычисляют U, при этом $U = T - n_м(n_м + 1)/2$, где $n_м$ — численность меньшей из групп. Об этом варианте критерия можно прочесть в книге S. Siegel, N. J. Castellan. Nonparametric Statistics for the Behavioral Sciences, 2nd ed. McGraw-Hill, N. Y., 1988. Подробный вывод Т-критерия и его связь с U-критерием приведены в книге F. Mosteller, R. Rourke. Sturdy Statistics: Nonparametrics and Order Statistics, Addison-Wesley, Reading, Mass., 1973.

Таблица 10.3. Критические значения критерия Манна—Уитни (двусторонний вариант)

Численность группы		Приблизительный уровень значимости α					
		0,05			0,01		
меньшей	большей	Критические значения		Точное значение α	Критические значения		Точное значение α
3	4	6	18	0,057			
	5	6	21	0,036			
	5	7	20	0,071			
	6	7	23	0,048	6	24	0,024
	7	7	26	0,033	6	27	0,017
	7	8	25	0,067			
	8	8	28	0,042	6	30	0,012
4	4	11	25	0,057	10	26	0,026
	5	11	29	0,032	10	30	0,016
	5	12	28	0,063			
	6	12	32	0,038	10	34	0,010
	7	13	35	0,042	10	38	0,012
	8	14	38	0,048	11	41	0,008
	8				12	40	0,016
5	5	17	38	0,032	15	40	0,008
	5	18	37	0,056	16	39	0,016
	6	19	41	0,052	16	44	0,010
	7	20	45	0,048	17	48	0,010
	8	21	49	0,045	18	52	0,011
6	6	26	52	0,041	23	55	0,009
	6				24	54	0,015
	7	28	56	0,051	24	60	0,008
	7				25	59	0,014
	8	29	61	0,043	25	65	0,008
	8	30	60	0,059	26	64	0,013
7	7	37	68	0,053	33	72	0,011
	8	39	73	0,054	34	78	0,009
8	8	49	87	0,050	44	92	0,010

число вариантов равно 184756. Поэтому лучше воспользоваться тем, что при численности групп, большей 8, распределение T приближается к нормальному со средним

$$\mu_T = \frac{n_м (n_м + n_б + 1)}{2}$$

и стандартным отклонением

$$\sigma_T = \sqrt{\frac{n_м n_б (n_м + n_б + 1)}{12}},$$

где $n_м$ и $n_б$ — объемы меньшей и большей выборок*.

В таком случае величина

$$z_T = \frac{T - \mu_T}{\sigma_T}$$

имеет стандартное нормальное распределение. Это позволяет сравнить z_T с критическими значениями нормального распределения (последняя строка табл. 4.1). Более точный результат обеспечивает поправка Йейтса:

$$z_T = \frac{|T - \mu_T| - \frac{1}{2}}{\sigma_T}.$$

Роды по Лебуайе

В последние десятилетия произошел коренной пересмотр взглядов на родовспоможение. Акушерская революция совершалась под лозунгом «Отец вместо седативных средств». Восторжество-

---

* Если некоторые значения совпадают, стандартное отклонение должно быть уменьшено согласно формуле:

$$\sigma_T = \sqrt{\frac{n_м n_б (N+1)}{12} - \frac{n_м n_б}{12N(N^2-1)} \sum (\tau_i - 1)\tau_i (\tau_i + 1)},$$

где $N = n_м + n_б$ — общее число членов обеих выборок, τ_i — число значений i-го ранга, а суммирование производится по всем совпадающим рангам.

вала точка зрения, согласно которой при нормальных родах следует прибегать к помощи психологических, а не лекарственных средств. Что делать конкретно, мнения расходились. Масла в огонь подлила книга Лебуайе «Рождение без насилия». Французский врач предлагал комплекс мер, призванных свести к минимуму потрясение, которое испытывает новорожденный при появлении на свет. Роды надлежит принимать в тихом затемненном помещении. Сразу после родов ребенка следует уложить на живот матери и не перерезать пуповину, пока та не перестанет пульсировать. Затем, успокаивая младенца легким поглаживанием, нужно поместить его в теплую ванну, чтобы «внушить, что разрыв с организмом матери — не шок, но удовольствие». Лебуайе указывал, что дети, рожденные по его методике, здоровее и радостнее других. Многие врачи считали, что предложенная методика не только противоречит общепринятой практике, но и создает дополнительную опасность для матери и ребенка. Тем не менее у Лебуайе нашлись и сторонники.

Как часто бывает в медицине, отсутствие достоверных данных могло затянуть спор на многие годы. Пока Н. Нелсон и соавт.* не провели клиническое испытание, материалы ограничивались «клиническим опытом» автора методики.

В эксперименте Нелсон, проведенном в клинике канадского университета Макмастер, участвовали роженицы без показаний к искусственному родоразрешению, срок беременности которых составлял не менее 36 недель и которые были согласны рожать как по обычной методике, так и по Лебуайе. Роженицы были случайным образом разделены на две группы. В контрольной роды проводились по общепринятой методике в нормально освещенном помещении с обычным уровнем шума; после рождения пуповина немедленно перерезалась, ребенка пеленали и отдавали матери. В экспериментальной группе роды принимались по методике Лебуайе. В обеих группах при родах присутствовали мужья, применение обезболивающих средств было ми-

* N. Nelson, M. Enkin, S. Saigal, K. Bennett, R. Milner, D. Sackett. A randomized clinical trial of the Leboyer approach to childbirth. *N. Engl. J. Med.*, 302: 655—660, 1980.

нимальным. Тем самым, группы различались только в том, в чем методика Лебуайе не совпадает с общепринятой.

То, в какую группу попала роженица, было известно самой роженице и всем, кто присутствовал при родах. На этом этапе эффект плацебо исключить было невозможно. Однако уже на этапе послеродового наблюдения одна из сторон, а именно врачи, которые оценивали состояние ребенка, не знали, по какой методике происходили роды. Таким образом исследование Нелсон было *простым слепым:* условия знала только одна из сторон, наблюдателю же они были неизвестны.

Для оценки развития детей была разработана специальная шкала. Из числа детей, рожденных по обычной методике, оценку «отлично» по этой шкале получали примерно 30%. Изучив труды Лебуайе, Нелсон и соавт. пришли к выводу, что предлагаемый метод, судя по заявлениям автора, гарантирует оценку «отлично» у 90% детей. Приняв уровень значимости $\alpha = 0{,}05$, исследователи рассчитали, что для обеспечения 90% вероятности выявить такие различия в каждой из групп должно быть по 20 детей.

Работа продолжалась целый год. За это время исследователи провели беседы с 187 потенциальными участницами, разъясняя им смысл предстоящего эксперимента. 34 женщины не подошли по состоянию здоровья, 97 отказались участвовать в эксперименте (из них 70 собирались рожать только по методике Лебуайе). Из оставшихся 56 женщин одна успела родить до рандомизации. В результате число участниц сократилось до 55. Их и разделили случайным образом на две группы. После того как из исследования выбыла одна из попавших в контрольную группу, в этой группе оказалось 26, а в экспериментальной 28 рожениц. Однако у 6 женщин в контрольной группе и у 8 в экспериментальной возникли осложнения, и их пришлось исключить из участия в эксперименте. В итоге в каждой из групп оказалось по 20 женщин. Вы видите, насколько трудно обеспечить достаточную численность групп даже в простом исследовании*.

Оценка по шкале развития производилось сразу после родов,

---

* D. Sackett, M. Gent. Controversy in counting and attributing events in clinical trials. *N. Engl. J. Med.*, 301:1410—1412, 1979.

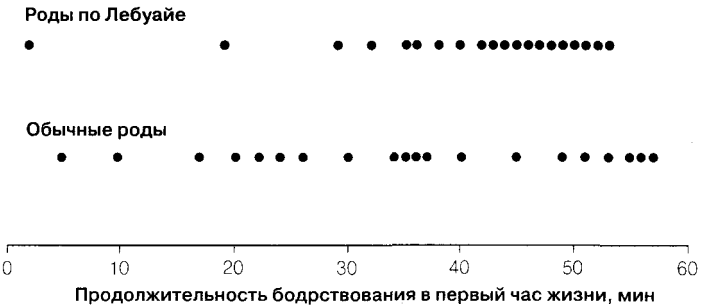


Рис. 10.2. Продолжительность бодрствования в первый час жизни после обычных родов и родов по Лебуайе. Обратите внимание, что в обеих группах распределение асимметрично — преобладают высокие значения.

а также спустя несколько месяцев. Мы остановимся на одном из показателей — времени бодрствования в первый час жизни. Предполагалось, что чем лучше состояние новорожденного, тем более он активен. Значит, у младенцев, рожденных по Лебуайе, время бодрствования должно быть продолжительнее, чем у рожденных по обычной методике.

Из рис. 10.2 видно, что данные не подчиняются нормальному распределению. Особенно это заметно в экспериментальной группе. Тем самым, параметрические методы, например критерий Стьюдента, к этим данным неприменимы. Поэтому воспользуемся непараметрическим критерием Манна—Уитни.

Объединим данные, относящиеся к обеим группам, и упорядочим их по возрастанию. В табл. 10.4 кроме суммарного времени бодрствования указан также его ранг. Поскольку численность групп одинакова, сумму рангов T можно вычислить для любой из них. Подсчитаем T для контрольной группы. Она равна 374. Размер групп достаточен, чтобы воспользоваться нормальным приближением для T. Поэтому перейдем от T к z_T. Итак, полагая истинной нулевую гипотезу, вычисляем среднее всех возможных значений T

$$\mu_T = \frac{n_м(n_м + n_б + 1)}{2} = \frac{20(20 + 20 + 1)}{2} = 410$$

Таблица 10.4. Продолжительность бодрствования в первый час жизни, мин

Роды по обычной методике	Ранг	Роды по Лебуайе	Ранг
5,0	2	2,0	1
10,1	3	19,0	5
17,7	4	29,7	10
20,3	6	32,1	12
22,0	7	35,4	15
24,9	8	36,7	17
26,5	9	38,5	19
30,8	11	40,2	20
34,2	13	42,1	22
35,0	14	43,0	23
36,6	16	44,4	24
37,9	18	45,6	26
40,4	21	46,7	27
45,5	25	47,1	28
49,3	31	48,0	29
51,1	33	49,0	30
53,1	36	50,9	32
55,0	38	51,2	34
56,7	39	52,5	35
58,0	40	53,3	37
	$T=374$		

и стандартное отклонение

$$\sigma_T = \sqrt{\frac{n_м n_б (n_м + n_б + 1)}{12}} = \sqrt{\frac{20 \times 20 \times 41}{12}} = 36{,}97.$$

Таким образом, с учетом поправки Йейтса,

$$z_T = \frac{|T - \mu_T| - \frac{1}{2}}{\sigma_T} = \frac{|374 - 410| - \frac{1}{2}}{36{,}9} = 0{,}962.$$

В табл. 4.1 находим 5% критическое значение для бесконеч-

ного числа степеней свободы. Найденное критическое значение равно 1,960, то есть больше полученного. Тем самым, имеющиеся данные не позволяют отклонить гипотезу о том, что младенцы, рожденные по методике Лебуайе, по своей активности ничем не отличаются от остальных.

Общая оценка развития также не показала существенной разницы между двумя группами детей. Исследование Нелсон и соавт. — пример тщательно спланированного и проведенного клинического испытания. На четко поставленный вопрос был получен ответ. Сегодня мало кто помнит о родах по Лебуайе. Не беда — на смену идут роды под водой. Оценка их влияния на развитие ребенка, быть может, станет темой будущих исследований.

СРАВНЕНИЕ НАБЛЮДЕНИЙ ДО И ПОСЛЕ ЛЕЧЕНИЯ: КРИТЕРИЙ УИЛКОКСОНА

В гл. 9 было описано использование парного критерия Стьюдента для сравнения состояния больных до и после лечения. Однако для применения этого критерия необходимо, чтобы изменения имели нормальное распределение. Существует критерий, основанный на рангах, не ограниченный этим условием, — это критерий Уилкоксона. Принцип критерия следующий. Для каждого больного вычисляют величину изменения признака. Все изменения упорядочивают по абсолютной величине (без учета знака). Затем рангам приписывают знак изменения и суммируют эти «знаковые ранги» — в результате получается значение критерия Уилкоксона W.

Как видим, используется информация об абсолютной величине изменения и его знаке (то есть уменьшении или увеличении наблюдаемого признака). Метод основан на рангах, поэтому не нуждается в предположениях о типе распределения изменений. Как в случае с критерием Манна—Уитни, здесь также можно перечислить все возможные величины W и найти критическое значение.

Обратите внимание, исходно ранги присваиваются в соответствии с абсолютной величиной изменения. Так, например,

Таблица 10.5. Действие диуретика

Участник	Суточный диурез, мл		Величина изменения	Ранг изменения	Знаковый ранг
	До приема	После приема			
1	1490	1600	110	5	5
2	1300	1850	550	6	6
3	1400	1300	−100	4	−4
4	1410	1500	90	3	3
5	1350	1400	50	2	2
6	1000	1010	10	1	1
					$W = 13$

величины 5,32 и −5,32 получат один и тот же ранг, а уже затем рангам будет присвоен знак изменения.

Рассмотрим пример. Допустим, мы исследуем некий препарат, предположительно диуретик. Дадим его 6 добровольцам и сравним диурез до и после приема препарата. Результаты представлены в табл. 10.5.

У 5 человек диурез увеличился. Значит ли это, что препарат является диуретиком?

Упорядочим изменения диуреза по абсолютной величине и присвоим им ранги от 1 до 6. Затем, приписав рангу каждого изменения соответствующий изменению знак, перейдем к знаковым рангам (последний столбец табл. 10.5). Наконец, вычислим сумму знаковых рангов $W = 13$.

Если препарат не оказывает действия, сумма рангов со знаком «+» должна быть примерно равна сумме рангов со знаком «–» и значение W окажется близким нулю. Напротив, если препарат увеличивает (или уменьшает) диурез, будут преобладать положительные (отрицательные) ранги и значение W будет отличным от нуля.

Чтобы найти критическое значение W, выпишем все 64 возможных исхода опыта (табл. 10.6 и рис. 10.3). В четырех случаях значение W по абсолютной величине равно или превосходит 19. Таким образом, отвергая нулевую гипотезу при $|W| > 19$, мы обеспечим уровень значимости 4/64 = 0,0625. Изменение диуреза в нашем опыте надо признать статистически не значимым:

Таблица 10.6. Возможные сочетания знаковых рангов для 6 пар измерений

Ранги						Сумма знаковых рангов
1	2	3	4	5	6	
–	–	–	–	–	–	–21
+	–	–	–	–	–	–19
–	+	–	–	–	–	–17
–	–	+	–	–	–	–15
–	–	–	+	–	–	–13
–	–	–	–	+	–	–11
–	–	–	–	–	+	–9
+	+	–	–	–	–	–15
+	–	+	–	–	–	–13
+	–	–	+	–	–	–11
+	–	–	–	+	–	–9
+	–	–	–	–	+	–7
–	+	+	–	–	–	–11
–	+	–	+	–	–	–9
–	+	–	–	+	–	–7
–	+	–	–	–	+	–5
–	–	+	+	–	–	–7
–	–	+	–	+	–	–5
–	–	+	–	–	+	–3
–	–	–	+	+	–	–3
–	–	–	+	–	+	–1
–	–	–	–	+	+	1
+	+	+	–	–	–	–9
+	+	–	+	–	–	–7
+	+	–	–	+	–	–5
+	+	–	–	–	+	–3
+	–	+	+	–	–	–5
+	–	+	–	+	–	–3
+	–	+	–	–	+	–1
+	–	–	+	+	–	–1
+	–	–	+	–	+	1
+	–	–	–	+	+	3

Таблица 10.6. Окончание

Ранги						Сумма знаковых рангов
1	2	3	4	5	6	
–	+	+	+	–	–	–3
–	+	+	–	+	–	–1
–	+	+	–	–	+	1
–	+	–	+	+	–	1
–	+	–	+	–	+	3
–	+	–	–	+	+	5
–	–	+	+	+	–	3
–	–	+	+	–	+	5
–	–	+	–	+	+	7
–	–	–	+	+	+	9
+	+	+	+	–	–	–1
+	+	+	–	+	–	1
+	+	+	–	–	+	3
+	+	–	+	+	–	3
+	+	–	+	–	+	5
+	+	–	–	+	+	7
+	–	+	+	+	–	5
+	–	+	+	–	+	7
+	–	+	–	+	+	9
+	–	–	+	+	+	11
–	+	+	+	+	–	7
–	+	+	+	–	+	9
–	+	+	–	+	+	11
–	+	–	+	+	+	13
–	–	+	+	+	+	15
+	+	+	+	+	–	9
+	+	+	+	–	+	11
+	+	+	–	+	+	13
+	+	–	+	+	+	15
+	–	+	+	+	+	17
–	+	+	+	+	+	19
+	+	+	+	+	+	21

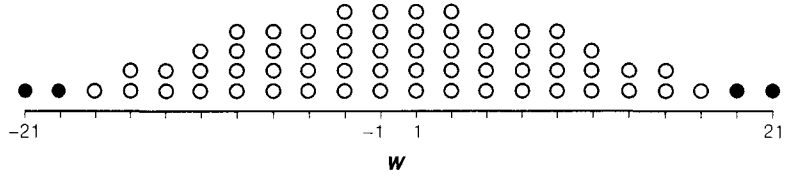


Рис. 10.3. 64 возможные суммы рангов для группы из 6 человек (см. табл. 10.6). 4 наибольших по абсолютной величине значения помечены черным.

$P < 0{,}0625$. На самом деле в таблице имеется 14 значений W, по абсолютной величине не меньших 13. Поскольку 14/64 = 0,219, мы могли бы записать $P < 14/64$.

Как и в случае критерия Манна—Уитни, распределение W не является непрерывным и поэтому нельзя указать критическое значение, для которого уровень значимости в точности равнялся бы, например, 5%. В табл. 10.7 приведены критические значения, наиболее близкие к 5 и 1% уровням значимости для случая, когда численность группы не превосходит 20.

Если число пар измерений больше 20, то распределение W достаточно близко к нормальному со средним $\mu_W = 0$ и стандартным отклонением

$$\sigma_W = \sqrt{\frac{n(n+1)(2n+1)}{6}},$$

где n — число пар наблюдений (то есть численность группы).

Можно, таким образом, использовать

$$z_W = \frac{W - \mu_W}{\sigma_W} = \frac{W}{\sqrt{\frac{n(n+1)(2n+1)}{6}}}.$$

Чтобы приближение было более точным, воспользуемся поправкой Йейтса на непрерывность:

$$z_W = \frac{|W| - \frac{1}{2}}{\sqrt{\frac{n(n+1)(2n+1)}{6}}}.$$

Таблица 10.7. Критические значения W (двусторонний вариант)

n	W	P	n	W	P
5	15	0,062	13	65	0,022
6	21	0,032		57	0,048
	19	0,062	14	73	0,020
7	28	0,016		63	0,050
	24	0,046	15	80	0,022
8	32	0,024		70	0,048
	28	0,054	16	88	0,022
9	39	0,020		76	0,050
	33	0,054	17	97	0,020
10	45	0,020		83	0,050
	39	0,048	18	105	0,020
11	52	0,018		91	0,048
	44	0,054	19	114	0,020
12	58	0,020		98	0,050
	50	0,052	20	124	0,020
				106	0,048

F. Mosteller and R. Rourke. Sturdy statistics: nonparametrics and order statistics, Addison-Wesley, Reading, Mass., 1973.

При анализе наблюдений до—после встречается два вида совпадений. Это, во-первых, совпадение величин, которым присваиваются ранги. Такая ситуация возникает при использовании любого рангового метода, будь то критерий Манна—Уитни или коэффициент корреляции Спирмена. Как всегда, совпадающим величинам присваивается общий ранг, равный среднему мест, занимаемых ими в упорядоченном наборе*.

Единственная особенность — то, что в случае наблюдений (до—после) речь идет о совпадении не самих величин наблюдае-

---

* Если некоторые значения совпадают, стандартное отклонение должно быть уменьшено в соответствии со следующей формулой:

$$\sigma_W = \sqrt{\frac{n(n+1)(2n+1)}{6} - \sum \frac{(\tau_i - 1)\tau_i(\tau_i + 1)}{12}},$$

где n — численность группы, τ_i - число значений i-го ранга.

мого признака, а их изменений. Другой вид совпадения — совпадение значений «до» и «после». Каждую такую пару наблюдений нужно исключать из расчета, соответственно уменьшая на единицу объем выборки.

Повторим последовательность шагов, позволяющую по наблюдениям, выполненным до и после лечения, проверить его эффективность.

- Вычислите величины изменений наблюдаемого признака. Отбросьте пары наблюдений, которым соответствует нулевое изменение.
- Упорядочите изменения по возрастанию их абсолютной величины и присвойте соответствующие ранги. Рангами одинаковых величин назначьте средние тех мест, которые они делят в упорядоченном ряду.
- Присвойте каждому рангу знак в соответствии с направлением изменения: если значение увеличилось — «+», если уменьшилось — «–».
- Вычислите сумму знаковых рангов W*.
- Сравните полученную величину W с критическим значением. Если она больше критического значения, изменение показателя статистически значимо.

А теперь применим критерий Уилкоксона к анализу рассмотренного в гл. 9 эксперимента Левина.

Курение и функция тромбоцитов

В гл. 9 мы разобрали исследование Левина, посвященное влиянию курения на функцию тромбоцитов. В частности, на рис. 9.2 приведены результаты опыта с выкуриванием сигареты: агрегация тромбоцитов до и после этого вредоносного воздействия. Рассмотрим еще раз эти данные (табл. 10.8). Обратим внимание на 4-й столбец: здесь показана величина изменения интересу-

---

* Существует вариант критерия Уилкоксона, в котором суммируют только положительные или только отрицательные знаковые ранги. На выводе это никак не сказывается, однако значение W, естественно, получается другим. Поэтому важно знать, на какой вариант критерия рассчитана имеющаяся в вашем распоряжении таблица критических значений.

Таблица 10.8. Агрегация тромбоцитов до и после выкуривания сигареты

Участник	До курения	После курения	Изменение	Ранг изменения	Знаковый ранг изменения
1	25	27	2	2	2
2	25	29	4	3,5	3,5
3	27	37	10	6	6
4	44	56	12	7	7
5	30	46	16	10	10
6	67	82	15	8,5	8,5
7	53	57	4	3,5	3,5
8	53	80	27	11	11
9	52	61	9	5	5
10	60	59	−1	1	−1
11	28	43	15	8,5	8,5
					W = 64

Рис. 10.4. Изменение агрегации тромбоцитов после выкуривания сигареты. Вряд ли мы имеем дело с нормальным распределением, об этом свидетельствует, в частности, «выпадающее» значение 27%. В таких случаях непараметрические методы, например критерий Уилкоксона, предпочтительнее параметрических, таких, как критерий Стьюдента.

ющего нас показателя. Можно ли считать распределение изменения нормальным? При большом желании да, но следует все же признать, что для суждения о типе распределения данных слишком мало. Смущает и «выскакивающее» значение 27% — оно наводит на мысль о возможной асимметрии распределения. В подобных случаях лучше не рисковать и воспользоваться непараметрическим критерием. Применим критерий Уилкоксона.

Выпишем абсолютные величины изменений в порядке возрастания. Полученные ранги приведены в пятом столбце табл. 10.8, а шестой столбец содержит те же ранги, но со знаками, соответствующими направлению изменения. Сумма знаковых

рангов $W = 2 + 3{,}5 + 6 + 7 + 10 + 8{,}5 + 3{,}5 + 11 + 5 + (-1) + 8{,}5 = 64$. В табл. 10.7 находим 1,8% критическое значение для суммы рангов. Оно равно 52, то есть меньше полученного нами. Поэтому мы признаем изменение агрегации тромбоцитов статистически значимым ($P < 0{,}018$).

СРАВНЕНИЕ НЕСКОЛЬКИХ ГРУПП: КРИТЕРИЙ КРУСКАЛА—УОЛЛИСА

В гл. 3 была рассмотрена задача сравнения нескольких выборок. Эта задача возникает, например, когда нужно определить, одинаково ли эффективны несколько методов лечения, каждый из которых испытывается на отдельной группе. Предполагалось, что данные, полученные для каждой из групп, подчиняются нормальному распределению, причем дисперсии по всем группам примерно одинаковы. На этом допущении и основан изложенный в гл. 3 однофакторный дисперсионный анализ. Сейчас мы познакомимся с его непараметрическим аналогом, не требующим предположения о нормальности распределения. Это критерий Крускала—Уоллиса.

Критерий Крускала—Уоллиса представляет собой обобщение критерия Манна—Уитни. Сначала все значения, независимо от того, какой выборке они принадлежат, упорядочивают по возрастанию. Каждому значению присваивается ранг — номер его места в упорядоченном ряду. (Совпадающим значениям присваивают общий ранг, равный среднему тех мест, которые эти величины делят между собой в общем упорядоченном ряду.) Затем вычисляют суммы рангов, относящихся к каждой группе, и для каждой группы определяют средний ранг. При отсутствии межгрупповых различий средние ранги групп должны оказаться близки. Напротив, если существует значительное расхождение средних рангов, то гипотезу об отсутствии межгрупповых различий следует отвергнуть. Значение критерия Крускала—Уоллиса H и является мерой такого расхождения средних рангов.

Для простоты положим, что групп всего три. Обобщение на большее число групп получится автоматически. Имеются результаты измерения некоторого признака в трех группах. Чис-

ленность групп — n_1, n_2 и n_3. Значения объединим, упорядочим и каждому присвоим ранг. Вычислим сумму рангов для каждой группы — R_1, R_2 и R_3. Найдем средние ранги: $\bar{R}_1 = R_1/n_1$, $\bar{R}_2 = R_2/n_2$ и $\bar{R}_3 = R_3/n_3$.

Общее число наблюдений $N = n_1 + n_2 + n_3$. Для объединенной группы рангами являются числа 1, 2, ..., N и общая сумма рангов равна

$$1 + 2 + ... + (N-1) + N = \frac{N(N+1)}{2}.$$

Тогда средний ранг R для объединенной группы равен

$$\bar{R} = \frac{1 + 2 + 3 + ... + N}{N} = \frac{N+1}{2}.$$

Теперь найдем величину D, равную

$$D = n_1(\bar{R}_1 - \bar{R})^2 + n_2(\bar{R}_2 - \bar{R})^2 + n_3(\bar{R}_3 - \bar{R})^2.$$

Это прямой аналог межгрупповой вариации, знакомой нам по гл. 9. Величина D зависит от размеров групп. Чтобы получить показатель, отражающий их различия, следует поделить D на $N(N+1)/12$. Полученная величина

$$H = \frac{D}{N(N+1)/12} = \frac{12}{N(N+1)} \sum n_м (\bar{R}_м - \bar{R})^2$$

является значением критерия Крускала—Уоллиса. Суммирование в приведенной формуле производится по всем группам.

Как найти критическое значение H? Можно было бы просто перечислить все сочетания рангов, как это делалось для критериев Манна—Уитни и Уилкоксона. Однако сделать это довольно трудно — число вариантов слишком велико. К счастью, если группы не слишком малы, распределение H хорошо приближается распределением χ^2 с числом степеней свободы $\nu = k - 1$, где k — число групп. Тогда для проверки нулевой гипотезы нужно просто вычислить по имеющимся наблюдениям значение H и сравнить его с критическим значением χ^2 из табл. 5.7. В случае трех групп приближение с помощью χ^2 пригодно, если численность каждой группы не меньше 5. Для четырех групп —

если общее число наблюдений не менее 10. Но если группы совсем малы, не остается ничего, кроме как обратиться к таблице точных значений распределения Крускала—Уоллиса (мы не приводим эту таблицу из-за ее громоздкости).

Итак, чтобы выяснить, одинаково ли действие нескольких методов лечения, каждый из которых испытывается на отдельной группе, нужно проделать следующее.

- Объединив все наблюдения, упорядочить их по возрастанию. Совпадающим значениям ранги присваиваются как среднее тех мест, которые делят между собой эти значения*.
- Вычислить критерий Крускала—Уоллиса H.
- Сравнить вычисленное значение H с критическим значением χ^2 для числа степеней свободы, на единицу меньшего числа групп. Если вычисленное значение H окажется больше критического, различия групп статистически значимы.

Приведем пример использования критерия Крускала—Уоллиса.

Влияние пероральных контрацептивов на выведение кофеина

Ряд лекарственных средств и пищевых продуктов (кофе, чай и прохладительные напитки) содержат кофеин. Беременным не следует увлекаться крепким кофе, поскольку кофеин может оказать неблагоприятное влияние на плод, а выведение кофеина у беременных замедлено. Существует предположение, что замедленное выведение кофеина обусловлено высоким уровнем половых гормонов во время беременности. Р. Патвардан и соавт.** решили косвенно подтвердить это предположение, определив скорость

---

* *При большом числе совпадающих рангов значение H следует поделить на*

$$1-\frac{\Sigma(\tau_i-1)\tau_i(\tau_i+1)}{N(N^2-1)},$$

где N — число членов всех групп, τ_i — как обычно, число рангов в i-й связке, а суммирование производится по всем связкам.

** R. Patwardhan, P. Desmond, R. Johnson, S. Schenker. Impaired elimination of caffeine by oral contraceptives. *J. Lab. Clin. Med.*, 95:603—608, 1980.

Таблица 10.9. Период полувыведения кофеина

		Женщины			
		Не принимающие пероральных контрацептивов		Принимающие пероральные контрацептивы	
Мужчины					
$T_{1/2}$, ч	Ранг	$T_{1/2}$, ч	Ранг	$T_{1/2}$, ч	Ранг
2,04	1	5,30	12	10,36	25
5,16	10	7,28	19	13,28	29
6,11	15	8,98	21	11,81	28
5,82	14	6,59	16	4,54	6
5,41	13	4,59	8	11,04	26
3,51	4	5,17	11	10,08	24
3,18	2	7,25	18	14,47	31
4,57	7	3,47	3	9,43	23
4,83	9	7,60	20	13,41	30
11,34	27				
3,79	5				
9,03	22				
7,21	17				
Сумма рангов	146		128		222
Средний ранг	11,23		14,22		24,67

выведения кофеина у женщин, принимающих пероральные контрацептивы. (При приеме пероральных контрацептивов уровень эстрогенов и прогестагенов в крови повышается — то же самое происходит и при беременности.)

Скорость выведения кофеина (как и других веществ) непостоянна — она прямо пропорциональна его концентрации в плазме. Поэтому нет смысла измерять скорость выведения, скажем, в миллиграммах в минуту. Вместо этого используют период полувыведения ($T_{1/2}$) — время уменьшения концентрации вдвое: после того как вещество всосется и поступит в кровь, эта величина остается постоянной, пока вещество не будет почти полностью выведено из организма.

$T_{1/2}$ определили у женщин, принимающих и не принимающих пероральные контрацептивы, а также у мужчин. Численность групп составила соответственно 9, 9 и 13 человек. Каждый участ-

ник эксперимента принимал 250 мг кофеина, что соответствует примерно 3 чашкам кофе, после чего дважды определяли концентрацию кофеина в крови и рассчитывали $T_{1/2}$. Результаты представлены в табл. 10.9.

Общий средний ранг

$$\overline{R} = \frac{1+2+3+\ldots+31}{31} = \frac{N+1}{2} = \frac{31+1}{2} = 16.$$

Вычисляем взвешенную сумму квадратов отклонений средних по группам от общего среднего

$$D = 13(11{,}23-16)^2 + 9(14{,}22-16)^2 + 9(24{,}67-16)^2 = 1000{,}82$$

и значение критерия Крускала—Уоллиса

$$H = \frac{12}{N(N+1)}D = \frac{12}{31(31+1)}1000{,}82 = 12{,}107.$$

По табл. 5.7 находим 1% критическое значение χ^2 с числом степеней свободы $\nu = k-1 = 3-1 = 2$. Оно равно 9,210, то есть меньше полученного нами. Таким образом, различия групп статистически значимы ($P < 0{,}01$).

Непараметрическое множественное сравнение

Потребность во множественном сравнении возникает всякий раз, когда с помощью дисперсионного анализа (или его непараметрического аналога — критерия Крускала—Уоллиса) обнаруживается различие нескольких выборок. В этом случае и требуется установить, в чем состоит это различие. В гл. 4 мы познакомились с параметрическими методами множественного сравнения. Они позволяют сравнить группы попарно и затем объединить их в несколько однородных наборов так, что различия между группами из одного набора статистически незначимы, а между группами из разных наборов — значимы. Кроме того, они позволяют сравнить все группы с контрольной.

К счастью, параметрические методы множественного сравнения легко преобразовать в непараметрические. Когда объемы выборок равны, для множественного сравнения используют не-

параметрические варианты критериев Ньюмена—Кейлса и Даннета. Когда же объемы выборок различны, применяется критерий Данна. Опишем вкратце эти методы.

Начнем с критериев для выборок равного объема. Критерии Ньюмена—Кейлса и Даннета совпадают практически полностью, поскольку критерий Даннета есть просто вариант критерия Ньюмена—Кейлса для сравнения всех выборок с одной контрольной.

Формула для непараметрического варианта критерия Ньюмена—Кейлса:

$$q = \frac{R_A - R_B}{\sqrt{\frac{n^2 l(nl+1)}{12}}},$$

где R_A и R_B — суммы рангов двух сравниваемых выборок, n — объем каждой выборки, l — интервал сравнения. Вычисленное q сравнивается с критическим значением в табл. 4.3 для бесконечного числа степеней свободы.

Значение непараметрического критерия Даннета определяется формулой:

$$q' = \frac{R_{\text{кон}} - R_A}{\sqrt{\frac{n^2 l(nl+1)}{6}}},$$

где $R_{\text{кон}}$ — сумма рангов контрольной выборки, а остальные величины те же, что в критерии q. Уточним только, что l — число всех выборок, включая контрольную. Значение q' сравнивается с критическим значением для бесконечного числа степеней свободы (табл. 4.4).

Наконец, для сравнения выборок разного объема используется критерий Данна. Впрочем, ничто не мешает применить его и к выборкам одинакового объема. Значение критерия Данна:

$$Q = \frac{\overline{R}_A - \overline{R}_B}{\sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}},$$

Таблица 10.10. Критические значения Q для попарного сравнения групп

Число сравниваемых выборок k	Уровень значимости α	
	0,05	0,01
2	1,960	2,576
3	2,394	2,936
4	2,639	3,144
5	2,807	3,291
6	2,936	3,403
7	3,038	3,494
8	3,124	3,570
9	3,197	3,635
10	3,261	3,692
11	3,317	3,743
12	3,368	3,789
13	3,414	3,830
14	3,456	3,868
15	3,494	3,902
16	3,529	3,935
17	3,562	3,965
18	3,593	3,993
19	3,622	4,019
20	3,649	4,044
21	3,675	4,067
22	3,699	4,089
23	3,722	4,110
24	3,744	4,130
25	3,765	4,149

где $\overline{R}_A$ и $\overline{R}_B$ — средние ранги двух сравниваемых выборок, n_A и n_B — их объемы, а N — общий объем всех сравниваемых выборок.

Критические значения Q приведены в табл. 10.10. «Стягивающее» сравнение проводится как в критерии Ньюмена—Кейлса.

Критерием Данна можно воспользоваться и для сравнения с контрольной выборкой. При этом формула для Q остается прежней, только критические значения находятся уже по табл. 10.11.

Еще одна чашка кофе

Вернемся к исследованию выведения кофеина. Мы уже установили, что между тремя группами (группа мужчин и две группы

Таблица 10.11. Критические значения Q для сравнения с контрольной группой

Число сравниваемых выборок k	Уровень значимости α	
	0,05	0,01
2	1,960	2,576
3	2,242	2,807
4	2,394	2,936
5	2,498	3,024
6	2,576	3,091
7	2,639	3,144
8	2,690	3,189
9	2,735	3,227
10	2,773	3,261
11	2,807	3,291
12	2,838	3,317
13	2,866	3,342
14	2,891	3,364
15	2,914	3,384
16	2,936	3,403
17	2,955	3,421
18	2,974	3,437
19	2,992	3,453
20	3,008	3,467
21	3,024	3,481
22	3,038	3,494
23	3,052	3,506
24	3,066	3,518
25	3,078	3,529

J. H. Zar, Biostatistical analysis. 2nd ed., Prentice-Hall, Englewood Cliffs, N. J., 1984.

женщин — принимающих и не принимающих пероральные контрацептивы) существует различие в скорости выведения кофеина. Однако осталось неизвестным, какие группы отличаются друг от друга, а какие похожи. Для ответа на этот вопрос предназначены методы множественного сравнения. Поскольку численность групп разная, применим критерий Данна.

Из табл. 10.9 видно, что сильнее всего различаются средние ранги в 3-й группе (женщины, принимающие пероральные контрацептивы) и в 1-й группе (мужчины). Вычисляем значение критерия Данна:

$$Q = \frac{\overline{R}_3 - \overline{R}_1}{\sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_3} + \frac{1}{n_1}\right)}} = \frac{24{,}67 - 11{,}23}{\sqrt{\frac{31(31+1)}{12}\left(\frac{1}{9} + \frac{1}{13}\right)}} = 3{,}409.$$

В табл. 10.10 находим 5% критическое значение для $k = 3$. Оно равно 2,394, то есть меньше выборочного. Тем самым, различия групп статистически значимы ($P < 0{,}05$). Продолжим стягивающие сравнения. Следующая пара групп — женщины, принимающие пероральные контрацептивы (3-я группа), и женщины, не принимающие пероральных контрацептивов (2-я группа):

$$Q = \frac{\overline{R}_3 - \overline{R}_2}{\sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_3} + \frac{1}{n_2}\right)}} = \frac{24{,}67 - 14{,}22}{\sqrt{\frac{31(31+1)}{12}\left(\frac{1}{9} + \frac{1}{9}\right)}} = 2{,}438.$$

Это значение также больше критического.

Наконец, для оставшейся пары групп:

$$Q = \frac{\overline{R}_2 - \overline{R}_1}{\sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_2} + \frac{1}{n_1}\right)}} = \frac{14{,}22 - 11{,}23}{\sqrt{\frac{31(31+1)}{12}\left(\frac{1}{9} + \frac{1}{13}\right)}} = 0{,}7583,$$

что меньше критического. Итак, выведение кофеина у женщин, принимающих пероральные контрацептивы, медленнее, чем у женщин, не принимающих пероральных контрацептивов, и у мужчин; последние же две группы по скорости выведения кофеина друг от друга не отличаются. Предположение о влиянии половых гормонов на выведение кофеина подтвердилось.

ПОВТОРНЫЕ ИЗМЕРЕНИЯ: КРИТЕРИЙ ФРИДМАНА

Если одна и та же группа больных последовательно подвергается нескольким методам лечения или просто наблюдается в разные моменты времени, применяют дисперсионный анализ повторных измерений (гл. 9). Но чтобы использование дисперсионного анализа было правомерно, данные должны подчиняться нор-

Таблица 10.12. Данные для расчета критерия Фридмана. Пример 1

Больной	Метод лечения			
	1	2	3	4
1	1	2	3	4
2	4	1	2	3
3	3	4	1	2
4	2	3	4	1
5	1	4	3	2
Сумма рангов	11	14	13	12

мальному распределению. Если вы в этом не уверены, лучше воспользоваться критерием Фридмана — непараметрическим аналогом дисперсионного анализа повторных измерений.

Логика критерия Фридмана очень проста. Каждый больной ровно один раз подвергается каждому методу лечения (или наблюдается в фиксированные моменты времени). Результаты наблюдений у *каждого* больного упорядочиваются. Обратите внимание, что если раньше мы упорядочивали группы, то теперь мы отдельно упорядочиваем значения у каждого больного независимо от всех остальных. Таким образом, получается столько упорядоченных рядов, сколько больных участвует в исследовании. Далее, для каждого метода лечения (или момента наблюдения) вычислим сумму рангов. Если разброс сумм велик — различия статистически значимы.

В табл. 10.12 описаны результаты испытания 4 методов лечения на 5 больных. В таблице указаны не сами значения, а их ранги среди данных, относящихся к одному больному. Каждая строка, кроме последней, соответствует одному больному. Последняя строка содержит суммы рангов для каждого из методов лечения. Различие сумм невелико; не похоже, чтобы эффективность какого-то метода отличалась от эффективности других.

Теперь обратимся к табл. 10.13. Различие в эффективности методов выражено предельно четко — упорядочение одинаково для всех больных. Во всех случаях наиболее эффективным оказался первый метод лечения, следующим — третий, за ним четвертый, и наконец, наименее эффективным — второй.

Таблица 10.13. Данные для расчета критерия Фридмана. Пример 2

	Метод лечения			
Больной	1	2	3	4
1	4	1	3	2
2	4	1	3	2
3	4	1	3	2
4	4	1	3	2
5	4	1	3	2
Сумма рангов	20	5	15	10

Перейдем к количественному оформлению наших впечатлений. Критерий Фридмана сходен с критерием Крускала—Уоллиса и вычисляется следующим образом. Сначала рассчитаем среднюю сумму рангов, присвоенных одному методу. (Именно этой величине равнялась бы сумма рангов любого из методов, если бы они были в точности равноэффективны.) Затем вычислим сумму квадратов S отклонений истинных сумм рангов, полученных каждым из методов, от средней суммы.

Разберем это на примере данных из табл. 10.12 и 10.13. Для каждого больного средний ранг равен $(1+2+3+4)/4 = 2{,}5$. В общем случае при k методах лечения средний ранг равен

$$\frac{1+2+3+\ldots+k}{k} = \frac{k+1}{2}.$$

Если каждым методом лечилось n больных, средняя сумма рангов равна $n(k+1)/2$. В нашем примере $n = 5$. Поэтому средняя сумма рангов равна $5(4+1)/2 = 12{,}5$.

Значение критерия S определяется формулой

$$S = \sum \left(R_{\text{м}} - \frac{n(k+1)}{2} \right)^2,$$

где $R_{\text{м}}$ — истинные суммы рангов для методов лечения.

Тогда для табл. 10.12 находим:

$$S = (11-12{,}5)^2 + (14-12{,}5)^2 + (13-12{,}5)^2 + (12-12{,}5)^2 =$$
$$= (-1{,}5)^2 + (1{,}5)^2 + (0{,}5)^2 + (-0{,}5)^2 = 5,$$

а для табл. 10.13:

$$S = (20-12{,}5)^2 + (5-12{,}5)^2 + (15-12{,}5)^2 + (10-12{,}5)^2 =$$
$$= (7{,}5)^2 + (-7{,}5)^2 + (2{,}5)^2 + (-2{,}5)^2 = 125.$$

Значение S для второй таблицы значительно превосходит значение для первой, что соответствует нашим первоначальным впечатлениям. Величина S позволяет судить, одинакова ли эффективность исследуемых методов.

Однако поделив ее на $nk(k+1)/12$, мы получим более удобный критерий:

$$\chi_r^2 = \frac{12}{nk(k+1)} S = \frac{12}{nk(k+1)} \sum \left(R_м - \frac{n(k+1)}{2} \right)^2 .$$

Это и есть критерий Фридмана. При большой численности группы его величина приблизительно следует распределению χ^2 с числом степеней свободы $\nu = k-1$. Однако при $k = 3$ и $n \le 9$ и при $k = 4$ и $n \le 4$ это приближение оказывается слишком грубым. В таком случае нужно воспользоваться приведенными в табл. 10.14 точными значениями χ_r^2.

Повторим порядок расчета критерия Фридмана.

- Расположите значения для каждого больного по возрастанию, каждому значению присвойте ранг.
- Для каждого из методов лечения подсчитайте сумму присвоенных ему рангов.
- Вычислите значение χ_r^2.
- Если число методов лечения и число больных присутствует в табл. 10.14, определите критическое значение χ_r^2 по этой таблице. Если число методов лечения и число больных достаточно велико (отсутствует в таблице), воспользуйтесь критическим значением χ^2 с числом степеней свободы $\nu = k-1$.
- Если рассчитанное значение χ_r^2 превышает критическое — различия статистически значимы.

Теперь применим критерий Фридмана для анализа уже знакомого исследования.

Таблица 10.14. Критические значения критерия Фридмана

$k=3$			$k=4$		
n	χ_r^2	P	n	χ_r^2	P
3	6,00	0,028	2	6,00	0,042
4	6,50	0,042	3	7,00	0,054
	8,00	0,005		8,20	0,017
5	5,20	0,093	4	7,50	0,054
	6,40	0,039		9,30	0,011
	8,40	0,008	5	7,80	0,049
6	5,33	0,072		9,96	0,009
	6,33	0,052	6	7,60	0,043
	9,00	0,008		10,20	0,010
7	6,00	0,051	7	7,63	0,051
	8,86	0,008		10,37	0,009
8	6,25	0,047	8	7,65	0,049
	9,00	0,010		10,35	0,010
9	6,22	0,048			
	8,67	0,010			
10	6,20	0,046			
	8,60	0,012			
11	6,54	0,043			
	8,91	0,011			
12	6,17	0,050			
	8,67	0,011			
13	6,00	0,050			
	8,67	0,012			
14	6,14	0,049			
	9,00	0,010			
15	6,40	0,047			
	8,93	0,010			

k — число методов лечения (моментов наблюдения), n — число больных, α — уровень значимости.

Owen. Handbook of statistical tables. US Department of Energy, Addison-Wesley, Reading, Mass., 1962.

Таблица 10.15. Легочное сосудистое сопротивление при лечении гидралазином

	Легочное сосудистое сопротивление					
	До лечения (контрольное)		Спустя 48 часов		Спустя 3—6 месяцев	
Больной	Величина	Ранг	Величина	Ранг	Величина	Ранг
1	22,2	3	5,4	1	10,6	2
2	17,0	3	6,3	2	6,2	1
3	14,1	3	8,5	1	9,3	2
4	17,0	3	10,7	1	12,3	2

Гидралазин при первичной легочной гипертензии

В табл. 10.15 воспроизведены данные о легочном сосудистом сопротивлении из табл. 9.5. В предыдущей главе мы применили к ним дисперсионный анализ повторных измерений. Это допустимо в случае нормального распределения. Но данных так мало, что судить о распределении невозможно. Поэтому прибегнем к критерию Фридмана, не требующему нормальности распределения.

Имеем три измерения ($k = 3$) у четырех больных ($n = 4$). Средний ранг для каждого наблюдения $1 + 2 + 3/3 = 2$. Средняя сумма рангов для каждого измерения равна $4 \times 2 = 8$. Сумма квадратов отклонений для трех наблюдений:

$$S = (12-8)^2 + (5-8)^2 + (7-8)^2 = (4^2) + (-3)^2 + (-1)^2 = 26,$$

$$\chi_r^2 = \frac{12}{nk(k+1)} S = \frac{12}{4 \times 3 \times 4} 26 = 6{,}5.$$

Эта величина совпадает с критическим значением χ_r^2 при $n = 4$ и $k = 3$. Соответствующий точный уровень значимости составляет 0,042. Таким образом, различия между измерениями статистически значимы ($P < 0{,}05$).

Множественное сравнение после применения критерия Фридмана

Как всегда, за выявлением различий между несколькими методами лечения должно последовать выяснение, в чем состоят эти

различия, то есть попарное сравнение методов лечения. Поскольку число больных, подвергшихся каждому методу лечения, одинаково, для этой цели легко приспособить критерий Ньюмена—Кейлса. Если считать один из методов лечения «контролем», то остальные можно сравнить с ним при помощи критерия Даннета. Если речь идет о повторных наблюдениях в ходе лечения, таким контролем естественно считать значения, полученные перед началом лечения.

Итак, для попарного сравнения методов лечения (или моментов наблюдения) применяется критерий Ньюмена—Кейлса:

$$q = \frac{R_A - R_B}{\sqrt{\frac{nl(l+1)}{12}}},$$

где R_A и R_B — суммы рангов для двух сравниваемых методов лечения, l — интервал сравнения, а n — число больных. Найденное значение q сравнивается с критическим из табл. 4.3 для бесконечного числа степеней свободы. Если найденное значение больше критического, различие методов лечения (моментов наблюдения) статистически значимо.

Для сравнения с контрольной группой применяется критерий Даннета:

$$q' = \frac{R_{\text{кон}} - R_A}{\sqrt{\frac{nl(l+1)}{6}}},$$

где l — число всех групп, включая контрольную, $R_{\text{кон}}$ — сумма рангов в контрольной группе. Остальные величины определяются, как в формуле для q. Значение q' сравнивается с критическим из табл. 4.4 для бесконечного числа степеней свободы.

Пассивное курение при ишемической болезни сердца

При ишемической болезни сердца коронарные артерии сужены атеросклеротическими бляшками. В отсутствие физической нагрузки, когда потребность миокарда в кислороде низка, это никак не сказывается на состоянии больного. Однако при физи-

ческой нагрузке, когда потребность миокарда в кислороде увеличивается, коронарные артерии уже не могут обеспечить соответствующего увеличения кровотока и развивается приступ стенокардии.

Курение для больных ишемической болезнью сердца особенно вредно. Тому есть несколько причин. Первая — при курении происходит сужение артерий и ухудшается кровоток. К сердцу поступает меньше кислорода и питательных веществ, затрудняется удаление продуктов метаболизма. Вторая причина — окись углерода из сигаретного дыма проникает в кровь и связывается с гемоглобином, замещая кислород. И наконец, третья причина — никотин и другие содержащиеся в табачном дыме вещества снижают сократимость миокарда, уменьшая кровоток и снабжение кислородом и питательными веществами всех органов, в том числе самого миокарда. В результате переносимость физической нагрузки снижается — приступы стенокардии возникают при менее интенсивной и продолжительной физической нагрузке.

Приводит ли к таким же последствиям пассивное курение? На этот вопрос попытался ответить У. Аронов*.

В эксперименте участвовали 10 больных ишемической болезнью сердца. Переносимость физической нагрузки определяли как время, в течение которого больной мог выполнять работу (крутить велотренажер) до возникновения приступа стенокардии.

У каждого больного определяли переносимость физической нагрузки, затем в течение 2 часов он отдыхал в отдельной комнате, где присутствовала специальная группа окуривателей из 3 человек. Окуриватели либо не курили, либо выкуривали по 5 сигарет, в последнем случае помещение либо проветривали, либо не проветривали. После такого отдыха переносимость физической нагрузки определяли вновь. Исследование продолжалось 3 дня, и каждый больной испытал (в случайном порядке) все три вида отдыха, по одному в день. Результаты представлены в табл. 10.16.

Сначала, рассматривая данные как 6 отдельных измерений,

---

* W. S. Aronow. Effect of passive smoking on angina pectoris. *N. Engl. J. Med.*, 299: 21—24, 1978.

оценим статистическую значимость различий между ними. Применим критерий Фридмана. Средний ранг равен

$$\frac{1+2+3+4+5+6}{6}=3{,}5.$$

Средняя сумма рангов по каждому измерению $3{,}5\times10=35$. Тогда:

$$S=(44-35)^2+(53-35)^2+(39-35)^2+$$
$$+(20-35)^2+(44-35)^2+(10-35)^2=1352,$$

$$\chi_r^2=\frac{12}{10\times6(6+1)}1352=38{,}629.$$

Полученное значение больше 20,517 — критического значения χ^2 для 0,1% уровня значимости при $\nu=k-1=6-1=5$ степенях свободы (см. табл. 5.7). Тем самым, различия статистически значимы.

Чтобы понять, в чем заключаются различия, применим критерий Ньюмена—Кейлса. Все измерения перенумеруем как показано в табл. 10.6, расположим по убыванию сумм рангов и приступим к попарному сравнению. Крайние суммы рангов — 53 при 2-м измерении и 10 при 6-м измерении. Интервал сравнения $l=6$, число больных $n=10$.

$$q=\frac{R_2-R_6}{\sqrt{\frac{nl(l+1)}{12}}}=\frac{53-10}{\sqrt{\frac{10\times6\times7}{12}}}=7{,}268.$$

Значение q превышает 4,030 — критическое значение q для уровня значимости $\alpha'=0{,}05$, интервала сравнения $l=6$ и бесконечного числа степеней свободы (табл. 4.3А). Различия статистически значимы. Остальные попарные сравнения приведены в табл. 10.17. Уровни четко разделяются на три группы. Первая группа (максимальная переносимость физической нагрузки) включает 1, 2, 3 и 5-е измерения, то есть все три измерения до отдыха, а также измерение после отдыха на свежем воздухе. Вторая группа представлена единственным измерением — после отдыха в прокуренном, но проветриваемом помещении. Нако-

Таблица 10.16. Продолжительность физической нагрузки у больных ишемической болезнью сердца до и после отдыха при пассивном курении разной интенсивности, с

	Свежий воздух				Пассивное курение, проветривание				Пассивное курение без проветривания			
	1		2		3		4		5		6	
	До отдыха		После отдыха		До отдыха		После отдыха		До отдыха		После отдыха	
Больной	Время	Ранг	Время	Ранг	Время	Ранг	Время	Ранг	Время	Ранг	Время	Ранг
1	193	4	217	6	191	3	149	2	202	5	127	1
2	206	5	214	6	203	4	169	2	189	3	130	1
3	188	4	197	6	181	3	145	2	192	5	128	1
4	375	3	412	6	400	5	306	2	387	4	230	1
5	204	5	199	4	211	6	170	2	196	3	132	1
6	287	3	310	5	304	4	243	2	312	6	198	1
7	221	5	215	4	213	3	158	2	232	6	135	1
8	216	5	223	6	207	3	155	2	209	4	124	1
9	195	4	208	6	186	3	144	2	200	5	129	1
10	231	6	224	4	227	5	172	2	218	3	125	1
Сумма рангов		44		53		39		20		44		10
Среднее время	231,6		241,9		232,3		181,1		233,7		145,8	

W. S. Aronow. Effect of passive smoking on angina pectoris. *N. Engl. J. Med.*, 299: 21—24, 1978.

Таблица 10.17. Попарные сравнения

Сравнение	Разность сумм рангов	l	q	Критическое значение q	$P<0{,}05$
2 и 6	53 – 10 = 43	6	7,268	4,030	Да
2 и 4	53 – 20 = 33	5	6,600	3,858	Да
2 и 3	53 – 39 = 14	4	3,430	3,633	Нет
2 и 5	53 – 44 = 9	3			Нет
2 и 1	53 – 44 = 9	2			Нет
1 и 6	44 – 10 = 34	5	6,800	3,858	Да
1 и 4	44 – 20 = 24	4	5,879	3,633	Да
1 и 3	44 – 39 = 5	3	1,581	3,314	Нет
1 и 5	44 – 44 = 0	2			Нет
5 и 6	44 – 10 = 34	4	8,329	3,633	Да
5 и 4	44 – 20 = 24	3	7,590	3,314	Да
5 и 3	44 – 39 = 5	2	2,236	2,772	Нет
3 и 6	39 – 10 = 29	3	9,171	3,314	Да
3 и 4	39 – 20 = 19	2	8,497	2,772	Да
4 и 6	20 – 10 = 10	2	4,472	2,772	Да

нец, третья группа (переносимость физической нагрузки минимальная) также содержит единственное измерение — после отдыха в прокуренном непроветриваемом помещении. Между измерениями, вошедшими в разные группы, различия статистически значимы (при $\alpha' = 0{,}05$). Общий вывод из работы Аронова: пассивное курение снижает переносимость физической нагрузки при ишемической болезни сердца.

ВЫВОДЫ

Изложенные в этой главе методы предназначены для проверки тех же гипотез, что критерий Стьюдента и дисперсионный анализ, но при этом не требуют, чтобы данные подчинялись нормальному распределению. Заменяя исходные данные рангами и избавляясь тем самым от необходимости делать какие-либо предположения относительно типа распределения, мы сохраняем большую часть информации о значениях признака и их изме-

нениях. Если распределение все же оказывается нормальным, то при этом происходит некоторое снижение чувствительности. Однако если распределение отлично от нормального, непараметрические методы чувствительнее параметрических.

Обратите внимание, что, оперируя не данными, а рангами, рассмотренные методы строятся, в сущности, по тому же принципу, что и рассмотренные ранее параметрические, такие, как критерий Стьюдента и дисперсионный анализ. Заменив данные рангами, мы делаем следующее.

- Формулируем нулевую гипотезу, то есть предполагаем, что наблюдаемые различия случайны.
- Выбираем критерий, то есть числовое выражение различий.
- Определяем, каким было бы распределение величины критерия при условии справедливости нулевой гипотезы.
- Находим критическое значение, то есть величину, которую при справедливости нулевой гипотезы значение критерия превышает достаточно редко (точнее, с вероятностью, равной уровню значимости α).
- Вычисляем значение критерия для наших данных и сравниваем его с критическим: если вычисленное значение больше, признаем различия статистически значимыми.

Выбор между параметрическими и непараметрическими методами определяется прежде всего характером данных. Имея дело с порядковыми признаками, не остается ничего, кроме как воспользоваться непараметрическими методами. Если признак числовой, стоит подумать, нормально ли его распределение. Тут могут помочь как общие соображения, так и графическое представление данных. Даже если нет веских оснований сомневаться в нормальности распределения, но данных мало, или вы не хотите делать никаких предположений о типе распределения — воспользуйтесь непараметрическими методами.

ЗАДАЧИ

10.1 Анализы, инструментальные исследования и лекарственные средства назначает врач, а платит за них главным образом больной. Многие врачи весьма смутно представляют себе

К задаче 10.1.

Врач	Среднегодовые расходы на обследование одного больного, долл.		Среднегодовые расходы на лечение одного больного, долл.	
	До ознакомления с расходами	После ознакомления с расходами	До ознакомления с расходами	После ознакомления с расходами
1	20	20	32	42
2	17	26	41	90
3	14	1	51	71
4	42	24	29	47
5	50	1	76	56
6	62	47	47	43
7	8	15	60	137
8	49	7	58	63
9	81	65	40	28
10	54	9	64	60
11	48	21	73	87
12	55	36	66	69
13	56	30	73	50

стоимость своих назначений и не озабочены тем, чтобы уменьшить расходы больного. Чтобы побудить врачей задуматься об этом, все шире практикуется учет затрат на обследование и лечение. Есть ли основания считать, что это сделает врача более экономным? Интересное исследование провели С. Шредер и соавт. (S. Schroeder et al. Use of laboratory tests and pharmaceuticals: variation among physicians and effect of cost audit on subsequent use. *JAMA*, 225:969—973, 1973). В течение трех месяцев они регистрировали расходы на обследование и лечение амбулаторных больных, которых наблюдали врачи из клиники Вашингтонского университета. Данные собирали по больным со сходными заболеваниями. Рассчитав для каждого врача среднегодовые расходы на обследование и лечение одного больного, составили общий список, который раздали врачам. Каждый врач знал свой номер в списке, но не знал номеров своих коллег, таким образом он мог сравнить свои расходы с расходами других, но не знал,

кого именно. Через некоторое время исследователи проверили, какие изменения произошли в расходовании средств у тех же врачей. Результаты представлены в таблице на предыдущей странице.

Произошли ли изменения в расходах на обследование и лечение? Есть ли связь между расходами на обследование и лечение? Как можно объяснить полученные результаты?

10.2. При заболеваниях сетчатки повышается проницаемость ее сосудов. Дж. Фишмен и соавт. (G. Fishman et al. Blood-retinal barrier function in patients with cone or cone-rod dystrophy. *Arch. Ophthalmol.*, 104:545—548, 1986) измерили проницаемость сосудов сетчатки у здоровых и у больных с ее поражением. Полученные результаты приведены в таблице.

Проницаемость сосудов сетчатки

Здоровые	Больные
0,5	1,2
0,7	1,4
0,7	1,6
1,0	1,7
1,0	1,7
1,2	1,8
1,4	2,2
1,4	2,3
1,6	2,4
1,6	6,4
1,7	19,0
2,2	23,6

С помощью непараметрического метода проверьте, подтверждают ли эти данные гипотезу о различии в проницаемости сосудов сетчатки? После этого воспользуйтесь соответствующим параметрическим методом. Если выводы окажутся иными, объясните, в чем причина различия.

10.3. Данные задачи 10.2 — часть более широкого исследования проницаемости сетчатки. Сравните данные, относящиеся к разным видам поражений.

Проницаемость сосудов сетчатки

Нормальная сетчатка	Поражение только в области центральной ямки	Аномалии в области центральной ямки и на периферии
0,5	1,2	6,2
0,7	1,4	12,6
0,7	1,6	12,8
1,0	1,7	13,2
1,0	1,7	14,1
1,2	1,8	15,0
1,4	2,2	20,3
1,4	2,3	22,7
1,6	2,4	27,7
1,6	6,4	
1,7	19,0	
2,2	23,6	

10.4. Решите задачи 9.5 и 9.6, используя непараметрические методы.

10.5. В гл. 3 на примере больных пиелонефритом была рассмотрена зависимость продолжительности госпитализации от правильности лечения. Д. Кнапп и соавт. решили выяснить, наблюдается ли такая зависимость при лечении пневмонии. Изучив 28 историй болезни, исследователи обнаружили следующее.

Продолжительность госпитализации, сут				
При правильном лечении	При неправильном лечении			
3,7	3,8	1,7	4,8	8,6
2,5	6,8	2,5	5,3	9,0
2,8	7,9	2,9	5,5	10,3
3,0	8,8	3,0	5,8	11,0
5,5	9,0	3,4	7,1	
6,4	9,3	3,7	6,6	

Есть ли разница в продолжительности госпитализации?

10.6. Предсердный натрийуретический гормон усиливает выведение натрия и воды почками. В. Хименес и соавт. (W. Jimenez

et al. Atrial natriuretic factor: reduced cardiac content in cirrhotic rats with ascites. *Am. J. Physiol.*, 250:F749—F752, 1986) исследовали его роль в задержке натрия и воды при циррозе печени. Крысам вводили экстракт предсердия: одной группе — экстракт, полученный от здоровых крыс, другой — от крыс с циррозом печени. Регистрировали изменение выделения натрия с мочой (в процентах от исходного). Результаты представлены в таблице. Какой вывод можно сделать по результатам опыта?

Экстракт от здоровых крыс	Экстракт от крыс с циррозом
760	80
1000	80
1370	80
1680	210
1970	210
2420	320
3260	500
5000	610
5400	760
7370	760
	890
	890
	1870
	1950

10.7. Введя изотоп внутривенно и наблюдая за его распространением с помощью гамма-камеры, можно определить кровенаполнение различных органов, в том числе легких. Р. Окада и соавт. (R. Okada et al. Radionuclide-determined change in pulmonary blood volume with exercise: improved sensitivity of multigated blood-pool scanning in detecting coronary-artery disease. *N. Engl. J. Med.*, 301:569—576, 1979) решили использовать этот метод для локализации поражения коронарных артерий при ишемической болезни сердца. Правая коронарная артерия снабжает кровью главным образом правый желудочек, левая — главным образом левый. Левый желудочек перекачивает кровь, которая поступает в него из легких, по всему телу. При поражении левой коронарной артерии кровоснабжение левого желудочка ухудшается. В покое, когда

объем перекачиваемой крови невелик, это никак не проявляется, однако при физической нагрузке это приводит к накоплению крови в легких. При поражении правой коронарной артерии этого не происходит. Примерно так рассуждали авторы, приступая к работе. Было обследовано 33 человека: 9 здоровых (1-я группа) и 24 больных ишемической болезнью сердца, из них 5 с поражением только правой коронарной артерии (2-я группа) и 19 с поражением обеих коронарных артерий или только левой (3-я группа). Рассчитывали отношение кровенаполнения легких при физической нагрузке к кровенаполнению в покое: по мысли авторов, в 3-й группе этот показатель должен быть выше, чем в первых двух. Результаты представлены в таблице.

Группа		
1	2	3
0,83	0,86	0,98
0,89	0,92	1,02
0,91	1,00	1,03
0,93	1,02	1,04
0,94	1,20	1,05
0,97		1,06
0,97		1,07
0,98		1,22
1,02		1,07
		1,23
		1,13
		1,08
		1,32
		1,10
		1,15
		1,37
		1,18
		1,12
		1,58

Различаются ли группы между собой? Если да, то как именно и достаточно ли велико различие, чтобы исследуемый показа-

тель можно было использовать для определения пораженной коронарной артерии?

10.8. Грезя о славе, автор этих строк предложил новый метод оценки эффективности лечения. Преимущество метода — его простота. Он состоит в следующем. Если у больного интересующий нас показатель увеличивается, ставится оценка +1, если уменьшается — 0 (допустим, случай неизменности показателя исключен). Сумма оценок по всем больным и есть значение критерия G. Вот пример расчета.

	Значение показателя			
Больной	до лечения	после лечения	Изменение показателя	Оценка
1	100	110	+10	+1
2	95	96	+1	+1
3	120	100	–20	0
4	111	123	+12	+1

Значение критерия $G = 1 + 1 + 0 + 1 = 3$. Является ли G полноценным критерием? Постройте распределение G и найдите критическое значение для случаев, когда число больных равно 4 и 6.

Глава 11

Анализ выживаемости

До сих пор мы имели дело только с полными данными: мы знали исход лечения у каждого больного. В гл. 5 мы разобрали работу, целью которой было определить влияние аспирина на риск тромбоза шунта у больных на гемодиализе. Мы подсчитали число больных с тромбозом и без тромбоза в группах аспирина и плацебо и свели результаты в таблицу сопряженности (см. табл. 5.1). Затем мы построили вторую таблицу сопряженности, содержавшую ожидаемые числа, которые наблюдались бы, если бы в группах аспирина и плацебо частота тромбозов была одинаковой. По двум этим таблицам мы вычислили величину χ^2. Полученное значение оказалось достаточно большим, чтобы отклонить гипотезу об отсутствии межгрупповых различий. В этом исследовании срок наблюдения всех больных был одинаковым и никто из них не выбыл из-под наблюдения до завершения исследования. То же самое можно сказать об исследовании галотановой и морфиновой анестезии, с которым мы впервые встретились в гл. 2. Тогда, говоря о трудностях, связанных с проспективными исследованиями, мы

упомянули о проблеме выбывания*, но в рассмотренных примерах мы с ней не сталкивались. Однако ситуация, когда исследование должно быть завершено до наступления исхода у всех больных, для проспективных исследований, в частности клинических испытаний, скорее правило, чем исключение. Понятно, что на этот случай нужны специальные статистические методы.

Наиболее типичный пример исследования такого рода — это изучение выживаемости, когда больных наблюдают от начала болезни до смерти. Обычно больных включают в исследование на всем его протяжении, поэтому оно всегда заканчивается до смерти последнего больного. Истинная продолжительность болезни выживших к концу исследования остается неизвестной. Кроме того, исследователь может потерять больного из виду до завершения исследования, если тот, к примеру, переехал в другой город. Наконец, больной может умереть по причине, не связанной с изучаемым заболеванием, например погибнуть в автокатастрофе. Во всех этих случаях длительность заболевания остается неизвестной, мы знаем только, что она превышает некоторый срок.

Сейчас мы займемся именно изучением выживаемости, однако будем иметь в виду, что те методы, которые мы освоим, пригодны и для других исследований, в том числе для контролируемых испытаний.

ПАССИВНОЕ КУРЕНИЕ НА ПЛУТОНЕ

Табачные дельцы, теснимые все дальше от Земли борцами за здоровый образ жизни, окопались на Плутоне. Они решили превратить эту девственную планету в оплот табакокурения. Многие наивные плутониане поддались навязчивой рекламе и закурили. Но это еще полбеды. Как известно, на Плутоне очень холодно, по-

* Здесь мы не говорим о *пропусках в данных*, причины которых — ошибка измерения, разбитая пробирка с пробой, потерянный анализ и т. п. К данным, содержащим пропуски, применяются обычные статистические методы с внесением необходимых вычислительных поправок. Подробнее об анализе данных с пропусками можно прочесть в книге S. Glantz, B. Slinker. Primer of applied regression and analysis of variance. McGraw-Hill, N.Y., 1990.

этому его обитатели редко покидают свои домики. Чрезвычайно деликатные по природе, плутониане не могут выставить курильщика на улицу и вынуждены дышать табачным дымом, который производит их несознательный соотечественник.

Плутониане вообще живут недолго, что же будет теперь, когда Плутон охватила эпидемия пассивного курения! Первое, что мы должны сделать в этой ситуации, — это оценить продолжительность жизни плутонианина после начала пассивного курения.

Вот как проводилось исследование. Мы попросили всех плутониан сообщать нам, как только в их домике появится активный курильщик. Выявленных таким образом пассивных курильщиков включали в группу наблюдения и дожидались (увы!) их смерти. Исследование длилось 15 плутонианских часов; за это время пассивными курильщиками стали 10 плутониан. Первыми сообщили о начале пассивного курения А и Б. Остальные участники вошли в группу наблюдения уже после начала исследования (что типично для исследований выживаемости); их звали В, Г, Д, Е, Ж, З, И и К. Периоды наблюдения за каждым из них показаны на рис 11.1А в виде горизонтальных отрезков. Из десяти участников к концу исследования умерли семь — А, Б, В, Е, Ж, З и К; в живых остались двое — Г и И. Еще одного участника, Д, местное начальство на 14-м часу исследования послало в командировку на Нептун; что с ним было дальше, нам неизвестно.

Таким образом, продолжительность жизни после начала пассивного курения нам известна в 7 случаях. В 3 случаях нам известно только, что наблюдаемые прожили не меньше такого-то срока*. Неважно, почему они не прослежены до конца жизни —

* В исследованиях выживаемости неполные данные называют также *цензурированными*. Данные о трех выбывших плутонианах цензурированы *справа* — известен момент начала наблюдения, но неизвестно, когда наблюдаемый умер. Если бы в исследовании участвовали плутониане, начавшие курить до его начала, то мы могли бы получить также данные, цензурированные *слева*, а также цензурированные с обеих сторон. Эти виды цензурирования и соответствующие методы анализа можно найти в D. Collett. Modelling survival in medical research. Chapman and Hall, London, 1994 и E. T. Lee. Statistical methods for survival data analysis. Wiley, 2nd ed., New York, 1992.

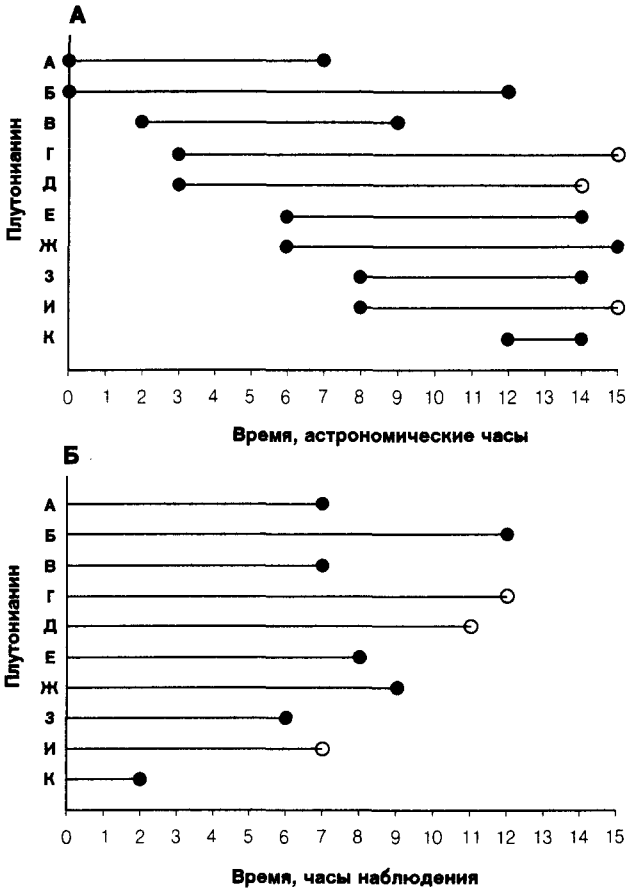


Рис. 11.1. Продолжительность жизни плутониан после начала пассивного курения. **А.** Ход исследования показан в обычной шкале времени. Жизнь плутонианина после начала пассивного курения представлена горизонтальным отрезком. Левый конец отрезка — это начало наблюдения. На правом конце отрезка — черный или белый кружок. Черный кружок означает, что плутонианин умер и, таким образом, продолжительность его жизни нам известна. Белый кружок означает, что исследование закончилось до его смерти либо он куда-то уехал — словом, *выбыл* из-под наблюдения. Относительно выбывших нам известно только, что они прожили *не меньше* определенного срока. **Б.** Ход исследования показан так, как будто все плутониане начали наблюдаться одновременно. Теперь на шкале времени не астрономические часы, а часы наблюдения. Такое представление данных облегчит нам дальнейшие расчеты.

прекратилось ли исследование, уехали они куда-то, — всех их мы будем называть *выбывшими*.

На рис. 11.1Б изображены те же данные, что и на рис. 11.1А. Теперь отрезки, соответствующие периоду наблюдения за каждым плутонианином, расположены так, как если бы все наблюдения были начаты в один момент. Это представление данных более удобно. Теперь сразу видно, кто сколько прожил после начала пассивного курения. Кружок на правом конце каждого из отрезков показывает, умер плутонианин за время наблюдения (кружок закрашен) или выбыл (кружок не закрашен).

Если бы продолжительность наблюдения была одинаковой, мы могли бы рассчитать долю выживших и применить методы, описанные в гл. 5. Однако поскольку участники входили в группу наблюдения на разных сроках исследования, это условие не выполняется. Если бы все наблюдаемые умерли, то можно было бы применить методы, изложенные в гл. 2 или 10. Однако и этого не произошло, как это обычно и бывает в исследованиях такого рода. Для анализа выживаемости нужны новые методы. Прежде чем с ними познакомиться, сформулируем требования, которым должны удовлетворять все исследования выживаемости.

- Для всех наблюдаемых известно время начала наблюдения.
- Для всех наблюдаемых известно время окончания наблюдения, а также — умер он или выбыл.
- Выбор наблюдаемых произведен случайно.

Для начала мы научимся строить кривую выживаемости, а затем перейдем к оценке статистической значимости различий кривых выживаемости.

КРИВАЯ ВЫЖИВАЕМОСТИ

Кривая выживаемости задает вероятность пережить любой из моментов времени после некоторого начального события. Эту вероятность обычно называют просто *выживаемостью*. В примере, который мы сейчас разбираем, кривая выживаемости применяется для изучения продолжительности жизни. Однако кривыми такого рода можно описать продолжительность самых разнообразных процессов. Тогда в качестве исхода будет выступать

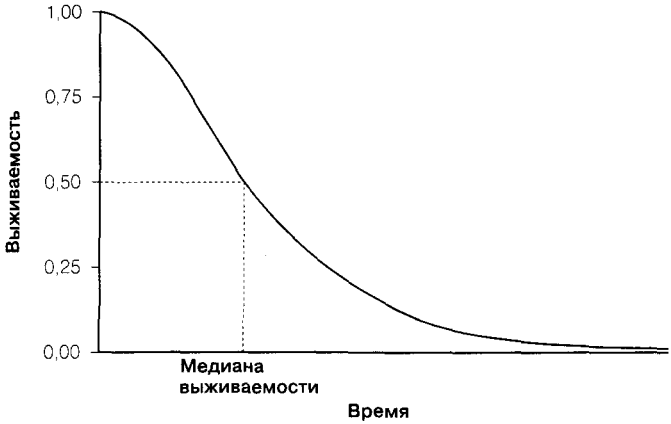


Рис. 11.2. Типичная кривая выживаемости. В начале значение функции выживаемости $S(t)$, естественно, равно 1. В дальнейшем оно уменьшается, постепенно приближаясь к нулю. Время, за которое значение функции выживаемости достигает значения 0,5, называется медианой выживаемости.

не смерть, а другое интересующее нас событие, не всегда нежелательное. Например, можно изучать срок лечения какого-либо заболевания (исход — ремиссия), длительность лечения бесплодия или эффективность контрацепции (исход в обоих случаях — наступление беременности), долговечность протеза (исход — поломка).

Для начала, как всегда, рассмотрим кривую выживаемости для *совокупности*. Такая кривая получилась бы, если бы мы проследили судьбу всех плутониан от рождения до смерти. Выживаемость к моменту времени t обозначим $S(t)$. Дадим определение.

Выживаемость $S(t)$ — это вероятность прожить более t с момента начала наблюдения.

Для совокупности эта вероятность выражается формулой:

$$S(t) = \frac{\text{Число переживших момент } t}{\text{Объем совокупности}}.$$

Типичная кривая выживаемости изображена на рис. 11.2. Понятно, что в точке 0, соответствующей начальному моменту, например моменту рождения, выживаемость равна 1. Затем кривая

постепенно понижается и, начиная с некоторой точки, становится равной 0. Возраст, до которого доживает ровно половина совокупности, называется *медианой выживаемости*.

Наша цель состоит в том, чтобы оценить выживаемость по выборке. Никакого другого способа ее оценки не существует. Если бы не выбывшие, это было бы очень просто:

$$\hat{S}(t) = \frac{\text{Число переживших момент } t}{\text{Объем выборки}}.$$

В тех случаях, когда имеет место выбывание (а это бывает почти всегда), мы не сможем воспользоваться этой формулой. Вместо этого поступим следующим образом. Для каждого момента времени, когда произошла хотя бы одна смерть, оценим вероятность *пережить этот момент*. Такой оценкой будет отношение числа переживших этот момент к числу наблюдавшихся к этому моменту. Тогда, согласно правилу умножения вероятностей, вероятность пережить некоторый момент времени для каждого вступившего в исследование будет равна произведению этих оценок от нулевого до данного момента. Рассмотрим эту процедуру более подробно на примере плутонианских пассивных курильщиков.

Будем считать, что все начали наблюдаться в момент времени $t = 0$, и от этого момента будем отсчитывать все сроки (рис. 11.1Б). Расположим плутониан по возрастанию длительности наблюдения (табл. 11.1) и укажем саму эту длительность во второй колонке таблицы. Длительность наблюдения выбывших плутониан пометим знаком «+» — это будет означать, что плутонианин прожил более такого-то срока, а на сколько — неизвестно. Первый плутонианин (К) умер через 2 часа, второй (З) — через 6 часов после начала наблюдения. На 7-м часу умерли двое — А и В, на этом же сроке выбыл из-под наблюдения плутонианин И.

Первый плутонианин умер в 2 часа. Наблюдались в это время все 10 плутониан. Значит, вероятность умереть в 2 часа — $d_2/n_2 = 1/10 = 0{,}1$. Соответственно, вероятность *не умереть* в 2 часа *для тех, кто дожил до этого времени*:

$$f_2 = 1 - \frac{d_2}{n_2} = 1 - \frac{1}{10} = \frac{9}{10} = 0{,}900.$$

Таблица 11.1. Результаты исследования продолжительности жизни плутониан после начала пассивного курения

	Время	Наблюдалось к моменту t	Умерло в момент t
Плутонианин	t	n_t	d_t
К	2	10	1
З	6	9	1
А и В	7	8	2
И	7+		–
Е	8	5	1
Ж	9	4	1
Д	11+		
Б	12	2	1
Г	12+		–

Следующий плутонианин умер в 6 часов. Наблюдалось к этому времени 9 плутониан. Для доживших до 6 часов вероятность умереть в 6 часов — $d_6/n_6 = 1/9 = 0{,}111$, а вероятность *не умереть* в 6 часов

$$f_6 = 1 - \frac{d_6}{n_6} = 1 - \frac{1}{9} = \frac{8}{9} = 0{,}889.$$

Теперь мы можем оценить вероятность, что плутонианин проживет более 6 часов, то есть $\hat{S}(6)$. Прожить более 6 часов — это значит не умереть в 2 часа и не умереть в 6 часов. То есть, по правилу умножения вероятностей,

$$\hat{S}(6) = f_2 \times f_6 = 0{,}900 \times 0{,}889 = 0{,}800.$$

Уже рискуя надоесть читателю однообразными рассуждениями, перейдем к следующему печальному событию. В 7 часов умерло сразу 2 плутонианина, наблюдалось к этому времени 8. Имеем

$$f_7 = 1 - \frac{d_7}{n_7} = 1 - \frac{2}{8} = \frac{6}{8} = 0{,}750,$$

$$\hat{S}(7) = f_2 \times f_6 \times f_7 = 0{,}900 \times 0{,}889 \times 0{,}750 = 0{,}600.$$

Внимательному читателю может показаться, что мы зря усложняем дело. Действительно, приведя сложные выкладки, мы получили то, что и так было очевидно: если через 7 часов умерло четверо из десяти плутониан, то дольше 7 часов прожило шестеро и выживаемость составляет $\hat{S}(7) = 6/10 = 0{,}600$.

Еще терпение! До сих пор у нас не было выбывших, поэтому результаты и совпадают. Посмотрим, что будет в 8 часов. В 8 часов умирает плутонианин Е. Наблюдаются к этому времени 5 плутониан (4 умерли, 1 выбыл: $10 - 4 - 1 = 5$).

$$f_8 = 1 - \frac{d_8}{n_8} = 1 - \frac{1}{5} = \frac{4}{5} = 0{,}800,$$

$$\hat{S}(8) = f_2 \times f_6 \times f_7 \times f_8 = 0{,}900 \times 0{,}889 \times 0{,}750 \times 0{,}800 = 0{,}480.$$

Если бы мы считали «долю выживших» старым способом, мы бы получили для $\hat{S}(8)$ оценку 0,5. В дальнейшем, чем больше будет выбывших, тем больше будет и расхождение.

Описанная процедура называется расчетом выживаемости моментным методом, или методом Каплана—Мейера.

Математическое выражение моментного метода:

$$\hat{S}(t) = \Pi\left(1 - \frac{d_t}{n_t}\right),$$

где d_t — число умерших в момент t, n_t — число наблюдавшихся к моменту t, П (большая греческая буква «пи») — символ произведения. В данном случае она означает, что надо перемножить значения $(1 - d_t/n_t)$ для всех моментов, когда произошла хотя бы одна смерть. В принципе, можно перемножать и по остальным моментам, однако, если $d_t = 0$, то $(1 - d_t/n_t) = 1$, а умножение на единицу на результате никак не скажется.

В табл. 11.2 расчет выживаемости моментным методом приведен полностью. Теперь мы можем представить результаты исследования выживаемости плутониан после начала пассивного курения в виде графика (рис. 11.3). Точки на графике соответствуют моментам, когда умер хотя бы один из наблюдавшихся. Эти точки обычно соединяют ступенчатой линией. В момент времени 0 выживаемость со-

Таблица 11.2. Расчет кривой выживаемости плутониан после начала пассивного курения

Плутониан	Время	Наблюдалось к моменту t	Умерло в момент t	Доля переживших момент t	Выживаемость
	t	n_t	d_t	$f_t = 1 - \frac{d_t}{n_t}$	$\hat{S}(t)$
К	2	10	1	0,900	0,900
З	6	9	1	0,889	0,800
А и В	7	8	2	0,750	0,600
И	7+	—	—		
Е	8	5	1	0,800	0,480
Ж	9	4	1	0,750	0,360
Д	11+	—	—		
Б	12	2	1	0,500	0,180
Г	12+	—	—		

ставляет 1,0, затем постепенно снижается. В данном случае умерли не все наблюдавшиеся — поэтому нуля линия не достигает.

Медиана выживаемости

Наиболее полная характеристика выживаемости — это кривая выживаемости, которую мы только что построили. Хотелось бы, однако, иметь и обобщенный показатель, характеризующий выживаемость в виде одного числа. Распределение по продолжительности жизни, как правило, асимметрично, поэтому лучше всего тут подходит медиана. Определение медианы выживаемости для совокупности мы дали выше. Для выборки медиана выживаемости определяется как *наименьшее время, для которого выживаемость меньше 0,5.*

Чтобы определить медиану выживаемости, нужно построить кривую выживаемости и посмотреть, где она впервые опускается ниже 0,5. Например, на рис. 11.3 это произошло в 8 часов. Аналогично медиане могут быть вычислены другие процентили выживаемости.

Если число умерших меньше половины числа наблюдаемых, медиану определить невозможно.

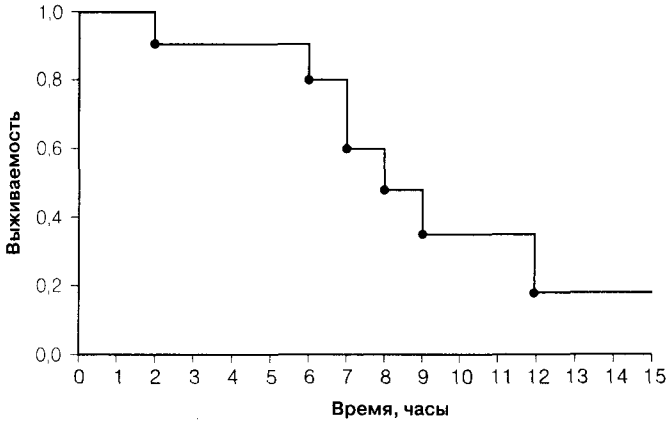


Рис. 11.3. Эта кривая выживаемости плутониан после начала пассивного курения рассчитана по данным с табл. 11.1; ход вычислений показан в табл. 11.2. Кривая представляет собой ступенчатую линию, каждой ступеньке соответствует момент смерти одного или нескольких плутониан.

Стандартная ошибка и доверительные интервалы выживаемости

Как всегда при исследовании выборки, полученная нами кривая выживаемости на самом деле представляет собой *оценку* кривой выживаемости. Если бы мы могли определить продолжительность жизни всех плутониан, подвергшихся пассивному курению, мы получили бы гладкую кривую вроде изображенной на рис. 11.2. Оценку точности приближения дает стандартная ошибка выживаемости; ее можно рассчитать по формуле Гринвуда*:

$$s_{\hat{S}(t)} = \hat{S}(t)\sqrt{\sum \frac{d_{t_i}}{n_{t_i}(n_{t_i} - d_{t_i})}},$$

где сумма берется по всем моментам t_i, от нуля до t включительно. На примере данных по выживаемости плутониан после на-

* Вывод этой формулы можно найти в: D. Collett. Modelling survival data in medical research. Chapman and Hall, London, 1994, pp. 22—26.

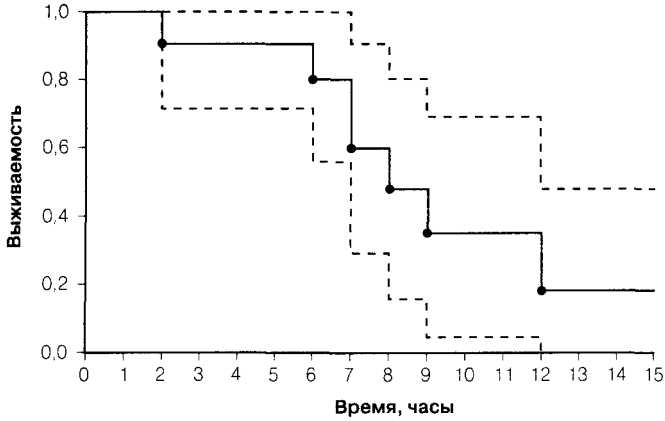


Рис. 11.4. Кривая выживаемости плутониан после начала пассивного курения и ее 95% доверительная область (ход вычислений показан в табл. 11.3). Границы доверительной области показаны пунктиром.

чала пассивного курения рассчитаем стандартную ошибку выживаемости для 7 часов:

$$s_{\hat{S}(7)} = 0{,}600\sqrt{\frac{1}{10(10-1)} + \frac{1}{9(9-1)} + \frac{2}{8(8-2)}} = 0{,}155.$$

В табл. 11.3 приведены значения стандартной ошибки для вычисленных по табл. 11.1 оценок функции выживаемости.

В гл. 7 было показано, как с помощью стандартной ошибки вычислить доверительные интервалы для долей. Точно так же ее используют для вычисления доверительного интервала для выживаемости. Напомним, что $100(1-\alpha)$-процентный доверительный интервал для доли p задается неравенством

$$\hat{p} - z_\alpha s_{\hat{p}} < p < \hat{p} + z_\alpha s_{\hat{p}},$$

где z_α — двустороннее критическое значение для стандартного нормального распределения, α — уровень значимости, $\hat{p}$ — выборочное значение доли, $s_{\hat{p}}$ — стандартная ошибка для этой доли. Доверительный интервал для выживаемости в момент t определяется аналогично:

$$\hat{S}(t) - z_\alpha s_{\hat{S}(t)} < S(t) < \hat{S}(t) + z_\alpha s_{\hat{S}(t)}.$$

Обычно определяют 95% доверительный интервал. Тогда $\alpha = 1 - 0{,}95 = 0{,}05$. Соответствующее значение $z_\alpha = 1{,}960$. Дальнейшие вычисления показаны в таблице 11.3. Отложив на графике доверительные интервалы (рис. 11.4), мы увидим расширяющийся «рукав» — доверительную область для выживаемости. Причина расширения доверительной области понятна: чем меньше остается наблюдаемых, тем больше ошибка.

Как вы помните, при расчете доверительных интервалов для долей существовало ограничение на использование нормального распределения. Аналогичное ограничение существует и при оценке доверительных интервалов для функции выживаемости. Дело в том, что нормальное приближение вносит сильные искажения, когда функция выживаемости принимает значение, близкое к граничным — к 0 или 1. В этом случае доверительный интервал должен быть *несимметричен* относительно *p*. (См. также рис. 7.4 и соответствующее обсуждение в гл. 7.) Приведенная выше формула, напротив, дает симметричную оценку, которая может выйти за граничные значения 1 и 0. Простейший способ подправить такую оценку состоит в том, чтобы значения, большие единицы, заменить на единицу, а меньшие нуля — на ноль. Существует и несколько более сложный способ, он позволяет рассчитать доверительный интервал точнее. Возьмем двойной логарифм $\ln[-\ln\hat{S}(t)]$. В отличие от $\hat{S}(t)$, эта величина не должна лежать в пределах от 0 до 1. Затем вычислим для нее стандартную ошибку, после чего вернемся к исходной функции $\hat{S}(t)$. Стандартная ошибка для логарифмической формы выживаемости:

$$s_{\ln[-\ln\hat{S}(t)]} = \sqrt{\frac{1}{[\ln\hat{S}(t)]^2}\sum\frac{d_t}{n_t(n_t - d_t)}}.$$

Тогда $100(1-\alpha)$-процентный доверительный интервал для $S(t)$ определяется неравенством:

$$\hat{S}(t)^{\exp\left(-z_\alpha s_{\ln[-\ln\hat{S}(t)]}\right)} < S(t) < \hat{S}(t)^{\exp\left(+z_\alpha s_{\ln[-\ln\hat{S}(t)]}\right)}.$$

Таблица 11.3. Расчет стандартной ошибки и 95% доверительного интервала кривой выживаемости плутониан после начала пассивного курения

Плутониан	Время	Наблюдалось в момент t	Умерло в момент t	Доля переживших момент t	Выживаемость		Стандартная ошибка	95% доверительный интервал	
								нижняя граница	верхняя граница
	t	n_t	d_t	$f_t = 1 - \frac{d_t}{n_t}$	$\hat{S}(t)$	$\frac{d_t}{n_t(n_t - d_t)}$	$s_{\hat{S}(t)}$		
К	2	10	1	0,900	0,900	0,011	0,095	0,716	1,000*
З	6	9	1	0,889	0,800	0,014	0,126	0,553	1,000*
А и В	7	8	2	0,750	0,600	0,042	0,155	0,296	0,904
И	7+	—	—						
Е	8	5	1	0,800	0,480	0,050	0,164	0,159	0,801
Ж	9	4	1	0,750	0,360	0,083	0,161	0,044	0,676
Д	11+	—	—						
Б	12	2	1	0,500	0,180	0,500	0,151	0,000*	0,475
Г	12+	—	—						

* Вычисленные значения были больше 1 либо меньше 0.

СРАВНЕНИЕ ДВУХ КРИВЫХ ВЫЖИВАЕМОСТИ

В клинических исследованиях часто возникает необходимость сравнить выживаемость разных групп больных. Посмотрим, как это делается в случае двух групп*. Нулевая гипотеза состоит в том, что в обеих группах выживаемость одинакова. Если бы не было выбывания и все больные наблюдались равное время, нам бы подошел анализ таблиц сопряженности (см. гл. 5). Если бы все больные наблюдались вплоть до смерти, можно было бы сравнить выживаемость в обеих группах с помощью изложенных в гл. 10 непараметрических методов, например рангового критерия Манна—Уитни или метода Крускала—Уоллиса. В реальной жизни подобные ситуации редки, и, как мы уже говорили, выбывание практически неизбежно. Для сравнения кривых выживаемости нужны специальные методы. Первым мы рассмотрим так называемый *логранговый* критерий.

Он основан на следующих трех допущениях.

- Две сравниваемые выборки независимы и случайны.
- Выбывание в обеих выборках одинаково.
- Функции выживаемости связаны соотношением: $S_2(t) = [S_1(t)]^{\Psi}$.

Величина Ψ («пси») называется отношением смертности. Если $\Psi = 1$, то кривые выживаемости совпадают. Если $\Psi < 1$, люди во 2-й выборке умирают позже, чем в 1-й. И наоборот, если $\Psi > 1$, позже умирают в 1-й выборке.

Трансплантация костного мозга при остром лимфобластном лейкозе взрослых

При остром лимфобластном лейкозе мутация предшественника лимфоцитов приводит к появлению клона лейкозных клеток, способных неограниченно делиться. В отличие от обычных лимфоцитов, лейкозные клетки функционально неактивны и не обладают защитными свойствами. Размножаясь в костном мозге, они подавляют нормальное кроветворение, в результате развива-

---

* Существуют методы сравнения и нескольких групп. Останавливаться на них мы не будем: они основаны на тех же принципах, но требуют громоздких вычислений.

ются иммунодефицит, анемия и тромбоцитопения. Без лечения острый лимфобластный лейкоз низбежно приводит к смерти.

Задача лечения — полностью уничтожить лейкозные клетки. Этого можно достичь с помощью облучения и химиотерапии. Однако при этом уничтожаются и нормальные кроветворные клетки. Чтобы компенсировать это побочное действие лечения, используют трансплантацию костного мозга. Для трансплантации лучше всего подходит костный мозг близкого родственника (*аллотрансплантация*). К сожалению, не всегда есть у кого его взять. Поэтому применяется и другой способ, так называемая *аутотрансплантация*, когда костный мозг берут у самого больного. Из полученного костного мозга специальныи методами удаляют лейкозные клетки и, по завершении курса лучевой и химиотерапии, его вновь вводят больному. Н. Вей с соавт. сравнили выживаемость после ауто- и аллотрансплантации*.

В исследование включали больных старше 15 лет с подтвержденным диагнозом острого лимфобластного лейкоза после достижения первой полной ремиссии. Больным, у которых не было подходящих родственников, проводили аутотрансплантацию (1-я группа), остальным — аллотрансплантацию (2-я группа). Исследование продолжалось 11 лет.

Полученные данные представлены в табл. 11.4. Как и ранее, выбывшие помечены знаком «+». В табл. 11.5 приведен расчет выживаемости для каждой из групп. Соответствующие кривые показаны на рис. 11.5. Выживаемость в 1-й группе хуже, чем во 2-й. Вопрос состоит в том, какова вероятность получить подобное различие выживаемости случайно.

Перейдем к построению логрангового критерия. Ход вычислений показан в табл. 11.6 (выбывших в таблице нет, показаны

* N. Vey, D. Blaise, A. Stoppa et al. Bone marrow transplantation in 63 adult patients with acute lymphoblastic leukemia in first complete remission. *Bone Marrow Transplantation*, 14:383—388, 1994. В этом исследовании выборки не были случайными: в группу аутотрансплантации попадали больные, у которых не нашлось близких родственников. Авторы указывают, однако, что по основным прогностическим признакам группы были сходны. Это лучшее, что можно сделать, когда рандомизация невозможна. Дальнейшее обсуждение этой темы вы найдете в гл. 12.

Таблица 11.4. Продолжительность жизни после трансплантации костного мозга

Аутотрансплантация (1-я группа, $n = 33$)		Аллотрансплантация (2-я группа, $n = 21$)	
Месяцы после пересадки	Число смертей или выбытий	Месяцы после пересадки	Число смертей или выбытий
1	3	1	1
2	2	2	1
3	1	3	1
4	1	4	1
5	1	6	1
6	1	7	1
7	1	12	1
8	2	15+	1
10	1	20+	1
12	2	21+	1
14	1	24	1
17	1	30+	1
20+	1	60+	1
27	2	85+	2
28	1	86+	1
30	2	87+	1
36	1	90+	1
38+	1	100+	1
40+	1	119+	1
45+	1	132+	1
50	3		
63+	1		
132+	2		

только моменты наступления смерти). Как видим, спустя месяц после трансплантации в 1-й группе умерли 3 из 33 больных, во второй — 1 из 21 больного. Каким бы было число умерших при условии справедливости нулевой гипотезы? Рассчитаем *ожидаемые числа* умерших, подобно тому, как мы это делали в гл. 5.

В первый месяц в обеих группах умерло $3+1=4$ из $33+21=54$ больных. Таким образом, смертность в обеих группах составила $4/54 = 0{,}074 = 7{,}4\%$. Если бы, согласно нулевой гипотезе, меж-

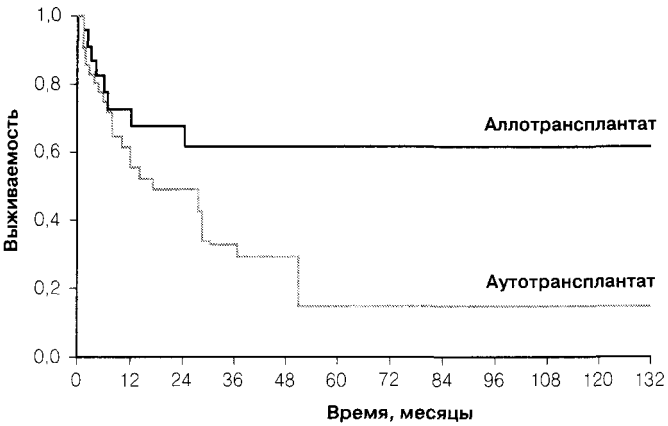


Рис. 11.5. Выживаемость при остром лимфобластном лейкозе взрослых после трансплантации костного мозга. Костный мозг брали у брата или сестры, совместимых по HLA (аллотрансплантация), либо у самого больного (аутотрансплантация). Данные приведены в табл. 11.4, ход вычислений — в табл. 11.5.

групповых различий не было, то в первой группе умерло бы $33 \times 0{,}074 = 2{,}442$ человека. Это число довольно близко к 3 — наблюдаемому числу умерших. Если нулевая гипотеза справедлива, ожидаемые и наблюдаемые числа и дальше будут близки.

Найдем таким же способом ожидаемое число умерших в 1-й группе в каждый из месяцев, когда кто-нибудь умирал хотя бы в одной группе.

$$E_{1t} = \frac{n_{1t} d_{\text{об}\,t}}{n_{\text{об}\,t}},$$

где E_{1t} — ожидаемое число умерших в первой группе в момент времени t; n_{1t} — число наблюдавшихся в 1-й группе к этому моменту, $d_{\text{об}\,t}$ — общее число смертей в этот момент в обеих группах, $n_{\text{об}\,t}$ — общее число наблюдавшихся к этому моменту.

Пока что не совсем понятно, как мы учитываем выбывших — ведь в формуле и в табл. 11.6 их число не фигурирует. Выбывшие учитываются косвенно — влияя на число наблюдавшихся. Например, во 2-й группе на сроке 17 мес никто не умер, однако число на-

Таблица 11.5. Вычисление выживаемости по данным из табл. 11.4

Аутотрансплантация				
Месяц	Умерли (выбыли) в месяц t	Наблюдались к началу месяца t	Доля переживших месяц t	Выживаемость
t	d_t	n_t	$f_t = 1 - \frac{d_t}{n_t}$	$\hat{S}(t)$
1	3	33	0,909	0,909
2	2	30	0,933	0,848
3	1	28	0,964	0,818
4	1	27	0,963	0,788
5	1	26	0,962	0,757
6	1	25	0,960	0,727
7	1	24	0,958	0,697
8	2	23	0,913	0,636
10	1	21	0,952	0,606
12	2	20	0,900	0,545
14	1	18	0,944	0,515
17	1	17	0,941	0,485
20+	1	16		
27	2	15	0,867	0,420
28	1	13	0,923	0,388
30	2	12	0,833	0,323
36	1	10	0,900	0,291
38+	1	9		
40+	1	8		
45+	1	7		
50	3	6	0,500	0,145
63+	1	3		
132+	2	2		

блюдавшихся уменьшилось с 13 до 11 человек. Это произошло потому, что 3 больных на этом сроке выбыли из-под наблюдения.

Просуммируем разности наблюдаемого и ожидаемого числа умерших:

$$U_L = \Sigma(d_{1t} - E_{1t}).$$

Таблица 11.5. Окончание

Аллотрансплантация				
Месяц	Умерли (выбыли) в месяц t	Наблюдались к началу месяца t	Доля переживших месяц t	Выживаемость
t	n_t	d_t	$f_t = 1 - \frac{d_t}{n_t}$	$\hat{S}(t)$
1	1	21	0,952	0, 952
2	1	20	0,950	0,905
3	1	19	0,947	0,857
4	1	18	0,944	0,810
6	1	17	0,941	0,762
7	1	16	0,938	0,714
12	1	15	0,933	0,667
15+	1	14		
20+	1	13		
21+	1	12		
24	1	11	0,909	0,606
30+	1	10		
60+	1	9		
85+	2	8		
86+	1	6		
87+	1	5		
90+	1	4		
100+	1	3		
119+	1	2		
132+	1	1		

Сумма берется по всем моментам t, когда хотя бы одна смерть наступала в любой из двух групп. Как видно из табл. 11.6, в нашем примере $U_L = 6{,}572$. Если U_L достаточно велико, гипотезу об отсутствии различий выживаемости следует отклонить.

U_L приближенно подчиняется нормальному распределению со стандартным отклонением

$$s_{U_L} = \sqrt{\sum \frac{n_{1t} n_{2t} d_{\text{об}t} (n_{\text{об}t} - d_{\text{об}t})}{n_{\text{об}t}^2 (n_{\text{об}t} - 1)}},$$

Таблица 11.6. Вычисление логрангового критерия по данным из табл. 11.4

Месяц	Аутотрансплантация (1-я группа)		Аллотрансплантация (2-я группа)		Объединенная группа		Ожидаемое число смертей в 1-й группе	Слагаемое для U_L	Слагаемое для $s^2_{U_L}$
	Умерли в месяц t	Наблюдались к началу месяца t	Умерли в месяц t	Наблюдались к началу месяца t	Умерли в месяц t	Наблюдались к началу месяца t			
t	d_{1t}	n_{1t}	d_{2t}	n_{2t}	$d_{об\,t} = d_{1t} + d_{2t}$	$n_{об\,t} = n_{1t} + n_{2t}$	$E_{1t} = n_{1t}\frac{d_{об\,t}}{n_{об\,t}}$	$d_{1t} - E_{1t}$	см. текст
1	3	33	1	21	4	54	2,444	0,556	0,897
2	2	30	1	20	3	50	1,800	0,200	0,691
3	1	28	1	19	2	47	1,191	−0,191	0,471
4	1	27	1	18	2	45	1,200	−0,200	0,469
5	1	26	0	17	1	43	0,605	0,395	0,239
6	1	25	1	17	2	42	1,190	−0,190	0,470
7	1	24	1	16	2	40	1,200	−0,200	0,468
8	2	23	0	15	2	38	1,211	0,789	0,465
10	1	21	0	15	1	36	0,583	0,417	0,243
12	2	20	1	15	3	35	1,714	0,286	0,691
14	1	18	0	14	1	32	0,563	0,438	0,246
17	1	17	0	13	1	30	0,567	0,433	0,248
24	0	15	1	11	1	26	0,576	−0,577	0,241
27	2	15	0	10	2	25	1,200	0,800	0,460
28	1	13	0	10	1	23	0,565	0,435	0,246
30	2	12	0	10	2	22	1,091	0,909	0,472
36	1	10	0	9	1	19	0,526	0,474	0,249
50	3	6	0	9	3	15	1,200	1,800	0,617
								$U_L = 6{,}572$	$s^2_{U_L} = 7{,}883$

где, как и раньше, сумма берется по всем моментам t, когда наблюдалась хотя бы одна смерть*. В последнем столбце табл. 11.6 приведены слагаемые $s^2_{U_L}$. Их сумма составляет 7,884, таким образом, $s_{U_L} = \sqrt{7,883} = 2,808$.

Разделив значение U_L на его стандартную ошибку (то есть стандартное отклонение выборочного распределения), получим

$$z = \frac{U_L}{s_{U_L}} = \frac{6,572}{2,808} = 2,341.$$

Распределение z приблизительно нормально, поэтому сравним эту величину с критическим значением для стандартного нормального распределения (см. последнюю строку табл. 4.1)**. Критическое значение для уровня значимости 2% в случае нормального распределения равно 2,326, то есть меньше полученного нами. Поэтому мы отклоняем нулевую гипотезу об отсутствии различий в выживаемости.

В заключение заметим, что совершенно неважно, для какой именно из групп вычисляется U_L. Для 2-й группы U_L равна по абсолютной величине U_L для 1-й, но имеет противоположный знак.

Поправка Йейтса для логрангового критерия

Мы уже сталкивались с ситуацией, когда дискретное распределение приближенно описывается нормальным, которое по сути своей непрерывно. Практически это приводит к излишней «мягкости» критерия: мы несколько чаще, чем следовало бы, отвергаем нулевую гипотезу. Чтобы компенсировать влияние дискретности, применяют поправку Йейтса. В случае логрангового критерия это делается таким образом:

---

* Вывод этой формулы приведен в книге D. Collett. Modelling survival data in medical research. Chapman & Hall, London, 1994, pp. 40—42.

** Иногда вместо U_L / s_{U_L} вычисляют $U_L^2 / s^2_{U_L}$. Эта величина имеет распределение χ^2 с одной степенью свободы. Оба варианта критерия приводят к одному результату. Точно так же к обоим вариантам в равной мере применима поправка Йейтса, о чем ниже.

$$z = \frac{|U_L| - \frac{1}{2}}{s_{U_L}}.$$

Для примера, который мы рассматриваем:

$$z = \frac{6{,}572 - 0{,}5}{2{,}808} = 2{,}162.$$

В результате применения поправки Йейтса величина z уменьшилась с 2,342 до 2,162, однако она по-прежнему больше 1,960 — критического значения для уровня значимости 0,05. В данном случае поправка Йейтса не изменила общий вывод — различия выживаемости статистически значимы.

КРИТЕРИЙ ГЕХАНА

Существует другой метод сравнения выживаемости. Он называется *критерием Гехана* и представляет собой обобщение критерия Уилкоксона. Он не требует постоянства отношения смертности, но на его результаты слишком сильно влияет число ранних смертей.

Критерий Гехана вычисляют так. Каждого больного из 1-й группы сравнивают с каждым больным из 2-й группы. Результат сравнения оценивают как +1, если больной из 1-й группы *наверняка* прожил дольше, –1, если он *наверняка* прожил меньше, и 0, если невозможно наверняка сказать, кто из них прожил дольше. Последнее возможно в трех случаях: если оба выбыли, если один выбыл до того, как другой умер, и если время наблюдения одинаково.

Результаты сравнения для каждого больного суммируют; эту сумму мы обозначим h. В свою очередь сумма всех h дает величину U_W, стандартная ошибка которой определяется по формуле:

$$s_{U_W} = \sqrt{\frac{n_1 n_2 \sum h^2}{(n_1 + n_2)(n_1 + n_2 - 1)}}.$$

И наконец, вычисляют

$$z = \frac{U_W}{s_{U_W}}.$$

Полученное значение нужно сравнить с критическим значением стандартного нормального распределения (см. последнюю строку табл. 4.1).

Поправка Йейтса применяется к критерию Гехана точно так же, как к логранговому критерию.

Какой критерий предпочесть? Логранговый критерий предпочтительнее критерия Гехана, если справедливо предположение о постоянном отношении смертности: $S_2(t) = [S_1(t)]^{\Psi}$. Установить, выполняется ли это условие, можно, нарисовав графики $\ln[-\ln\hat{S}_1(t)]$ и $\ln[-\ln\hat{S}_2(t)]$ — они должны быть параллельны. Во всяком случае, кривые выживаемости не должны пересекаться.

ЧУВСТВИТЕЛЬНОСТЬ И ОБЪЕМ ВЫБОРКИ

Как вы помните, чувствительность любого критерия зависит от трех величин — величины различия, которую он должен уловить, уровня значимости и численности групп. И наоборот, численность групп, необходимая для того, чтобы уловить различия, не меньшие некоторой величины, определяется уровнем значимости и необходимой чувствительностью. Логранговый критерий не является исключением. Чем меньшее различие выживаемости нужно выявить, тем большим должно быть число наблюдений.

Для простоты ограничимся случаем равной численности групп*. Заметим, что, как и всегда, при заданном числе обследованных именно равная численность групп обеспечивает максимальную чувствительность.

Прежде всего следует оценить необходимое число исходов (смертей, рецидивов и т. д.). Имеем

$$d = (z_\alpha + z_{1-\beta})^2 \left(\frac{1+\Psi}{1-\Psi}\right)^2,$$

---

* Вывод формул можно найти в работе L. S. Freedman. Tables of number of patients required in clinical trials using the log-rank test. *Statist. Med.*, 1:121—129, 1982.

где Ψ — отношение смертности, а z_α и $z_{1-\beta}$ — соответствующие α и $1-\beta$ значения стандарного нормального распределения (их можно найти в последней строке табл. 4.1). Как определить Ψ? Поскольку при всех t соблюдается равенство $S_2(t) = [S_1(t)]^\Psi$, этот параметр можно оценить как

$$\Psi = \frac{\ln S_2(\infty)}{\ln S_1(\infty)},$$

где $S_1(\infty)$ и $S_2(\infty)$ — выживаемость в 1-й и 2-й группах к концу наблюдения. Теперь мы можем найти n — численность каждой из групп:

$$n = \frac{d}{2 - S_1(\infty) - S_2(\infty)}.$$

Таким образом, по ожидаемым долям доживших до завершения эксперимента мы можем найти объем n каждой из выборок.

Рассмотрим пример. Пусть мы предполагаем, что выживаемость должна повыситься с 30 до 60% или более. Эти различия мы хотим выявить с вероятностью 80% (то есть чувствительность $1-\beta = 0{,}8$). Уровень значимости $\alpha = 0{,}05$. По табл. 4.1 находим $z_\alpha = z_{0,05} = 1{,}960$ и $z_{1-\beta} = z_{0,80} = 0{,}840$.

Оценив

$$\Psi = \frac{\ln S_2(\infty)}{\ln S_1(\infty)} = \frac{\ln 0{,}6}{\ln 0{,}3} = \frac{-0{,}511}{-1{,}203} = 0{,}425,$$

подставим значения в формулу для числа исходов

$$d = (z_\alpha + z_{1-\beta})^2 \left(\frac{1+\Psi}{1-\Psi}\right)^2 = (1{,}960 + 0{,}840)^2 \left(\frac{1+0{,}425}{1-0{,}425}\right)^2 = 48{,}1$$

и рассчитаем численность каждой группы:

$$n = \frac{d}{2 - S_1(\infty) - S_2(\infty)} = \frac{48{,}1}{2 - 0{,}3 - 0{,}6} = 43{,}7.$$

Итак, в каждую из групп должно входить по 44 человека.

ЗАКЛЮЧЕНИЕ

К анализу выживаемости неприменимы обычные способы оценки различий, такие, как сравнение долей и средних величин. Необходимы методы, учитывающие выбывание, которое неизбежно имеет место в исследованиях такого рода. Мы рассмотрели простейшие методы сравнения выживаемости, а именно сравнение выживаемости в двух группах. Соответствующие методы для произвольного числа групп основаны примерно на тех же принципах. Как логранговый критерий, так и критерий Гехана относятся к непараметрическим — они не исходят из предположения об определенной форме кривой выживаемости. Существуют и параметрические методы анализа выживаемости.

Значение анализа выживаемости чрезвычайно велико. В гл. 4 мы говорили о показателях процесса и показателях результата. Если, например, препарат снижает уровень холестерина, то это еще не значит, что он позволяет продлить жизнь больного или отдалить появление стенокардии, — речь, следовательно, идет о *показателе процесса*. Напротив, если доказано, что препарат продлевает жизнь, то речь идет о *показателе результата*, имеющем несомненную клиническую значимость.

Сегодня, когда требования к доказательствам эффективности лечения ужесточаются, изучение выживаемости (и вообще течения заболеваний) приобретает все большее значение. Исследования такого рода, в отличие от простой регистрации показателей процесса, столь же трудны, сколь и необходимы. В следующей главе мы подробнее обсудим разные типы исследований и их роль в медицине.

ЗАДАЧИ

11.1. Амбулаторное лечение пожилых людей дешевле стационарного. Однако позволяет ли амбулаторное наблюдение достаточно надежно выявлять тех, кто нуждается в госпитализации? Для оценки общего состояния пожилого человека предложена так называемая шкала повседневной работы по дому (IADL, Instrumental Activities of Daily Living). Один из разделов иссле-

дования Б. Келлер и Дж. Поттер (B. Keller, J. Potter. Predictors of mortality in outpatient geriatric evaluation and management clinic patients. *J. Gerontology,* 49:M246—M251, 1994) был посвящен изучению прогностической ценности этой шкалы.

В исследование были включены люди примерно одного возраста (средний возраст 78,4 года, стандартное отклонение 7,2 года), разделенные на 2 группы: с высокой и низкой оценкой по шкале повседневной работы по дому. В результате 4-летнего наблюдения были получены следующие данные:

Высокая оценка		Низкая оценка	
Время, мес	Умерли или выбыли	Время, мес	Умерли или выбыли
14	1	6	2
20	2	12	2
24	3	18	4
25+	1	24	1
28	1	26+	1
30	2	28	4
36+	1	32	4
37+	1	34+	2
38	2	36	4
42+	1	38+	3
43+	1	42	3
48	2	46+	2
48+	62	47	3
		48	2
		48+	23

Оцените статистическую значимость различий в выживаемости двух групп.

11.2. Ф. Джирард и соавт. (P. Girard et al. Surgery for pulmonary metastases: who are the 10 years survivors? *Cancer,* 74:2791—2797, 1994) изучили выживаемость 34 больных после резекции легкого по поводу метастазов. Результаты приведены в таблице на следующей странице. Постройте кривую выживаемости и ее 95% доверительную область.

Выживаемость после резекции легкого по поводу метастазов

Месяц после операции	Число умерших и выбывших
1	1
2	1
3	3
4	1
5	1
6	1
7	2
8	1
9	1
10+	1
11+	2
12	2
13	1
15	1
16	3
20	3
21	1
25+	1
28	1
34	1
36+	1
48+	1
56	1
62	1
84	1

11.3. Основная причина детской смертности в Японии — онкологические заболевания. Позволяют ли современные методы лечения продлить жизнь детей? В. Аджики и соавт. (W. Ajiki et al. Survival rates of childhood cancer patients in Osaka, Japan, 1975–1984. *Jpn. J. Cancer Res.,* 86:13–20, 1995) сравнили выживаемость (с момента постановки диагноза) детей с онкологическими заболеваниями в период 1975—1979 гг. с выживаемостью в период 1980—1984 гг.

1975—1979 гг.		1980—1984 гг.	
Время, мес	Умерли или выбыли	Время, мес	Умерли или выбыли
2	3	2	4
4	4	4	1
6	3	6	3
8	4	8	10
10+	1	12	4
12	2	14	3
14	3	18+	1
16+	1	20+	1
18	2	22	2
22+	1	24	1
24	1	30	2
30	2	36	3
36	1	48	2
52+	1	54+	1
54	1	56	2
56	1	60	1
60	1	60+	9
60+	18		

(а) Постройте кривые выживаемости и 95% доверительные интервалы. (б) Найдите медианы выживаемости. (в) Оцените статистическую значимость различий выживаемости. (г) Определите чувствительность логрангового критерия с уровнем значимости $\alpha = 0{,}05$, предполагая, что $S(\infty) = S(60)$. (д) Вычислите общее число смертей и численность групп, при которых чувствительность логрангового критерия составит 0,80 при условии, что $S(\infty)$ снизилась с 0,40 в период 1975—1979 гг. до 0,20 или 0,15 в 1980—1984 гг.

Глава 12

Как построить исследование

Мы познакомились со многими статистическими методами, узнали о принципах, лежащих в их основе, и получили некоторый навык в расчетах. Каждый метод основан на собственной математической модели, и применение его тем успешнее, чем ближе эта модель к действительности. Чтобы правильно выбрать статистической метод, необходимо учитывать прежде всего характер интересующего нас признака (количественный, порядковый или качественный) и тип распределения (нормальное или нет). Ниже мы кратко суммируем все, что узнали о выборе статистического метода. Однако существует еще одно обстоятельство, о котором мы упоминали лишь вскользь, но которое решающим образом влияет на практическую ценность результата исследования. Это представительность выборки. Любой статистический метод исходит из предположения, что выборка извлечена из совокупности *случайно*. Если это условие не выполняется (то есть если выборка непредставительна), никакой, даже самый изощренный статистический метод не даст правильного результата.

Далее, если выборка представительна, то какую совокупность она представляет? Как мы увидим, больные в крупных медицинских центрах, где обычно проводятся клинические испытания, мало напоминают тех, с которыми встречается врач общей практики. И наконец, мы еще раз напомним об опасности эффекта множественных сравнений. Интересно, что этот многоликий враг исследователей в наибольшей степени угрожает самым любознательным из них.

КАКИМ КРИТЕРИЕМ ВОСПОЛЬЗОВАТЬСЯ

В этой книге мы не стремились охватить все статистические методы: многие из них остались вне поля зрения. Так, не были рассмотрены *многофакторные* методы, в которых исследуются результаты одновременного использования нескольких способов лечения или две группы сравниваются по нескольким показателям.

Однако мы выстроили костяк из статистических методов, вокруг которого естественным образом наращиваются более общие. Охватив широкий круг типов задач, внутри каждого типа мы рассмотрели простейшую модель. Встретившись с более сложной задачей того же или сходного типа, вы без труда сами подберете подходящий метод. Тем не менее освоенные нами методы открывают достаточно большие возможности для решения практических задач.

С помощью табл. 12.1 вы легко найдете, каким критерием следует воспользоваться в зависимости от вида исследования и изучавшегося признака (количественный, порядковый или качественный). Виду исследования (применялись ли сравниваемые методы лечения к общей группе больных или каждый испытывался на отдельной группе, равно ли число сравниваемых методов двум и т. д.) соответствуют столбцы таблицы. Строки таблицы определяют, какие признаки изучались — числовые, порядковые или качественные. Данные о выживаемости мы выделили в отдельный тип, поэтому получилось четыре типа данных. Выбор статистического критерия в случае числовых признаков требует пояснения. Если известно, что распределение признака

Таблица 12.1. Каким критерием воспользоваться

	Исследование				
Признак	Две группы	Более двух групп	Одна группа до и после лечения	Одна группа, несколько видов лечения	Связь признаков
Количественный (распределение нормальное*)	Критерий Стьюдента (гл. 4)	Дисперсионный анализ (гл. 3)	Парный критерий Стьюдента (гл. 9)	Дисперсионный анализ повторных измерений (гл. 9)	Линейная регрессия, корреляция или метод Блэнда—Алтмана (гл. 8)
Качественный	Критерий χ^2 (гл. 5)	Критерий χ^2 (гл. 5)	Критерий Мак-Нимара (гл. 9)	Критерий Кокрена (в нашем курсе рассмотрен не был)	Коэффициент сопряженности (в нашем курсе рассмотрен не был)
Порядковый	Критерий Манна—Уитни (гл. 10)	Критерий Крускала—Уоллиса (гл. 10)	Критерий Уилкоксона (гл. 10)	Критерий Фридмана (гл. 10)	Коэффициент ранговой корреляции Спирмена (гл. 8)
Выживаемость	Критерий Гехана (гл. 11)				

* Если совокупность имеет иное распределение, примените аналогичные непараметрические методы.

в совокупности нормально, можно использовать параметрический метод, указанный в таблице (иногда необходимы дополнительные условия, например, в случае дисперсионного анализа требуется равенство дисперсий). Если распределение далеко от нормального, или если у вас нет желания использовать параметрические методы, следует воспользоваться их непараметрическими аналогами.

Табл. 12.1 — это своего рода путеводитель по статистическим критериям. Но прежде чем им воспользоваться, примите во внимание три вещи. Во-первых, обнаружив, что нулевая гипотеза об отсутствии эффекта не может быть отвергнута, выясните почему. Для этого определите чувствительность критерия (гл. 6). Если чувствительность мала, причиной может быть малый объем выборки. Но если чувствительность велика, то эффект действительно отсутствует. Во-вторых, обнаружив *статистически* значимый эффект, не забудьте вычислить его величину и доверительные интервалы (гл. 7 и 8), по которым можно судить о его *клинической* значимости. И, наконец, в-третьих, обязательно попытайтесь понять, в самом ли деле процедура получения данных обеспечивает их представительность, в противном случае все последующие выкладки потеряют смысл. Тема представительности данных заслуживает более подробного рассмотрения.

РАНДОМИЗАЦИЯ И СЛЕПОЙ МЕТОД

Все статистические методы исходят из предположения, что данные извлечены из совокупности *случайно*. Что значит «извлечены случайно»? Это значит, что вероятность оказаться выбранным одинакова для всех членов совокупности. Например, если групп две (экспериментальная и контрольная) и их размеры равны, то любой член совокупности может *равновероятно попасть в любую из групп*.

Обеспечить равную вероятность попадания в любую из групп совсем не так просто, как кажется на первый взгляд. (Предназначенные для этого методы называются рандомизацией, с этим понятием мы встречались в гл. 3.) Прежде всего необходимо исключить всякое влияние человека, что довольно сложно. Врачи,

участвующие в исследовании, изобретательны и хитроумны. Любой недочет в системе рандомизации они обязательно используют, чтобы повлиять на формирование групп. При этом они, скорее всего, будут исходить из самых добрых побуждений; тем не менее такое вмешательство неизбежно приведет к нарушению сопоставимости групп и к искажению результатов исследования. Следует тщательно продумать, как сделать такое влияние невозможным для всех участников исследования, и прежде всего для себя самого.

Задача рандомизации — обеспечить такой подбор больных, чтобы контрольная группа *ни в чем не отличалась* от экспериментальной, кроме метода лечения. Однако этого мало. На этапе оценки результатов вновь появляется пристрастный исследователь. Велика и роль больного, его веры в новый способ лечения. Обоих следует лишить возможности влиять на результаты. Для этого предназначен слепой метод. В идеале это *двойной слепой метод*: ни больной, ни наблюдающий его врач не знают, какой из способов лечения был применен. Двойной слепой метод не всегда осуществим, поэтому используют также простой слепой (примененный способ лечения известен врачу, но не больному или наоборот) и частично слепой (и врач, и больной располагают лишь частью информации) методы. В любом случае информацию, которой располагают участники исследования, следует свести к минимуму.

Строго говоря, применение рандомизации и слепого метода — две разные проблемы, однако они настолько тесно связаны, что примеры, которые мы рассмотрим, приложимы к обеим.

Перевязка внутренней грудной артерии при стенокардии

Идея этой операции возникла еще в 30-е годы. При ишемической болезни сердца сосуды, питающие миокард, частично закупориваются атеросклеротическими бляшками. Миокард не получает достаточно кислорода, и при физической нагрузке, когда потребность в кислороде увеличена, возникает приступ стенокардии. Если перевязать внутренние грудные артерии, то кровь, которая раньше текла по ним, устремится (по крайней мере частично) в коронарные сосуды — примерно так рассуждали авторы

метода. Кровоснабжение миокарда улучшится, приступы стенокардии прекратятся. Сама же операция достаточно проста, ее можно выполнить под местной анестезией. Идея была осуществлена, и в 1958 г. Р. Митчелл и соавт.* опубликовали результаты. Операция была проведена 50 больным. Продолжительность послеоперационного наблюдения составляла от 2 до 6 месяцев. У 34 больных (68% общего числа) состояние улучшилось (у 18 приступы стенокардии прекратились полностью, у 16 стали реже). У 11 больных (22%) состояние осталось прежним, умерли 5 больных (10%). На первый взгляд, превосходные результаты.

Еще до публикации работы Митчелла на страницах журнала «Ридерс Дайджест» появилась восторженная статья «Хирург спасает сердце», принесшая этому способу лечения больше известности, чем все публикации в медицинских журналах.

Однако в наши дни мало кто слышал о перевязке внутренних грудных артерий. Что стало с этим многообещающим методом лечения? В 1959 г. Л. Кобб и соавт.** опубликовали результаты проверки эффективности двусторонней перевязки внутренних грудных артерий, полученные двойным слепым методом. Ни больной, ни врач, оценивавший результат операции, не знали, были ли перевязаны внутренние грудные артерии или нет. Больному делали надрезы и выделяли сосуды. Затем вскрывали конверт, в котором говорилось, нужно ли выполнить перевязку. К какой группе — экспериментальной или контрольной — принадлежал больной, покинувший операционную, знал только оперировавший его хирург. По данным послеоперационного наблюдения группы не различались ни по частоте приступов, ни по переносимости физической нагрузки. Чем было обусловлено обнаруженное Митчеллом улучшение состояния — отбором для операции наиболее легких больных, их энтузиазмом в отношении разрекламированного метода лечения или пристрастностью

---

* J. Mitchell, R. Glover, R. Kyle. Bilateral internal mammary artery ligation for angina pectoris: preliminary clinical considerations. *Am. J. Cardiol.*, 1:46—50, 1958.

** L. Cobb, G. Thomas, D. Dillard, K. Merendino, R. Bruce. An evaluation of internal-mammary-artery ligation by a double-blind technic. *N. Engl. J. Med.*, 260:1115—1118. 1959.

оценки результатов — судить трудно. Вывод же прост: результаты исследования без контрольной группы, без применения слепого метода несостоятельны.

Портокавальное шунтирование при циррозе печени

При алкоголизме часто развивается цирроз печени. Одно из его проявлений — портальная гипертензия: повышение давления в воротной вене из-за затруднения кровотока через печень. Повышение давления в воротной вене приводит к варикозному расширению вен пищевода. Это чрезвычайно опасное состояние: из-за разрыва варикозно расширенных вен в любой момент может возникнуть смертельное кровотечение. Для снижения давления в воротной вене применяют *портокавальное шунтирование:* воротную и нижнюю полую вены соединяют в обход печени.

Ранние работы по оценке результатов этой операции относятся к концу 40-х годов. Типичный план исследования в ту эпоху предусматривал набор определенного числа оперированных и подсчет доли выживших, каковая и рассматривалась в качестве результата. То обстоятельство, что больной мог бы выжить и без операции (а также умереть *в результате* операции), во внимание не принималось. Контрольные группы больных, *не подвергавшихся портокавальному шунтированию,* использовались редко.

В 1966 г., через двадцать лет после первой операции, Н. Грейс и соавт.* провели анализ полусотни исследований эффективности этого метода. Предметом анализа была связь между наличием контрольной группы и применением рандомизации, с одной стороны, и оценкой эффективности — с другой. Табл. 12.2 показывает, как распределились исследования по этим признакам. Проявилась любопытная закономерность. Если исследование выполнялось без контрольной группы или последняя формировалась не случайно, метод, как правило, получал высокую оценку. В тех немногих исследованиях, где использовалась контрольная группа и больные равновероятно распределялись между нею и экспериментальной, метод оценивался невысоко.

---

* N. Grace, H. Muench, T. Chalmers. The present status of shunts for portal hypertension in cirrhosis. *Gastroenterology*, 50:684—691, 1966.

Таблица 12.2. Оценки эффективности портокавального шунтирования (по результатам 51 исследования)

	Оценка		
Исследование	высокая	средняя	низкая
Без контрольной группы	24	7	1
С нерандомизированной контрольной группой	10	3	2
С рандомизированной контрольной группой	0	1	3

Причина высоких оценок в исследованиях без контрольной группы ясна, ведь само суждение об эффективности метода здесь совершенно произвольно. Сложнее с оценками, основанными на использовании нерандомизированных групп. Даже при кажущейся беспристрастности отбора сама *возможность* влиять на него толкает исследователя на построение неравноценных групп. В результате в одну группу попадают более тяжелые больные, в другую — более легкие.

Исследователь редко стремится обмануть других, но легко становится жертвой самообмана. При этом форма самообмана может быть весьма изощренной. Рассмотрим такой пример: больных, госпитализированных по нечетным дням месяца, определяют в экспериментальную группу, по четным — в контрольную. Можно ли считать такую рандомизацию достаточной? Разумеется, нет. Врач может влиять на срок госпитализации, следовательно, состав групп будет неслучайным.

Если у кого-либо из участников исследования есть возможность влиять на построение групп, эта возможность будет использована.

Для рандомизации недостаточно, чтобы выбор не зависел от *исследователя*. Он должен быть независим и от самих *подопытных*. Приведем пример из области лабораторных исследований. Двадцать крыс, сидящих в клетке, нужно разделить на две группы. Выпустим из клетки десять крыс и назовем их контрольной группой. Представительна ли она? Скорее всего, нет. Вероятно, первыми из клетки выбегут самые сильные и агрессивные особи.

Есть только один способ получить случайную выборку — воспользоваться для этого достоверно случайным процессом, на-

пример бросанием игральной кости или таблицей (генератором) случайных чисел.

Мы видели, что среди всех исследований эффективности портокавального шунтирования лишь те, в которых применялась рандомизация, показали истинную степень его эффективности. Остальные приводили к оценкам, смещенным в пользу операции. Общим правилом является следующее.

Чем лучше проведено исследование, тем менее вероятно его результат смещен в пользу исследуемого метода.

Влияние качества рандомизации на результаты клинических испытаний исследовали К. Шульц и соавт*. Рассмотрев 250 контролируемых клинических испытаний, они разделили их на хорошо и плохо рандомизированные. Хорошо рандомизированным считалось испытание, в котором распределение по группам основывалось на использовании случайных чисел. В остальных случаях участники исследования могли влиять на распределение по группам и испытание считалось плохо рандомизированным. Так, плохо рандомизированным считалось распределение, зависящее от момента включения в исследование. Шульц обнаружил, что доля методов лечения, признанных по итогам испытания эффективными, оказалась в плохо рандомизированных испытаниях на 41% выше, чем в хорошо рандомизированных. Некачественная рандомизация привела к почти полуторному завышению числа эффективных методов!

Этична ли рандомизация?

Итак, только рандомизация позволяет надежно оценить эффективность нового метода лечения. Но этична ли она, когда речь идет о жизни и здоровье людей? В гл. 3 мы уже говорили о психологических трудностях, связанных с рандомизацией. Рандомизация лишает права выбора и врача-экспериментатора, и самого больного. Простое решение состоит том, что *если достоверно не известно, какой метод лучше, то лечить можно любым.*

---

* K. F. Schulz, I. Chalmers, R. J. Hayes, D. G. Altman. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA*, 273:408—412, 1995.

К сожалению, на деле все не так просто. У любого метода найдутся сторонники и противники (иначе кто бы взялся за проверку.) Не будем говорить о мнении авторов метода. Но свои воззрения есть и у привлеченного к эксперименту врача, человека обычно просвещенного и не чуждого гуманизма. Почему, нередко спрашивает врач, я должен, подобно язычнику, слепо следовать воле неких случайных чисел, требующих лишить больного лучшего лечения? Этично ли в глазах поборников перевязки грудных артерий было использование Коббом рандомизированной контрольной группы? Однако, как мы видели, неэтичной оказалась скорее не рандомизация, а операция. Слыша мнения о нецелесообразности рандомизированных испытаний, задайте вопрос: *на чем, кроме веры и интуиции, основано убеждение в достоинствах одного и недостатках другого метода?* Ведь сравнительная проверка еще только предстоит.

Мы привели примеры неэффективных методов, которые успели стать достоянием практической медицины, но все же не превратились в общепринятые. К сожалению, опровергнуть укоренившийся метод почти так же невозможно, как невозможно опровергнуть традицию. Самое тщательное доказательство неэффективности давно прижившегося метода в лучшем случае ускорит его естественное отмирание. Так невозможно доказать отсутствие лечебного действия пиявок, этих священных коров практической медицины.

Мы уже говорили о том, что не следует путать достоверность и статистическую значимость. Именно в совершенно недостоверных работах уровень значимости, как правило, не оставляет желать лучшего. Нередко приходится слышать о «высоко достоверных результатах, $P < 0{,}01$», тогда как речь идет о нерандомизированном исследовании, применительно к которому, как мы показали, вообще бессмысленно говорить о значении P. И наоборот, если в результате правильно проведенного исследования мы получили значение $P < 0{,}1$, то это значит, что вероятность ошибочно признать существование различий не превышает 10% — и это утверждение *истинно*. Какой практический вывод сделать из этого истинного утверждения, каждый может решить сам. Считать ли вероятность ошибки 10% слишком большой — это вообще не вопрос статистики. Многое тут зависит от того, чем мы

рискуем, признав или отвергнув предлагаемый метод лечения. Меньше всего следует фетишизировать уровень значимости и придавать ему смысл критерия истинности. В конце концов, различие между 5 и 10% чисто количественное. Гораздо важнее тщательно продумывать, какую совокупность должна представлять ваша выборка, как обеспечить случайность формирования групп и уберечься от невольного самообмана при оценке результатов.

Всегда ли нужна рандомизация?

Следует признать, что великие открытия, изменившие облик медицины в середине XX века, такие, как открытие пенициллина, не подвергались проверке в рандомизированных исследованиях.

Порой сами обстоятельства способны натолкнуть на переоценку общепринятых методов лечения. Так, французский военный хирург Амбруаз Парэ в полном соответствии с предписаниями лечил огнестрельные раны кипящим маслом. Однажды, в одну из битв 1536 г., масла на всех раненых не хватило. Части солдат Парэ сделал перевязку, не обработав рану этим целительным средством. Утром он с удивлением обнаружил, что солдаты, чьи раны перед перевязкой были обработаны по всем правилам, корчатся от боли, тогда как просто перевязанные «прекрасно отдохнули и не испытывали болей»*. История умалчивает, подал ли Парэ рапорт о необходимости проведения рандомизированных клинических испытаний эффективности кипящего масла как средства лечения пулевых ранений. Но нам не кажется, что, соверши он свое открытие в наши дни, ему потребовалось бы детальная проверка.

Наконец, рандомизация не всегда возможна. Так, в гл. 11 мы рассмотрели выживаемость после трансплантации костного мозга при остром лимфобластном лейкозе взрослых. Одним больным пересаживался костный мозг близких родственников, дру-

* Пример заимствован из книги H. R. Wulff. Rational Diagnosis and Treatment, Blackwell, Oxford, 1976. В этой небольшой по объему и блистательно написанной книге вы найдете многое идей, перекликающихся с нашим обсуждением.

гим — их собственный. Случайно распределить больных по двум этим группам невозможно, поскольку не у каждого найдется родственник-донор. К счастью для экспериментаторов, само по себе наличие или отсутствие близких родственников не влияет на течение заболевания. Ситуация, когда разделить больных случайным образом невозможно, в медицинских исследованиях возникает довольно часто. В таких случаях надо стремиться сделать группы максимально схожими по всем известным прогностическим факторам.

ДОСТАТОЧНО ЛИ РАНДОМИЗАЦИИ?

Контролируемые рандомизированные клинические испытания сегодня стали эталоном медицинского исследования. Но всегда ли они приводят к верным заключениям? Нет, не всегда. Нередко в исследовании скрыто присутствует множественное сравнение. Исследователь не учитывает эту множественность и в результате, сам того не подозревая, многократно занижает вероятность ошибочно выявить мнимый эффект. Рассмотрим три типичных случая.

Проверкой нового метода лечения независимо друг от друга занимаются несколько исследователей. Получив положительный результат, исследователь опубликует его. А получив отрицательный? Вероятно, воздержится от публикации, но, кроме того, еще и предпримет повторную проверку. В конце концов в одной из *многих* проверок будет обнаружен желанный «эффект». В гл. 4 мы описали эту ситуацию и привели оценки истинной вероятности ошибиться, многократно превышающей вероятность ошибки в единичном испытании.

В медицине приняты широкомасштабные исследования различных методов лечения, используемых прежде всего при хронических болезнях, таких, например, как ишемическая болезнь сердца и сахарный диабет. Результатом исследования является описание огромного числа разнообразных признаков. Данные подвергаются различным группировкам с целью выяснения наиболее информативных признаков, в наибольшей степени влияющих на конечный показатель — выживаемость. Понятно, что

при значительном числе возможных группировок не составит труда выделить группы, на которых тот или иной метод лечения будет наиболее эффективен. Эту плодотворную деятельность мог бы омрачить учет множественности сравнений, например применение поправки Бонферрони. Приведем пример. Администрация по делам ветеранов провела рандомизированное исследование коронарного шунтирования*. Среди наблюдавшихся больных *в целом* не было выявлено статистически значимых различий в выживаемости между оперированными и неоперированными больными. Однако стоило разделить наблюдения на подгруппы, как оказалось, что хирургическое вмешательство обеспечивает более высокую выживаемость среди «больных с поражением ствола левой коронарной артерии». Интерпретация подобных находок требует крайней осторожности.

Сходная картина наблюдается, когда в данных, полученных для анализа одних факторов, обнаруживается связь между другими. Возможно, это реально существующая связь, но, возможно, и злая шутка эффекта множественных сравнений, когда, попарно сравнивая все со всем, исследователь непременно найдет какую-нибудь статистическую зависимость. Поэтому для проверки такой попутно обнаруженной связи нужно выполнить отдельное исследование.

К чему может привести вольная группировка данных, полученных в безупречно выполненном рандомизированном исследовании, было убедительно показано Ли и соавт.** Они воспроизвели достаточно типичное исследование. Взяв истории болезни 1073 больных ишемической болезнью сердца, они случайным образом разделили их на две группы. Одну группу назвали контрольной, а другую экспериментальной (представим себе, что попавшие в нее получали волшебный препарат «рандоми-

---

* M. Murphy, H. Hultgren, K. Detre, J. Thomsen, T. Takaro. Treatment of chronic stable angina: a preliminary report of survival data of the Randomized Veterans Administration Cooperative Study. *N. Engl. J. Med.*, 297:621—627, 1977.

** K. Lee, F. McNeer, F. Starmer, P. Harris, R. Rosati. Clinical judgement and statistics: lessons from a simulated randomized trial in coronary artery disease. *Circulation*, 61:508—515, 1980.

зин»). Между группами не было обнаружено значимых различий по таким признакам, как возраст, пол, число пораженных коронарных артерий и т. д. По одному признаку — сократимости левого желудочка — статистически значимое различие наблюдалось. Несомненно, пытливый исследователь не преминул бы связать это различие с использованием «рандомизина». Однако, увы, по самому важному признаку — выживаемости — различие было статистически не значимым (см. рис.12.1А).

В этой ситуации исследователь наверняка продолжил бы поиск различий, разделив больных на более мелкие группы. Так и поступил Ли. Больные были разделены (стратифицированы) по двум признакам: числу пораженных коронарных артерий (1, 2 или 3) и сократимости левого желудочка (нормальной или сниженной). В результате получилось 6 подгрупп. Влияние рандомизина на выживаемость изучалось в каждой из этих подгрупп. Но этого мало. Каждая подгруппа была разделена еще на две в зависимости от наличия или отсутствия сердечной недостаточности. В каждой из получившихся 12 подгрупп вновь оценивалась эффективность рандомизина. Упорные усилия были вознаграждены. В одной из подгрупп (больные с поражением 3 коронарных артерий и сниженной сократимостью левого желудочка) рандомизин оказался эффективен: различия выживаемости «леченых» и «нелеченых» были статистически значимыми, $P < 0{,}025$ (рис. 12.1Б).

Рандомизин — выдумка. Но многочисленные препараты, эффективность которых была доказана совершенно таким же способом, существуют в действительности. Секрет их «эффективности» очень прост — это множественность сравнений. В исследовании рандомизина было построено 18 пар подгрупп и выполнено 18 сравнений. Чему равна вероятность получить хотя бы один значимый результат в 18 сравнениях, уровень значимости в каждом из которых равен 0,05? Находим: $\alpha' = 1-(1-\alpha)^k = 1-(1-0{,}05)^{18} = 1-0{,}40 = 0{,}60$. Таким образом, истинная вероятность ошибки I рода оказалась в 12 раз выше той, о которой доложил бы исследователь.

Как избежать несостоятельных выводов, не отказываясь от возможности группировать данные? Для этого достаточно в уровне значимости каждого отдельного сравнения учесть, что их

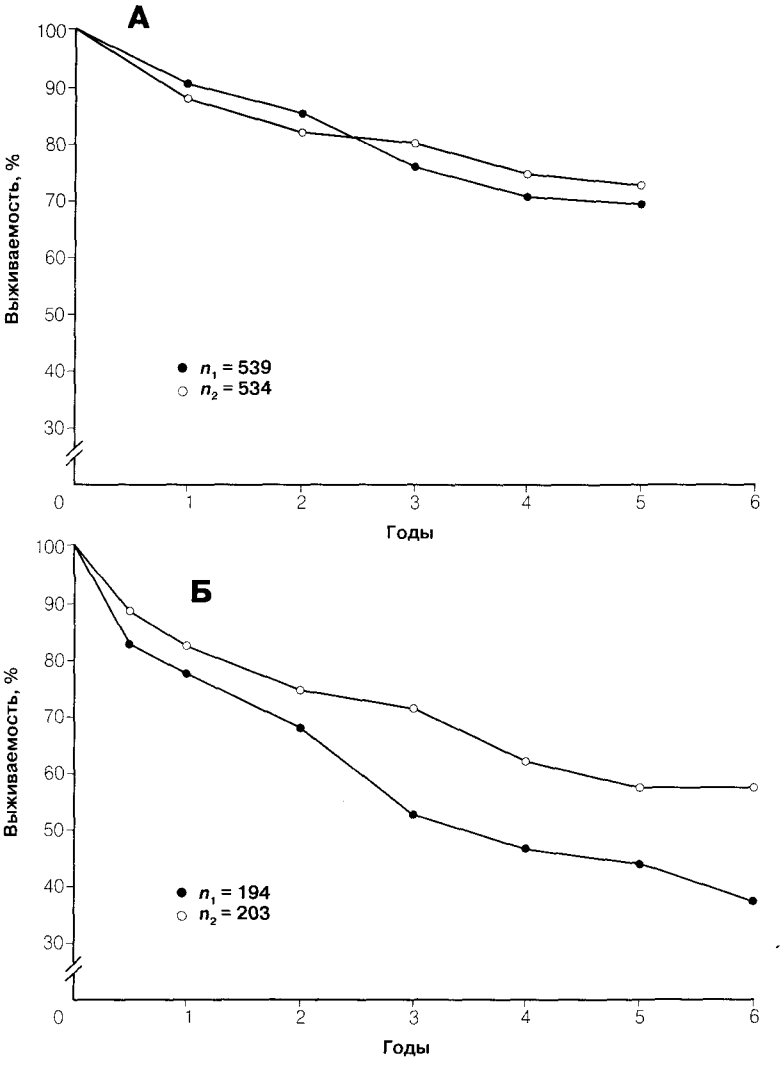


Рис. 12.1. А. Больных с ишемической болезнью сердца (1073 человека) случайным образом разделили на 2 группы. Статистически значимых различий выживаемости не обнаружено. **Б.** Выделив больных с поражением 3 коронарных артерий и сниженной сократимостью левого желудочка, их вновь случайным образом разделили на 2 группы. На этот раз различия выживаемости статистически значимы ($P < 0{,}025$). Выделяя все новые подгруппы, мы в конце концов всегда найдем различия там, где их нет.

более одного. Поправка Бонферрони дает уровень значимости, равный a'/k, где a' — выбранный уровень значимости для всего набора из k сравнений. Это чрезмерно жесткая, заниженная оценка. Наиболее продуктивный подход состоит в применении многофакторных статистических методов*. Помимо прочего, они позволяют обнаружить *одновременное* влияние более чем двух методов лечения, что в принципе недоступно методам, изложенным ранее.

КОГО МЫ ИЗУЧАЕМ

В лабораторных исследованиях, в исследованиях общественного мнения или потребительского спроса существует достаточная определенность, что представляет собой исследуемая совокупность. Понятно и как организовать представительную выборку из нее. Иначе обстоит дело в клинических исследованиях. Здесь нет ясности ни в том, какова изучаемая совокупность, ни в том, как построить представительную выборку из нее.

Чаще всего исследования проводятся в крупных клиниках, куда попадают далеко не все больные. При всей своей условности рис. 12.2, тем не менее, отражает реальную картину. Из 1000 больных госпитализируется лишь девять и *только один* попадает в клинику. Ясно, что сложный путь больного по медицинским учреждениям далеко не случаен — он определяется прежде всего тяжестью, сложностью случая или редкостью болезни. Поэтому при всем желании больных в клиниках трудно признать представительной выборкой. Это несоответствие обязательно нужно иметь в виду, решая, на какую совокупность больных могут быть (и в какой мере) распространены полученные в исследовании результаты.

Данные, относящиеся к госпитализированным больным, и прежде всего к больным из крупных клиник, не отражают ни общий спектр болезней и их стадий, ни их взаимосвязь. Исследователи вынуждены изучать взаимосвязь болезней, опираясь на дан-

* С ними вы можете познакомиться в нашей книге: S. A. Glantz, B. K. Slinker. Primer of Applied Regression and Analysis of Variance. McGraw-Hill, N.Y., 1990.

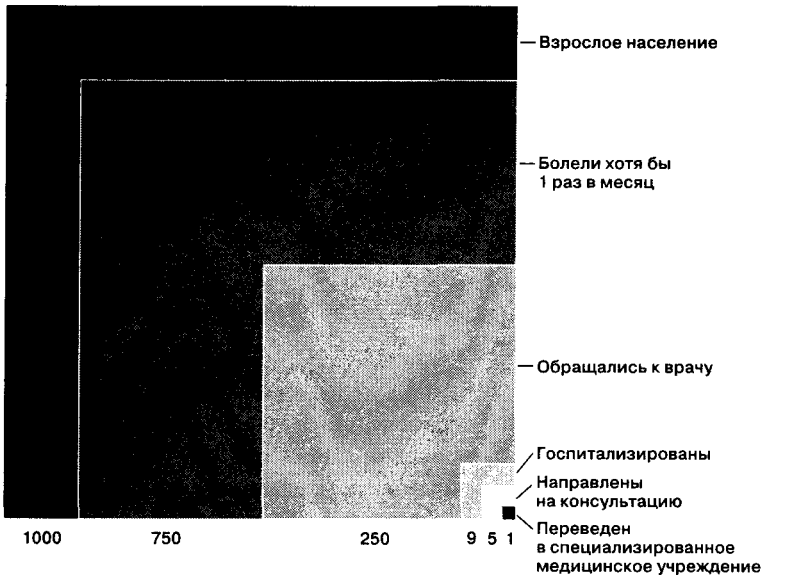


Рис. 12.2. В специализированных медицинских учреждениях оказывается лишь очень незначительная доля больных — обычно они лечатся амбулаторно или не лечатся вообще. На рисунке показано, сколько человек на 1000 населения болеют, обращаются к врачу и попадают в больницу в течение месяца.

ные, относящиеся к госпитализированным или амбулаторным больным. Но разные заболевания и разные стадии одного заболевания требуют разных форм лечения. В результате связь заболеваний представляется искаженной. Человек, страдающий несколькими болезнями, имеет больше шансов попасть в больницу, чем человек с одной болезнью. Поэтому наиболее частый вид искажения — это мнимое обнаружение связи заболеваний или преувеличение действительно существующей связи. В задаче 5.10 мы встретились с более сложным видом искажения, когда из-за неравной вероятности госпитализации создается впечатление о более сильной связи болезни X с болезью Z, чем с болезнью Y. Данные о связи заболеваний, полученные при изучении госпитализированных больных, следует оценивать с чрез-

вычайной осторожностью. Эта проблема названа по имени Берксона*, первым обратившего на нее внимание.

КАК УЛУЧШИТЬ ПОЛОЖЕНИЕ

Способность применить статистический подход в медицине не сводится к заучиванию нескольких формул и умению отыскать табличное значение. Как и любая творческая деятельность, применение статистических методов и интерпретация полученных результатов требуют глубокого проникновения в суть дела — понимания как возможностей и ограничений используемых методов, так и существа решаемой клинической задачи. В гл. 1 мы говорили, что значение статистических методов возрастает по мере ужесточения требований к обоснованию эффективности предлагаемых методов лечения. Статистическое обоснование зачастую оказывается важнейшим фактором, определяющим решение в пользу предлагаемого лечения.

В то же время сами медики редко занимаются статистическим обоснованием своих исследований в силу того, что их познания в этой области столь же скромны, сколь и оторваны от практики. Обычно вся статистическая сторона дела перепоручается консультантам, нередко действительно разбирающимся в статистике, но имеющим довольно смутное представление о медицинских вопросах. Единственный выход состоит в том, чтобы медики наконец сами занялись статистическим анализом, поскольку именно они знают цели исследования и несут за него ответственность.

---

* J. Berkson. Limitations of the applications of fourfold table analysis to hospital data. *Biometrics,* 2:47—53, 1946. Менее формальное обсуждение вы найдете в работе D. Mainland. The risk of fallacious conclusions from autopsy data on the incidence of diseases with application to heart disease. *Am. Heart. J.*, 45:644—654, 1953. Пример того, сколь различны выводы, полученные в результате наблюдения больных из конкретной клиники, всех госпитализированных больных и, наконец, всех больных, приведен в комментарии Мюнча (*N. Engl. J. Med.* 272:1134, 1965) к работе H. Binder, A. Clement, W. Thayer, H. Spiro. Rarity of hiatus hernia in achalasia. *N. Engl. J. Med.*, 272:680—682, 1965.

Увы, проблема усугубляется еще и тем, что у немалой части исследователей сбор данных предшествует формулировке вопроса, на который они должны бы ответить. На этом пути исследователя неизменно подстерегают малоприятные открытия. Всякий раз исследователь попадает в ситуацию, когда данные собраны и остается только вычислить значение *P*, но тут обнаруживается, что это значение существует не само по себе, а лишь в связи с *проверкой гипотезы*. Но самое обескураживающее — чтобы проверить гипотезу, ее, оказывается, нужно иметь.

Не многие исследователи обременяют себя необходимостью еще до начала сбора данных осознать цели исследования и подлежащие проверке гипотезы. Например, лишь 20% протоколов, одобренных комитетом по клиническим исследованиям одного крупного научно-медицинского центра, содержали четко сформулированные гипотезы*.

Попытайтесь понять, что вы хотите от исследования, какой вопрос вы хотите решить. И когда у вас будет конкретная гипотеза, станет понятно, каким должен быть тип предстоящего эксперимента и какие потребуются данные. Тогда по табл. 12.1 вы легко определите нужный метод анализа. Придерживаясь этих правил, вы всегда соберете данные, необходимые и достаточные для анализа.

Лишь очень немногие поступают таким образом. Поэтому неудивительно, что, когда настает время вычислить значение *P*, исследователь обнаруживает, что собранные им данные мало связаны с проверяемой гипотезой, да к тому же нарушают предпосылки известных ему статистических методов. Но не начинать же все с начала. Поэтому для устранения и сглаживания статистических несообразностей на этом, *завершающем* этапе призывается специалист, который оставляет от Монблана данных немногое, хоть как-то пригодное для анализа, заменяет неприменимые параметрические методы неприхотливыми, но менее чув-

---

* Подробнее об этой проблеме и той роли, которую могли бы сыграть в ее решении комитеты по клиническим исследованиям, говорится в работе M. Giammona, S. Glantz. Poor statistical design in research on humans: the role of Committees on Human Research. *Clin. Res.*, 31:571—577, 1983.

ствительными непараметрическими или предлагает вместо одной гипотезы перейти к нескольким, пригодным для статистической проверки. Отчет об исследовании приобретает приемлемый вид. Однако само исследование не становится более осмысленным. Способ избежать этого прост и состоит в том, чтобы задуматься о том, как анализировать данные, в начале, а не в конце исследования.

С примерами несостоятельных работ мы неоднократно встречались в этой книге. Еще чаще они встречаются в жизни. Поэтому серьезный врач, особенно исследователь, не должен принимать за чистую монету все, что пишется в журналах.

Знакомясь с материалами очередного исследования, обратите внимание, названы ли:

- подлежащая проверке гипотеза;
- использованные данные и способ их получения (включая метод рандомизации);
- совокупность, которую представляют используемые в исследовании выборки;
- статистические методы, использованные для оценки гипотезы.

Очень трудно найти публикацию, которая бы содержала все это. Но чем ближе она к такому идеалу, тем вернее можно положиться на приведенные в ней выводы. Напротив, очень мало доверия заслуживает статья, в которой использованные методы не указаны вовсе или упоминаются некие «стандартные методы».

Возвращаясь к вопросу об этичности исследований на людях, хочется подчеркнуть, что чем менее грамотно и добросовестно исследование, тем менее оно этично, как по отношению к тем больным, которые в нем участвовали, так и ко всем больным, лечение которых напрямую зависит от его результатов. Неэтичен любой вводящий в заблуждение результат. Неэтично подвергать людей страданиям и мучить лабораторных животных ради получения данных, на основании которых невозможно сделать какой-либо вывод. Неэтично выполнять такие исследования, опровержение которых потребует чьих-то сил, здоровья и средств.

Конечно, тщательная проработка статистической стороны исследования не освобождает исследователя от обязанности тща-

тельно продумать эксперимент с врачебной точки зрения, свести риск и страдания больных к минимуму. Больше того, она даже не гарантирует, что в исследовании будут получены глубокие и новаторские результаты. Иными словами, статистическая корректность — это необходимое, но еще не достаточное условие успеха исследования.

Как же изменить исследовательскую практику к лучшему?

Прежде всего, будьте активны. Если это от вас зависит, не подпускайте к исследованиям людей, несведущих в статистике, как не подпускаете тех, кто не смыслит в медицине. Встретив статистические несуразности в журнале, пишите редактору*. Не стесняйтесь задавать вопросы своим коллегам. Не поддавайтесь гипнозу наукообразия — докапывайтесь до сути дела. Когда вас осыпают мудреными терминами, спросите, что в данном случае означает *P*.

Но самое главное, чтобы ваши собственные исследования были безупречны с точки зрения планирования и применения статистических методов.

---

* Если редактор не утратил интерес к жизни и профессии, он обязательно среагирует. Так, в 1978 г., еще никому неизвестным медиком, я написал в *Circulation Research* о случаях неверного использования критерия Стьюдента для множественного сравнения (об этом см. гл. 1 и 4). Редакторы получили отзыв на мое письмо у специалиста, после чего пересмотрели требования редакции к изложению в публикуемых статьях статистических методов и методов проведения эксперимента. Два года спустя редакция сообщила о «значительном улучшении применения методов проверки статистической значимости публикуемых в журнале результатов». Желающих ознакомиться с перепиской по этому вопросу отошлем к работам M. Rosen, B. Hoffman. Editorial: statistics, biomedical scientists, and circulation research. *Circ. Res.*, 42:739, 1978 и S. Glantz. Biostatistics: how to detect, correct, and prevent errors in the medical literature. *Circulation*, 61:1—7, 1980; S. Wallenstein, C. Zucker, J. Fleiss. Some statistical methods useful in circulation research. *Circ. Res.*, 47:1—9, 1980.

Формулы для вычислений

ДИСПЕРСИЯ

$$s^2 = \frac{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}{n-1}.$$

ДИСПЕРСИОННЫЙ АНАЛИЗ

Расчет по групповым средним и стандартным отклонениям

Имеется k групп; n_i — численность i-й группы, $\overline{X}_i$ — среднее в i-й группе, s_i — стандартное отклонение в i-й группе.

$$N = \Sigma n_i.$$

$$S_{\text{вну}} = \Sigma (n_i - 1) s_i^2.$$

$$\nu_{\text{вну}} = N - k.$$

$$S_{\text{меж}} = \Sigma n_i \overline{X}_i^2 - \frac{(\Sigma n_i \overline{X}_i)^2}{N}.$$

$$\nu_{меж} = k - 1.$$

$$F = \frac{S_{меж} / \nu_{меж}}{S_{вну} / \nu_{вну}}.$$

Расчет по исходным данным

n_i — численность i-й группы, X_{ij} — значение признака у j-го больного i-й группы.

$$C = \frac{\left(\sum_i \sum_j X_{ij} \right)^2}{N}.$$

$$S_{общ} = \sum_i \sum_j X_{ij}^2 - C.$$

$$S_{меж} = \sum_i \frac{\left(\sum_j X_{ij}^2 \right)^2}{n_i} - C.$$

$$S_{вну} = S_{общ} - S_{меж}.$$

Число степеней свободы и величина F вычисляются как при расчете по групповым средним и стандартным отклонениям.

КРИТЕРИЙ СТЬЮДЕНТА

Расчет по групповым средним и стандартным отклонениям

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}},$$

где

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{n_1 + n_2}{n_1 n_2 (n_1 + n_2 - 2)} [(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]}.$$

$$\nu = n_1 + n_2 - 2.$$

Расчет по исходным данным

$$s_{\bar{X}_1-\bar{X}_2}=\sqrt{\frac{n_1+n_2}{n_1n_2(n_1+n_2-2)}\left[\Sigma X_1^{\,2}-\frac{(\Sigma X_1^{\,2})}{n_1}+\Sigma X_2^{\,2}-\frac{(\Sigma X_2^{\,2})}{n_2}\right]}.$$

Значения t и ν вычисляются как при расчете по групповым средним и стандартным отклонениям.

ТАБЛИЦА СОПРЯЖЕННОСТИ 2×2

Имеется таблица сопряженности

A	B
C	D

$$\chi^2=\frac{N\left(|AD-BC|-\frac{N}{2}\right)^2}{(A+B)(C+D)(A+C)(B+D)},$$

где $N=A+B+C+D$.

$\nu=1$.

Критерий Мак-Нимара

Значения двух качественных признаков «есть—нет» определены у одних и тех же больных:

		Признак 1	
		+	–
Признак 2	+	A	B
	–	C	D

Тогда

$$\chi^2 = \frac{(|B-C|-1)^2}{B+C}.$$

$$\nu = 1.$$

Точный критерий Фишера

1. Вычислить

$$P' = \frac{R_1!R_2!C_1!C_2!}{N!A!B!C!D!},$$

где R_1 и R_2 — суммы по строкам, C_1 и C_2 — суммы по столбцам.

2. Найти наименьшее из чисел A, B, C и D. Допустим, это число A.

3. Уменьшить A на единицу.

4. Пересчитать числа в остальных клетках так, чтобы суммы по строкам и столбцам остались прежними.

5. Вычислить P' по приведенной формуле.

6. Повторять шаги 3—5, пока A не станет равным 0.

7. Сложить все значения P', которые не превышают P' для исходной таблицы (включая P' для исходной таблицы).

Полученная сумма представляет собой значение P для одностороннего варианта точного критерия Фишера. Чтобы получить значение P для двустороннего варианта, нужно продолжить вычисления в следующем порядке.

8. Вернуться к исходной таблице.

9. Увеличить A на единицу.

10. Пересчитать числа в остальных клетках так, чтобы суммы по строкам и столбцам остались прежними.

11. Вычислить P'.

12. Повторять шаги 9—11, пока одно из чисел в клетках не станет равным 0.

13. Сложить значения P', которые не превышают P' для исходной таблицы, и прибавить значение P для одностороннего варианта. Полученная сумма представляет собой значение P для двустороннего варианта точного критерия Фишера.

Факториалы чисел от 0 до 20

n	$n!$
0	1
1	1
2	2
3	6
4	24
5	120
6	720
7	5040
8	40320
9	362880
10	3628800
11	39916800
12	479001600
13	6227020800
14	87178291200
15	1307674368000
16	20922789888000
17	355687428096000
18	6402373705728000
19	121645100408832000
20	2432902008176640000

При $n > 20$ используйте формулу

$$n! \approx \left(\frac{n}{e}\right)^n \sqrt{2\pi n},$$

где $e = 2{,}71828$ (основание натуральных логарифмов), $\pi = 3{,}14159$ (число «пи»).

КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ

$$S_{\text{общ}} = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n}.$$

$$S_{\text{рег}} = b\left(\Sigma XY - \frac{\Sigma X \Sigma Y}{n}\right).$$

$$r = \sqrt{\frac{S_{\text{рег}}}{S_{\text{общ}}}} = \frac{\Sigma XY - n\overline{X}\,\overline{Y}}{\sqrt{(\Sigma X^2 - n\overline{X}^2)(\Sigma Y^2 - n\overline{Y}^2)}}.$$

ДИСПЕРСИОННЫЙ АНАЛИЗ ПОВТОРНЫХ ИЗМЕРЕНИЙ

k — число измерений, n — число больных. Подстрочные индексы: i — номер измерения, j — номер больного, например X_{ij} — результат i-го измерения у j-го больного.

$$A = \frac{\left(\sum_i \sum_j X_{ij}\right)^2}{kn}.$$

$$B = \sum_i \sum_j X_{ij}^2.$$

$$C = \frac{\sum_i \left(\sum_j X_{ij}\right)^2}{n}.$$

$$D = \frac{\sum_j \left(\sum_i X_{ij}\right)^2}{k}.$$

$$S_{\text{ле}} = C - A.$$

$$S_{\text{ост}} = A + B - C - D.$$

$$\nu_{\text{ле}} = k - 1.$$

$$\nu_{\text{ост}} = (n-1)(k-1).$$

$$F = \frac{S_{\text{ле}} / \nu_{\text{ле}}}{S_{\text{ост}} / \nu_{\text{ост}}}.$$

КРИТЕРИЙ КРУСКАЛА—УОЛЛИСА

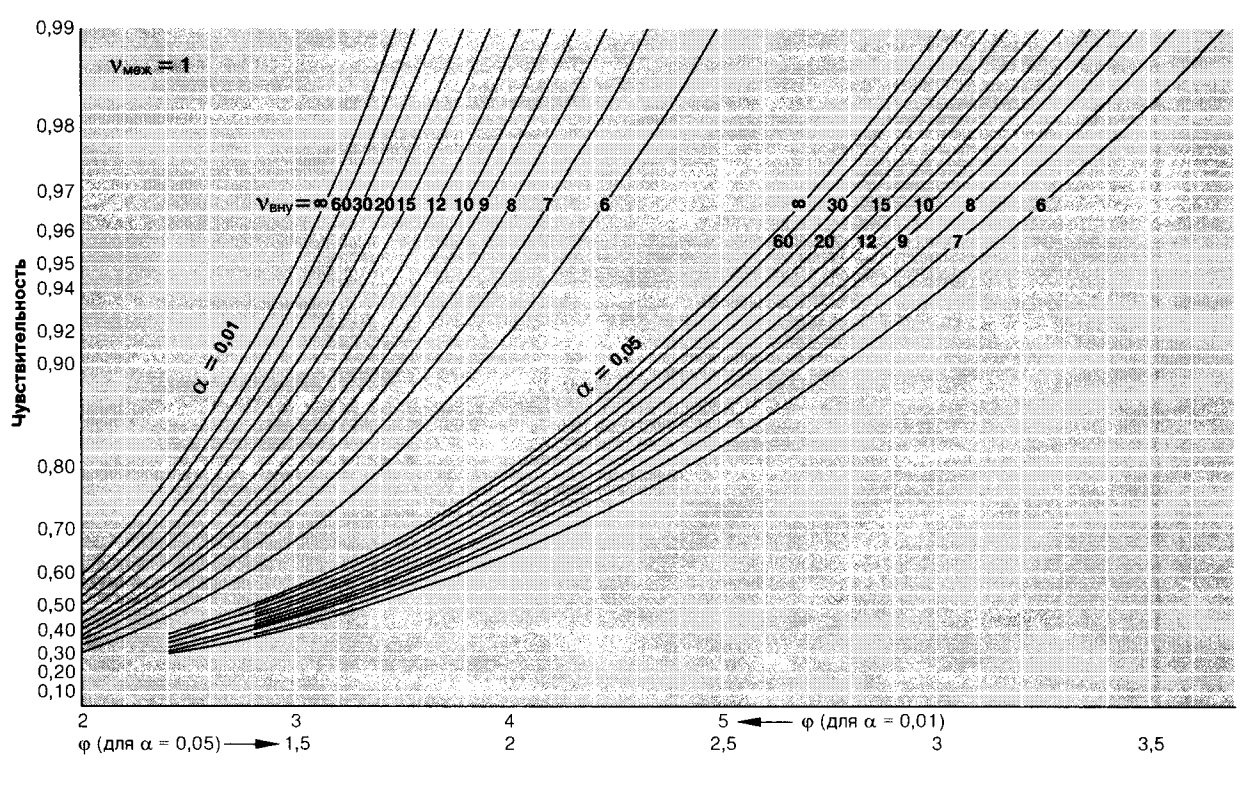
$$H = \frac{12}{N(N+1)} \sum \left(\frac{R_i^2}{n_i} \right) - 3(N+1),$$

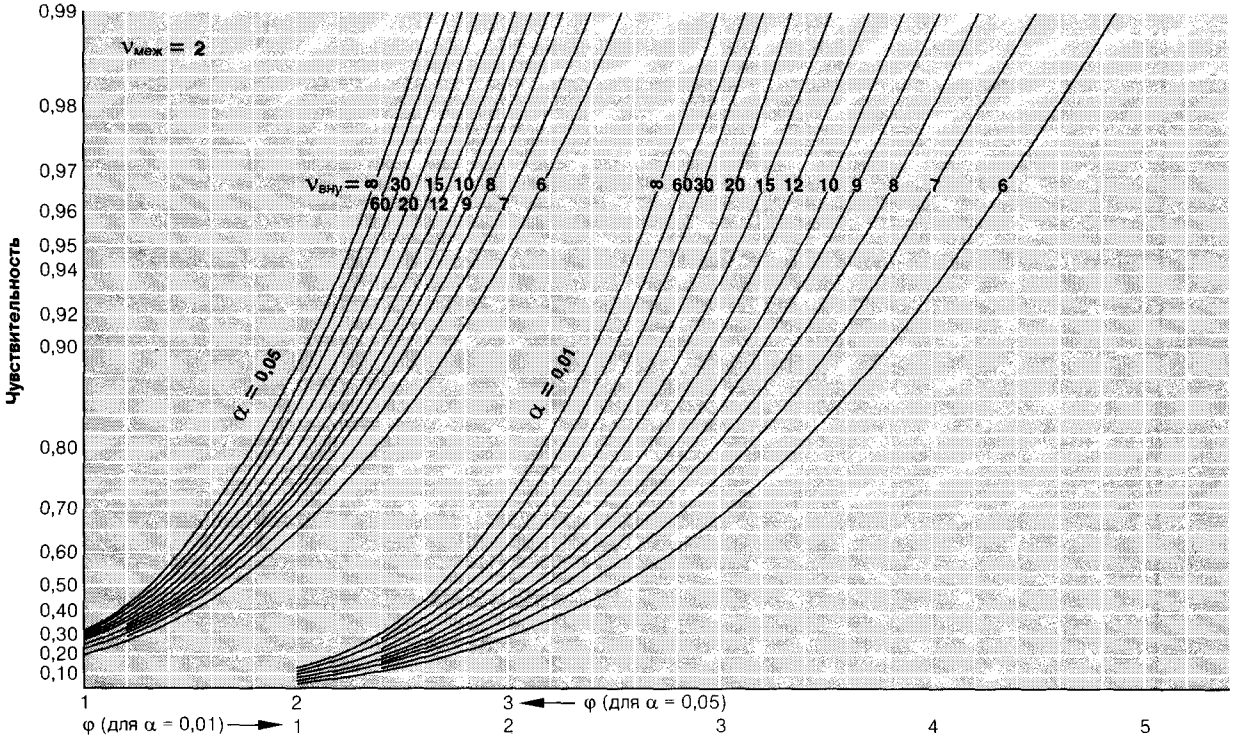
$$\chi_r^2 = \frac{12}{nk(k+1)} \Sigma R_i^2 - 3n(k+1),$$

где R_i — сумма рангов i-го измерения.

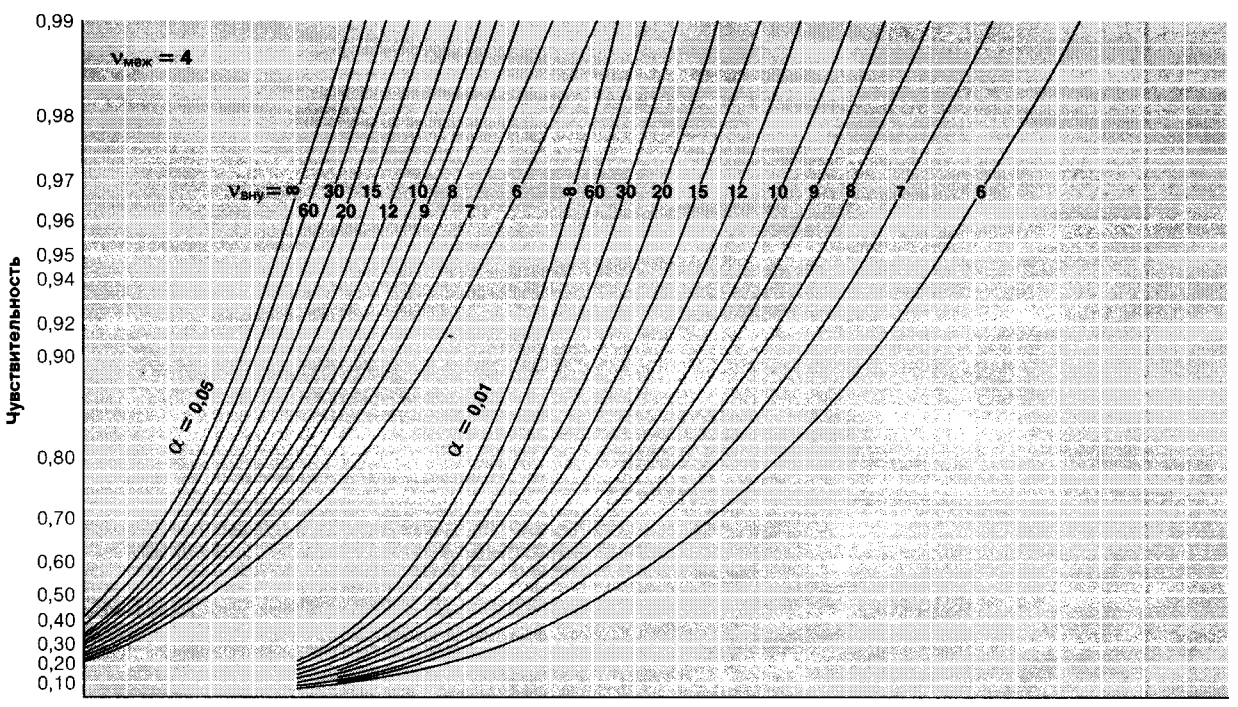
Приложение Б

Диаграммы чувствительности дисперсионного анализа


Чувствительность
0,99
0,98
0,97
0,96
0,95
0,94
0,92
0,90
0,80
0,70
0,60
0,50
0,40
0,30
0,20
0,10
$\nu_{меж} = 1$
$\nu_{вну} = \infty$ 60 30 20 15 12 10 9 8 7 6
$\alpha = 0{,}01$
$\alpha = 0{,}05$
∞ 30 15 10 8 6
60 20 12 9 7
2 3 4 5 ← φ (для α = 0,01)
φ (для α = 0,05) → 1,5 2 2,5 3 3,5

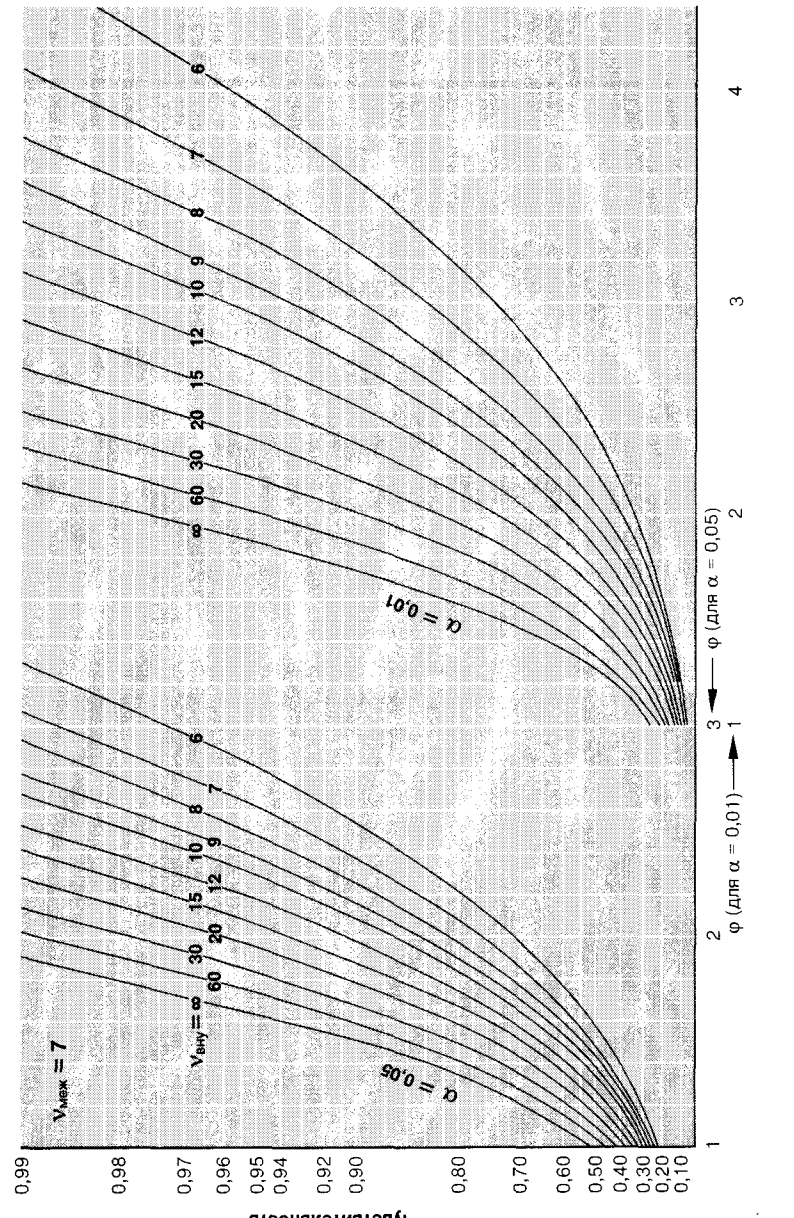

Чувствительность
0,99
0,98
0,97
0,96
0,95
0,94
0,92
0,90
0,80
0,70
0,60
0,50
0,40
0,30
0,20
0,10
$\nu_{меж} = 2$
$\nu_{вну} = \infty$ 30 15 10 8 6
60 20 12 9 7
$\alpha = 0{,}05$
∞ 60 30 20 15 12 10 9 8 7 6
$\alpha = 0{,}01$
1 2 3 ← φ (для α = 0,05)
φ (для α = 0,01) → 1 2 3 4 5

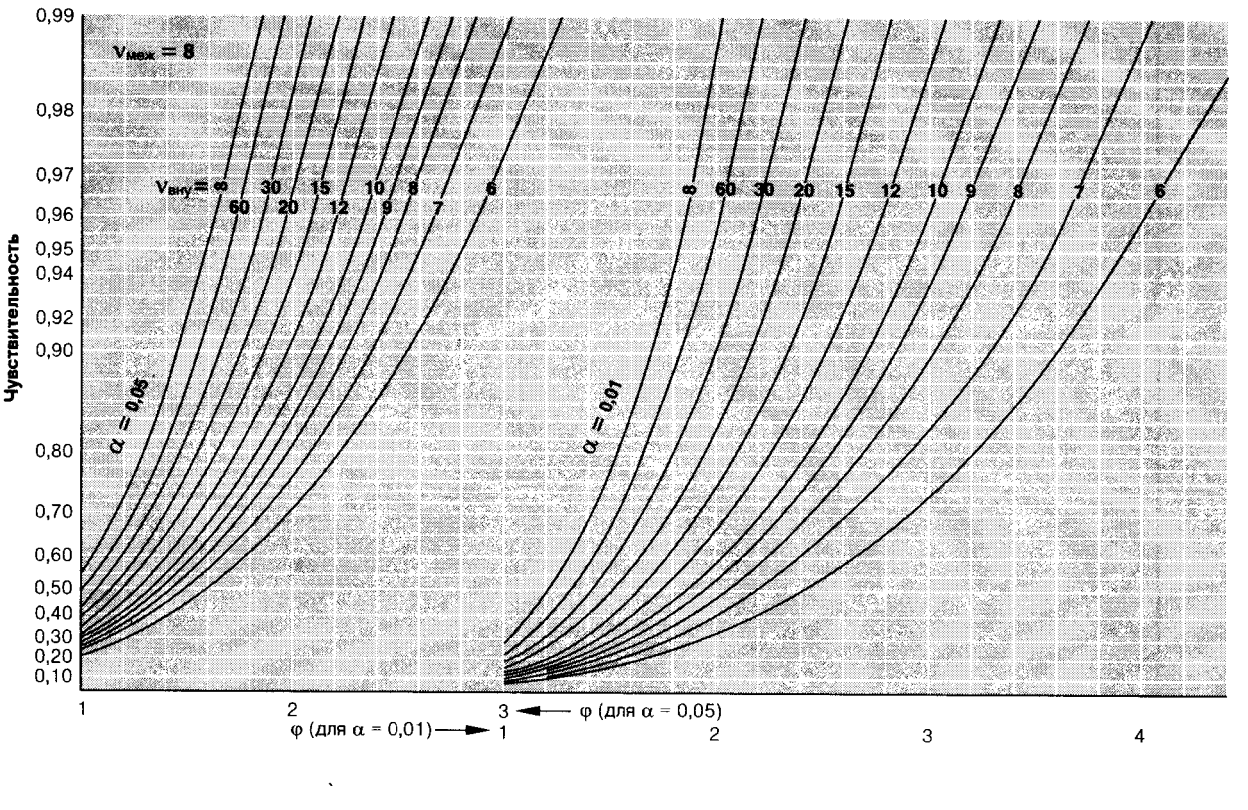
Чувствительность
0,99
0,98
0,97
0,96
0,95
0,94
0,92
0,90
0,80
0,70
0,60
0,50
0,40
0,30
0,20
0,10
$V_{меж} = 3$
$V_{вну} = \infty$ 30 15 10 8 7 6
60 20 12 9
$\alpha = 0,05$
$\alpha = 0,01$
∞ 60 30 20 15 12 10 9 8 7 6
1 2
φ (для α = 0,01)
3 2
φ (для α = 0,05)
1 2 3 4 5


Чувствительность
0,99
0,98
0,97
0,96
0,95
0,94
0,92
0,90
0,80
0,70
0,60
0,50
0,40
0,30
0,20
0,10
$\nu_{меж} = 4$
$\nu_{вну} = \infty$ 60 30 20 15 12 10 9 8 7 6
$\alpha = 0,05$
$\alpha = 0,01$
1 2 3 4 5
φ (для α = 0,01) → 1 2 3
← φ (для α = 0,05)

Чувствительность
0,99
0,98
0,97
0,96
0,95
0,94
0,92
0,90
0,80
0,70
0,60
0,50
0,40
0,30
0,20
0,10
$\nu_{max} = 5$
$\nu_{env} = \infty$
60
30
20
15
12
10
9
8
7
6
5
4
3
$\alpha = 0{,}05$
$\alpha = 0{,}01$
1
2
φ (для α = 0,01)
3
2
1
φ (для α = 0,05)
3
4
5

Чувствительность
0,99
0,98
0,97
0,96
0,95
0,94
0,92
0,90
0,80
0,70
0,60
0,50
0,40
0,30
0,20
0,10
$\nu_{max} = 6$
$\nu_{вну} = \infty$ 60 30 20 15 12 10 9 8 7 6
$\alpha = 0{,}05$
$\alpha = 0{,}01$
1 2 3
φ (для α = 0,01)
1 2 3
φ (для α = 0,05)
1 2 3 4


Чувствительность
0,99
0,98
0,97
0,96
0,95
0,94
0,92
0,90
0,80
0,70
0,60
0,50
0,40
0,30
0,20
0,10
$\nu_{меж} = 7$
$\nu_{вну} = \infty$ 60 30 20 15 12 10 9 8 7 6
$\alpha = 0{,}05$
$\alpha = 0{,}01$
1 2 3
φ (для α = 0,01)
1 2 3 4
φ (для α = 0,05)


Чувствительность
0,99
0,98
0,97
0,96
0,95
0,94
0,92
0,90
0,80
0,70
0,60
0,50
0,40
0,30
0,20
0,10
$V_{нвж} = 8$
$V_{вну} = \infty$ 60 30 20 15 12 10 9 8 7 6
$\alpha = 0,05$
$\alpha = 0,01$
∞ 60 30 20 15 12 10 9 8 7 6
1 2 3
φ (для α = 0,01) →
← φ (для α = 0,05)
1 2 3 4

Решения задач

2.1. Среднее — 3,09; стандартное отклонение — 2,89; медиана — 2; 25-й процентиль — 1; 75-й процентиль — 5. Вряд ли данные извлечены из совокупности с нормальным распределением: среднее довольно сильно отличается от медианы, медиана гораздо ближе к 25-му процентилю, чем к 75-му, а значит, распределение асимметрично. Поскольку среднее почти равно стандартному отклонению, в случае нормального распределения примерно 15% значений было бы меньше нуля. Поэтому отсутствие отрицательных значений также говорит против нормальности распределения.

2.2. Среднее — 244; стандартное отклонение — 43; медиана — 235,5; 25-й процентиль — 211; 75-й процентиль — 246. Выборка вполне может быть извлечена из совокупности с нормальным распределением: медиана близка к среднему и находится примерно посредине между 25-м и 75-м процентилями. Сравните с предыдущей задачей.

2.3. Среднее — 5,4; стандартное отклонение — 7,6; медиана —

2,0; 25-й процентиль — 1,6; 75-й процентиль — 2,4. Выборку нельзя считать извлеченной из нормально распределенной совокупности: среднее не только не равно медиане, но даже превышает 75-й процентиль. Стандартное отклонение превышает среднее, при этом среди данных нет отрицательных значений (и не может быть по самой природе данных). Высокие значения среднего и стандартного отклонения обусловлены главным образом двумя «выпадающими» значениями — 19,0 и 23,6.

2.4. Это равномерное распределение: все значения от 1 до 6 выпадают с равной вероятностью. Среднее число очков — 3,5.

2.5. Это распределение выборочных средних, вычисленных по выборкам объемом 2, извлеченным из совокупности, описанной в предыдущей задаче. Среднее этого распределения равно среднему в совокупности, то есть 3,5, а стандартное отклонение (примерно 1,2) — это оценка стандартной ошибки среднего, вычисленного по выборке объемом 2.

2.6. Распределение по числу авторов не может быть нормальным уже потому, что нормальное распределение непрерывно, а число авторов всегда целое. Кроме того, все 4 средних меньше двух стандартных отклонений. Это значит, что в случае нормального распределения какое-то число статей должно было бы иметь отрицательное число авторов. Следовательно, мы имеем дело с асимметричным распределением наподобие распределения юпитериан по росту. К 1976 г. среднее число авторов резко возросло, однако стандартное отклонение возросло еще больше, так что теперь среднее меньше одного стандартного отклонения. Это говорит об увеличении асимметрии. Обратите внимание, что если бы Р. и С. Флетчеры привели не стандартное отклонение, а стандартную ошибку, мы не смогли бы прийти к этим выводам.

3.1. $F = 15{,}74$; $\nu_{меж} = 1$; $\nu_{вну} = 40$. Полученное значение F превышает критическое для данного числа степеней свободы и уровня значимости 0,01 (7,31). Различия статистически значимы. Можно утверждать, что гель с простагландином E_2 сокращал продолжительность родов.

3.2. $F = 64{,}18$; $\nu_{меж} = 4$; $\nu_{вну} = 995$. Различия статистически значимы (максимальную объемную скорость середины выдоха нельзя считать одинаковой во всех группах, $P < 0{,}01$).

3.3. $F = 35{,}25$; $\nu_{меж} = 2$; $\nu_{вну} = 207$; $P < 0{,}01$.

3.4. $F = 60{,}37$; $\nu_{меж} = 6$; $\nu_{вну} = 245$; $P < 0{,}01$.
3.5. $F = 2{,}52$; $\nu_{меж} = 1$; $\nu_{вну} = 70$; $P > 0{,}05$.
3.6. $F = 3{,}85$; $\nu_{меж} = 5$; $\nu_{вну} = 90$; $P < 0{,}01$.
3.7. $F = 8{,}19$; $\nu_{меж} = 3$; $\nu_{вну} = 79$; $P < 0{,}01$.
3.8. $F = 0{,}41$; $\nu_{меж} = 4$; $\nu_{вну} = 101$; $P > 0{,}05$.

4.1. Для среднего артериального давления $t = -1{,}97$, для общего периферического сосудистого сопротивления $t = -1{,}29$. Число степеней свободы в обоих случаях $\nu = 23$, при $\alpha = 0{,}05$ ему соответствует критическое значение $t = 2{,}069$. Следовательно, различия обоих гемодинамических показателей статистически не значимо.

4.2. $t = 3{,}14$; $\nu = 20$; $P < 0{,}01$. Различия статистически значимы, однако, вопреки первоначальным предположениям, нифедипин не повышает, а снижает артериальное давление.

4.3. Нет. $t = 1{,}33$; $\nu = 20$; $P > 0{,}05$. Нифедипин не влияет на диаметр коронарных артерий.

4.4. Задача 3.1: $t = 3{,}97$; $\nu = 40$; $P < 0{,}001$. Задача 3.5: $t = 1{,}59$ $\nu = 70$; $P > 0{,}05$.

4.5. Вот некоторые результаты попарных сравнений. Некурящие, работающие в помещении, где не курят, и пассивные курильщики — $t = 6{,}21$, выкуривающие небольшое число сигарет и выкуривающие среднее число сигарет — $t = 4{,}72$, выкуривающие среднее число сигарет и выкуривающие большое число сигарет — $t = 2{,}39$. Применим поправку Бонферрони. Поскольку имеется 5 групп, можно провести 10 попарных сравнений. Чтобы истинный уровень значимости остался равным 0,05, в каждом из сравнений уровень значимости следует принять равным $0{,}05/10 = 0{,}005$. Число степеней свободы $n = 995$. Таким образом, критическое значение t составляет 2,807. Отличия проходимости дыхательных путей у некурящих, работающих в помещении, где не курят, и пассивных курильщиков статистически значимы.

4.6. Некурящие, работающие в накуренном помещении (пассивные курильщики): $q' = 6{,}249$; $l = 5$. Выкуривающие небольшое число сигарет: $q' = 7{,}499$; $l = 5$. Выкуривающие среднее число сигарет: $q' = 12{,}220$; $l = 5$. Выкуривающие большое число сигарет: $q' = 14{,}580$; $l = 5$. Критическое значение q' при уровне значимости 0,01, числе степеней свободы 995 и $l = 5$ составляет 3,00. Следовательно, отличие некурящих, работающих в помещении, где не

курят, от пассивных курильщиков и от собственно курильщиков всех степеней злостности статистически значимо.

4.7. Не занимающиеся спортом и бегуны трусцой: $t = 5{,}616$. Не занимающиеся спортом и бегуны-марафонцы: $t = 8{,}214$. Бегуны трусцой и бегуны-марафонцы: $t = 2{,}598$. Чтобы истинный уровень значимости остался равным 0,05, в каждом из сравнений уровень значимости следует принять равным $0{,}05/3 = 0{,}017$. Число степеней свободы $\nu = 207$. Критический уровень t составляет 2,42. Все три группы различаются статистически значимо.

4.8. Бегуны трусцой: $t = 5{,}616$. Бегуны-марафонцы: $t = 8{,}214$. Поскольку в данном случае возможно только два парных сравнения, в каждом из них уровень значимости следует принять равным $0{,}05/2 = 0{,}025$. Число степеней свободы $\nu = 207$. Критический уровень t составляет 2,282. Таким образом, не занимающиеся спортом статистически значимо отличаются как от бегунов трусцой, так и от марафонцев. Обратите внимание, что мы получили те же значения t, что и в предыдущей задаче, но число возможных сравнений уменьшилось до 2, благодаря чему критический уровень t снизился. Однако при таком методе анализа мы не можем сделать никакого вывода о различиях бегунов трусцой и марафонцев.

4.9. Контрольная группа, 15 и 30 сигарет; 75 сигарет без тетрагидроканнабинолов и 50 сигарет; 75 и 150 сигарет.

4.10. Всего можно провести 6 сравнений. Контроль и дофамин в низкой дозе: $t = 0$. Контроль и дофамин в высокой дозе: $t = 3{,}171$. Контроль и нитропруссид натрия: $t = 4{,}228$. Дофамин в низкой дозе и дофамин в высокой дозе: $t = 2{,}569$. Дофамин в низкой дозе и нитропруссид натрия: $t = 3{,}426$. Дофамин в высокой дозе и нитропруссид натрия: $t = 0{,}964$. Уровень значимости в каждом из сравнений $0{,}05/6 = 0{,}0083$, число степеней свободы $\nu = 79$, соответствующий критический уровень t составляет 2,72. Итак, группы довольно четко разделились на контроль и дофамин в низкой дозе, с одной стороны, и дофамин в высокой дозе и нитропруссид натрия, с другой. Картину несколько портит сравнение дофамина в низкой и высокой дозе: значение t не достигает критического уровня, хотя и близко к нему. В такой ситуации большинство исследователей, вероятно, все же сочтет различие

этих групп статистически значимым, учитывая «жесткость» поправки Бонферрони, их вряд ли можно за это упрекнуть.

4.11. Результаты попарных сравнений:

Сравнение	Разность средних	q	l	Критическое значение q
Контроль и нитропруссид натрия	15 – 7 = 8	5,979	4	3,7
Контроль и дофамин в высокой дозе	15 – 9 = 6	4,485	3	3,4
Контроль и дофамин в низкой дозе	15 – 15 = 0	0,000	2	2,8
Дофамин в низкой дозе и нитропруссид натрия	15 – 7 = 8	4,845	3	3,4
Дофамин в низкой дозе и дофамин в высокой дозе	15 – 9 = 6	3,634	2	2,8
Дофамин в высокой дозе и нитропруссид натрия	9 – 7 = 2	1,365	2	2,8

Критические значения q для уровня значимости $\alpha' = 0{,}05$, числа степеней свободы $\nu = 79$ и соответствующих значений l приведены в правой колонке. Общий вывод тот же, что и в предыдущей задаче, при этом различие дофамина в низкой и высокой дозе теперь статистически значимо.

4.12. Групп слишком много, чтобы применить поправку Бонферрони: она окажется слишком «строгой». Применим поэтому критерий Ньюмена—Кейлса.

Упорядочим группы по убыванию среднего.

Группа	3	2	1	1	2	3
Отделение	Тер.	Хир.	Тер.	Хир.	Тер.	Хир.
Среднее	65,2	57,3	51,2	49,9	46,4	43,9
Стандартное отклонение	20,5	14,9	13,4	14,3	14,7	16,5

Проделаем стягивающие сравнения. Результат приведен в таблице на следующей странице. В правом столбце — критическое значение для уровня значимости $\alpha' = 0{,}05$.

Значение q превышает критическое только в первых 4 сравнениях. Таким образом, все группы можно объединить в две кате-

гории. К категории высокой опустошенности относятся медсестры 3-й группы терапевтических отделений и 2-й группы хирургических отделений, к категории умеренной опустошенности — все остальные. Отнесение медицинских сестер 2-й группы хирургических отделений к категории высокой опустошенности довольно условно — их можно было бы отнести и к категории умеренной опустошенности. При множественных сравнениях подобные ситуации встречаются, к сожалению, нередко.

Сравнение					
Группа, отделение	Группа, отделение	Разность средних	q	Интервал сравнения	Критическое значение q
3, тер.	3, хир.	65,2 – 43,9 = 21,3	5,362	6	4,1
3, тер.	2, тер.	65,2 – 46,4 = 18,8	4,733	5	3,9
3, тер.	1, хир.	65,2 – 49,9 = 15,3	3,852	4	3,7
3, тер.	1, тер.	65,2 – 51,2 = 14,0	3,525	3	3,4
3, тер.	2, хир.	65,2 – 57,3 = 7,9	1,989	2	2,8
2, хир.	3, хир.	57,3 – 43,9 = 13,4	3,374	5	3,9
2, хир.	2, тер,	57,3 – 46,4 = 10,9	2,744	4	3,7
2, хир.	1, хир.	57,3 – 49,9 = 7,4	1,863	3	3,4
2, хир.	1, тер.	57,3 – 51,2 = 6,1	1,536	2	2,8
1, тер.	3, хир.	51,2 – 43,9 = 7,3	1,838	4	3,7
1, тер.	2, тер.	51,2 – 46,4 = 4,8	1,208	3	3,4
1, тер.	1, хир.	51,2 – 49,9 = 1,3	0,327	2	2,8
1, хир.	3, хир.	49,9 – 43,9 = 6,0	1,511	3	3,4
1, хир.	2, тер.	49,9 – 46,4 = 3,5	0,881	2	2,8
2, тер.	3, хир.	46,4 – 43,9 = 2,5	0,629	2	2,8

5.1. Да, позволяют: $\chi^2 = 17{,}878$; $\nu = 1$; $P < 0{,}001$.

5.2. Значения χ^2 для исследованных признаков следующие: возраст матери — 11,852 ($P < 0{,}001$), время от окончания предыдущей беременности — 10,506 ($P < 0{,}005$), планировалась ли беременность — 3,144 ($P > 0{,}05$), повторная беременность — 1,571 ($P < 0{,}05$), курение во время беременности — 17,002 ($P < 0{,}001$), посещения врача во время беременности — 4,527 ($P < 0{,}05$), самый низкий гемоглобин во время беременности — 0,108

($P > 0{,}05$), раса — 0,527 ($P > 0{,}05$). (Число степеней свободы для расы — 2, для остальных признаков — 1.) Таким образом, факторы риска: возраст матери меньше 25 лет, время от окончания предыдущей беременности менее 1 года, курение во время беременности, возможно также менее 11 посещений врача во время беременности.

5.4. $\chi^2 = 7{,}288$; $\nu = 2$; $P < 0{,}05$, различия эффективности статистически значимы. Сравним ампициллин и цефалексин.

	Рецидив	
	есть	нет
Ампициллин	20	7
Цефалексин	14	2

$\chi^2 = 0{,}433$; $\nu = 1$; $P > 0{,}05$ (с поправкой Бонферрони), различия статистически не значимы. Объединим соответствующие строки и сравним ампициллин или цефалексин с триметопримом/сульфаметоксазолом.

	Рецидив	
	есть	нет
Ампициллин или цефалексин	34	9
Триметоприм/сульфаметоксазол	24	21

$\chi^2 = 5{,}387$; $\nu = 1$; $P < 0{,}05$ (с поправкой Бонферрони), различия статистически значимы. Итак, триметоприм/сульфаметоксазол превосходит как ампициллин, так и цефалексин, которые друг от друга не отличаются.

5.5. $\chi^2 = 74{,}925$; $\nu = 2$; $P < 0{,}001$. Связь заболеваемости с количеством выпитой воды статистически значима. Сравнив группы попарно (используя поправку Бонферрони), можно убедиться, что заболеваемость растет с количеством выпитой воды.

5.6. $\chi^2 = 48{,}698$; $\nu = 3$; $P < 0{,}001$, в целом различие долей статистически значимо. Разбиение таблицы показывает, что не отличаются 1946 от 1956 г. и 1966 от 1976 г. Далее, объединенная группа 1946 и 1956 гг. отличаются в лучшую сторону от объединенной группы 1966 и 1976 гг. Таким образом, между 1956 и 1966 г. ситуация изменилась к худшему.

5.7. $\chi^2 = 5{,}185$; $\nu = 1$; $P < 0{,}025$. Различия (в пользу хирургического лечения) статистически значимы.

5.8. Без антиангинальной терапии: в двух клетках ожидаемые числа меньше 5, поэтому следует применить точный критерий Фишера, он дает $P = 0{,}151$. Различия статистически не значимы. На фоне антиангинальной терапии: можно было бы применить критерий χ^2, однако для единообразия применим точный критерий Фишера: $P = 0{,}094$. Различия статистически не значимы.

5.9. $\chi^2 = 2{,}273$; $\nu = 1$; $P > 0{,}05$. Теперь статистически значимых различий нет.

5.10. $\chi^2 = 8{,}812$; $\nu = 1$; $P < 0{,}005$. Различия статистически значимы: в больнице среди страдающих болезнью Z доля больных X выше, чем среди страдающих болезнью Y. Как мы видели, эти различия обусловлены исключительно разной вероятностью госпитализации при этих болезнях.

6.1. $\delta/\sigma = 1{,}1$; $n = 9$, чувствительность — 63% (рис. 6.9).

6.2. $\delta/\sigma = 0{,}55$, чувствительность — 80%, $n = 40$ (рис. 6.9).

6.3. Среднее артериальное давление: $\delta = 0{,}25 \times 76{,}8 = 19{,}2$; $\sigma = 17{,}8$ (объединенная оценка); $\delta/\sigma = 1{,}08$; $n = 9$ (численность меньшей из групп). По рис. 6.9 находим чувствительность — 63%. Общее периферическое сосудистое сопротивление: $\delta/\sigma = 553/1154 = 0{,}48$; $n = 9$; чувствительность примерно 13%.

6.4. Примерно 70%.

6.5. Примерно 50 крыс в каждой группе.

6.6. Обозначим истинную долю p, а ее выборочную оценку $\hat{p}$. Наименьшее различие долей, которое мы хотим выявить, обозначим Δp. Объем каждой из выборок равен n.

Если нулевая гипотеза об отсутствии различий верна, то величина $z = \Delta\hat{p}/s_{\Delta\hat{p}}$ подчиняется стандартному нормальному распределению. Кроме того, при справедливости нулевой гипотезы, $\hat{p}_1$ и $\hat{p}_2$ — это две оценки одной и той же доли. Тогда ее объединенная оценка — $\hat{p} = (\hat{p}_1 + \hat{p}_2)/2 = (0{,}3 + 0{,}9)/2 = 0{,}6$, а стандартная ошибка разности:

$$s_{\Delta\hat{p}} = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n}+\frac{1}{n}\right)} = \frac{0{,}692}{\sqrt{n}}.$$

При уровне значимости $\alpha = 0{,}05$ критическое значение z составляет $z_\alpha = 1{,}960$. Ему соответствует

$$\Delta\hat{p} = z_\alpha s_{\Delta\hat{p}} = 1{,}960\frac{0{,}692}{\sqrt{n}} = \frac{1{,}356}{\sqrt{n}}.$$

Истинные доли p_1 и p_2 составляют соответственно 0,3 и 0,9, тогда их разность $\Delta p = p_2 - p_1 = 0{,}9 - 0{,}3 = 0{,}6$, а ее стандартная ошибка

$$s_{\Delta p} = \sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n}} = \frac{0{,}547}{\sqrt{n}}.$$

Величина $z = (\Delta\hat{p} - \Delta p)/s_{\Delta p}$ подчиняется стандартному нормальному распределению. Поскольку необходимая чувствительность 90%, найдем по таблице 6.4 значение z, правее которого лежит 90% всех значений. Это $z = -1{,}282$. Ему соответствует

$$\Delta\hat{p} = \Delta p + z_\beta s_{\Delta p} = 0{,}6 + (-1{,}282)\frac{0{,}547}{\sqrt{n}} = 0{,}6 - \frac{0{,}701}{\sqrt{n}}.$$

Приравняем обе оценки $\Delta\hat{p}$:

$$\frac{1{,}356}{\sqrt{n}} = 0{,}6 - \frac{0{,}701}{\sqrt{n}}.$$

Тогда $n = 11{,}7$, то есть в каждой группе должно быть 12 больных.

6.7. 80%.

6.8. На 5 мг% — 36%, на 10 мг% — 95%.

6.9. 183.

6.10. При данной численности групп и ожидаемом эффекте лечения мы получим следующие доли в клетках.

	Рецидив		
	Есть	Нет	Всего
Ампициллин	0,205	0,102	0,307
Триметоприм/сульфаметоксазол	0,341	0,170	0,511
Цефалексин	0,061	0,121	0,182
Всего	0,607	0,393	1

$\varphi = 1{,}4$; $\nu_{меж} = (3-1)(2-1) = 2$; по рис. 6.10 находим чувствительность — 58%.

6.11. 135.

7.1. 90% доверительные интервалы: 1,8—2,2; 2,1—2,5; 2,6—3,0; 3,9—5,9. 95% доверительные интервалы: 1,8—2,2; 2,0—2,6; 2,6—3,0; 3,7—6,1. (С округлением до 1 знака после запятой.)

7.2. Воспользовавшись рис. 7.4, найдем: для контрольной группы 6—42%, для группы, получавшей гель с простагландином E_2, — 5—40%. 95% доверительный интервал для разности долей от −15 до 33% (можно использовать приближение с помощью нормального распределения). Разность долей статистически не значима.

7.3. 95% доверительный интервал разности средней продолжительности родов — от 2,7 до 8,1. Различия статистически значимы ($P < 0{,}05$).

7.4. При включенном приборе не чувствовали боли 80%, по рис. 7.4 находим 95% доверительный интервал — от 60 до 90%. При выключенном приборе доля — 15%, 95% доверительный интервал — примерно от 3 до 40%. Доверительные интервалы не перекрываются, поэтому различия статистически значимы.

7.5. Некурящие, работающие в помещении, где не курят, — 3,07—3,27; пассивные курильщики — 2,62—2,82; выкуривающие небольшое число сигарет — 2,53—2,73; выкуривающие среднее число сигарет — 2,19—2,39; выкуривающие большое число сигарет — 2,02—2,22. Объединив группы с перекрывающимися доверительными интервалами, получим 3 категории: первая — некурящие, работающие в помещении, где не курят, вторая — пассивные курильщики и выкуривающие небольшое число сигарет, третья — выкуривающие среднее и большое число сигарет.

7.6. 1946 г. — 17—31%; 1956 г. — 22—36%; 1966 г. — 43—59%; 1976 г. — 48—64%.

7.7. Для 90% значений: 121—367, для 95% значений: 108—380.

8.1. а) $a = 3{,}0$; $b = 1{,}3$; $r = 0{,}79$; б) $a = 5{,}1$; $b = 1{,}2$; $r = 0{,}94$; в) $a = 5{,}6$; $b = 1{,}2$; $r = 0{,}97$. С увеличением диапазона данных растет и коэффициент корреляции.

8.2. а) $a = 24{,}3$; $b = 0{,}36$; $r = 0{,}561$; б) $a = 0{,}5$; $b = 1{,}15$; $r = 0{,}599$. Первый пример показывает, сколь большое влияние может иметь единственная точка. Второй пример показывает, как важ-

но нанести данные на график, прежде чем приступить в регрессионному анализу: здесь выборка явно разнородна и может быть описана двумя различными зависимостями. Условия применимости регрессионного анализа не соблюдены, и попытка выразить связь единственной линией регрессии несостоятельна.

8.3. Во всех четырех экспериментах $a = 3{,}0$; $b = 0{,}5$; $r = 0{,}82$. Условия применимости регрессионного анализа соблюдены только в первом эксперименте.

8.4. Да. $r = -0{,}68$; $P < 0{,}05$.

8.5. Применим метод Блэнда—Алтмана. Для конечно-диастолического объема: средняя разность −3 мл, стандартное отклонение 14 мл. Для конечно-систолического объема: средняя разность 4 мл, стандартное отклонение 10 мл. Это говорит о хорошей согласованности по обоим показателям. При графическом анализе видно, что в обоих случаях разность увеличивается с ростом среднего показателя.

8.6. При калорийности 37 ккал/кг: $a = -44{,}3$; $b = 0{,}34$; при калорийности 33 ккал/кг: $a = -34{,}8$; $b = 0{,}35$. Для разности коэффициентов сдвига $t = 1{,}551$; $n = 20$; $P > 0{,}05$, для разности коэффициентов наклона: $t = 0{,}097$; $\nu = 20$; $P > 0{,}05$. При калорийности 37 ккал/кг нулевой азотистый баланс достигается при поступлении азота 130 мг/кг.

8.7. Оценки согласованы достаточно хорошо: коэффициент ранговой корреляции Спирмена $r_s = 0{,}89$; $P < 0{,}002$. Впрочем, тут можно применить и коэффициент корреляции Пирсона, он даст $r = 0{,}94$; $P < 0{,}001$.

8.8. Коэффициент ранговой корреляции Спирмена $r_s = 0{,}899$; $P < 0{,}001$. Визуальная оценка достаточно хорошо соответствует результатам взвешивания. Однако, если нанести данные на график, можно заметить, что при большом налете визуальная оценка занижает результат. Дополнительный вопрос: нельзя ли в этом случае воспользоваться методом Блэнда—Алтмана?

8.9. Коэффициент ранговой корреляции Спирмена $r_s = 0{,}85$; $P < 0{,}001$. Данные подтверждают гипотезу о связи между адгезивностью эритроцитов и тяжестью серповидноклеточной анемии.

8.10. 0,999.

8.11. 20.

8.12. Для коэффициентов наклона $t = -2{,}137$; $\nu = 26$; $P < 0{,}05$. Для коэффициентов сдвига $t = -2{,}396$; $\nu = 26$; $P < 0{,}05$. При сравнении линий регрессии в целом имеем: $F = 6{,}657$; $\nu_{меж} = 2$; $\nu_{вну} = 2$. Различия линий регрессии статистически значимы.

9.1. Применив парный критерий Стьюдента, получим: $t = 4{,}69$; $\nu = 9$; $P < 0{,}002$. Полоскание с хлоргексидином более эффективно.

9.2. Антитела к пневмококкам: $t = 3{,}2$; $\nu = 19$; $P < 0{,}01$, изменение статистически значимо. Антитела к стрептококкам: $t = 1{,}849$, $\nu = 19$; $P > 0{,}05$, изменение статистически не значимо.

9.3. Антитела к пневмококкам: $\delta = 306$ (средний начальный уровень), $\sigma = 621$ (стандартное отклонение изменения), $\varphi = 0{,}49$. По рис. 6.9 находим чувствительность — примерно 50%. Антитела к стрептококкам: $\delta = 0{,}74$; $\sigma = 2{,}85$; $\varphi = 0{,}26$, чувствительность около 20%.

9.4. Антитела к пневмококкам: $F = 10{,}073$. Антитела к стрептококкам: $F = 3{,}422$. В общем случае $F = t^2$.

9.5. Дисперсионный анализ повторных наблюдений дает $F = 184{,}50$; $\nu_{меж} = 3$; $\nu_{вну} = 33$. Различия статистически значимы. Попарные сравнения с помощью критерия Стьюдента и поправки Бонферрони показывают, что результаты до курения и вдыхания окиси углерода статистически значимо не отличаются друг от друга, но отличаются от результатов после курения и вдыхания окиси углерода; те, в свою очередь, статистически значимо отличаются друг от друга.

9.6. Применив дисперсионный анализ повторных наблюдений, получим $F = 5{,}04$. Критический уровень F при $\alpha = 0{,}05$ и числе степеней свободы $\nu_{меж} = 2$ и $\nu_{вну} = 6$ составляет 5,14, то есть несколько превышает полученное.

9.7. Дисперсионный анализ повторных измерений дает $F = 4{,}56$; $\nu_{меж} = 2$; $\nu_{вну} = 12$. Различия статистически значимы. Критерий Стьюдента с поправкой Бонферрони показывает, что объем пищи при исходном давлении в поясе 20 мм рт. ст. меньше, чем при давлении 0 и 10 мм рт. ст. Результаты при 0 и 10 мм рт. ст. друг от друга статистически значимо не отличаются.

9.8. $\delta = 100$, в качестве σ возьмем квадратный корень из остаточной дисперсии, равный 74. Тогда $\varphi = 1{,}35$, чувствительность примерно 50%.

9.9. Применим критерий Мак-Нимара: $\chi^2 = 4{,}225$; $\nu = 1$, $P < 0{,}05$. Индометацин эффективен.

9.10. Теперь данные представлены в виде обычной таблицы сопряженности; $\chi^2 = 2{,}402$; $\nu = 1$, $P > 0{,}05$. Игнорируя парность наблюдений, мы теряем часть информации, в результате чувствительность снижается.

10.1. Изменение расходов на обследование: $W = -72$, $n = 12$ (одно нулевое изменение), $P < 0{,}02$. Изменение расходов на лечение: $W = -28$, $n = 13$, $P > 0{,}048$. Расходы на обследование снизились, на лечение остались прежними. Статистически значимой связи между расходами на обследование и лечение нет: $r_s = 0{,}201$, $P > 0{,}05$.

10.2. Критерий Стьюдента дает $t = 1{,}908$, $\nu = 22$, $P > 0{,}05$. Статистически значимых различий нет. Применим критерий Манна—Уитни. $T = 203$, $n = 12$. Можно применить приближение нормальным распределением: $z = 3{,}041$, $P < 0{,}005$. Различия статистически значимы. Распределение далеко от нормального, поэтому параметрический критерий проигрывает в чувствительности непараметрическому.

10.3. $H = 20{,}66$; $\nu = 2$, $P < 0{,}001$. Различия статистически значимы.

10.4. Задача 9.5: $\chi_r^2 = 32{,}4$; $\nu = 3$; $P < 0{,}001$. Задача 9.6: $\chi_r^2 = 6{,}5$; $k = 3$; $n = 4$; $P = 0{,}042$. Различия статистически значимы.

10.5. $T = 54$; $n_6 = 6$; $n_б = 22$; $z_T = -1{,}848$; $P > 0{,}05$.

10.6. Применим критерий Манна—Уитни с поправкой Йейтса: $z_T = 3{,}425$; $P < 0{,}001$. Различия статистически значимы.

10.7. $H = 18{,}36$; $\nu = 2$; $P < 0{,}001$. Различия групп статистически значимы. Попарное сравнение с помощью критерия Данна показывает следующее:

Сравнение групп	Q	$P < 0{,}05$
3 и 1	4,112	Да
3 и 2	2,229	Нет
2 и 1	0,975	Нет

Группы не распадаются на различающиеся категории, кроме того, различия 2-й группы (поражение только правой коронарной артерии) и 3-й (поражение левой или обеих коронарных

артерий) статистически не значимы. Предполагавшееся диагностическое значение исследуемого показателя не доказано.

10.8. Да, критерий G ничем не хуже прочих (если не считать проблемы: что делать, если показатель не изменился).

Для $n = 4$ распределение его значений таково:

G	Вероятность
0	1/16
1	4/16
2	6/16
3	4/16
4	1/16

Для случая $n = 6$:

G	Вероятность
0	1/64
1	6/64
2	15/64
3	20/64
4	15/64
5	6/64
6	1/64

При $n = 4$ вероятность получить даже самые маловероятные значения — 0 или 64 составляет $1/16 + 1/16 = 1/8 = 0{,}125$. В этом случае мы не можем указать критическое значение для 5% уровня значимости (обратите внимание, что при этой численности группы критерий Уилкоксона тоже не даст результата). При $n = 6$ критические значения — 0 и 6, соответствующий уровень значимости $1/64 + 1/64 = 1/32 = 0{,}31$.

11.1. Воспользуемся логранговым критерием. Сумма разностей ожидаемого и наблюдаемого числа смертей $U_L = -13{,}243$, ее стандартная ошибка $s_{U_L} = 3{,}090$. Таким образом, $z = -4{,}285$ (с поправкой Йейтса $z \doteqdot -4{,}124$). Различия выживаемости статистически значимы ($P < 0{,}001$). Выживаемость приведена в таблице.

Высокие оценки		Низкие оценки активности	
Месяцы	Выживаемость	Месяцы	Выживаемость
14	0,988	6	0,967
20	0,963	12	0,934
24	0,925	18	0,867
28	0,912	24	0,850
30	0,887	28	0,782
38	0,861	32	0,714
48	0,834	36	0,643
		42	0,584
		47	0,521
		48	0,479

11.2. Выживаемость представлена в таблице.

			95% доверительный интервал	
Время, месяцы	Выживаемость	Стандартная ошибка	нижняя граница	верхняя граница
1	0,971	0,029	0,914	1,000
2	0,941	0,040	0,862	1,000
3	0,853	0,061	0,734	0,972
4	0,824	0,065	0,695	0,952
5	0,794	0,069	0,658	0,930
6	0,765	0,073	0,622	0,907
7	0,706	0,078	0,553	0,859
8	0,676	0,080	0,519	0,834
9	0,647	0,082	0,486	0,808
12	0,579	0,086	0,410	0,748
13	0,545	0,088	0,373	0,717
15	0,511	0,089	0,337	0,684
16	0,409	0,088	0,235	0,582
20	0,307	0,084	0,143	0,470
21	0,272	0,081	0,114	0,431
28	0,234	0,078	0,080	0,387
34	0,195	0,074	0,049	0,340
56	0,130	0,072	0,000	0,272
62	0,065	0,058	0,000	0,179
84	0,000	0,000	0,000	0,000

11.3. (а) Выживаемость и 95% доверительные интервалы представлены в таблице.

Месяцы	Выживаемость	95% доверительный интервал	
		нижняя граница	верхняя граница
1975—1979 гг.			
2	0,940	0,873	1,000
4	0,860	0,764	0,956
6	0,800	0,688	0,912
8	0,720	0,597	0,843
12	0,679	0,550	0,808
14	0,617	0,482	0,752
18	0,574	0,435	0,713
24	0,552	0,413	0,691
30	0,508	0,367	0,649
36	0,486	0,345	0,627
54	0,463	0,322	0,604
56	0,440	0,299	0,581
60	0,417	0,276	0,558
1980—1984 гг.			
2	0,920	0,846	0,994
4	0,900	0,818	0,982
6	0,840	0,738	0,942
8	0,640	0,507	0,773
12	0,560	0,423	0,697
14	0,500	0,361	0,639
18	0,457	0,318	0,596
22	0,435	0,296	0,574
24	0,391	0,254	0,528
30	0,326	0,193	0,459
36	0,283	0,156	0,410
48	0,236	0,114	0,358
60	0,212	0,094	0,330

(б) Медиана выживаемости составила 36 мес в 1975—1979 гг. и 14 мес в 1980—1984 гг. (в) Логранговый критерий дает $z = -1{,}777$ (с поправкой Йейтса $z = -1{,}648$), что ниже критического значе-

ния для $\alpha = 0{,}05$; различия выживаемости статистически не значимы. (г) Чувствительность составляет 0,62. (д) Число смертей 104, суммарная численность групп 149 (для снижения $S(\infty)$ до 0,20); число смертей 65, суммарная численность групп 89 (для снижения $S(\infty)$ до 0,15).

Предметный указатель