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 венецианцев. 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 чаше 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

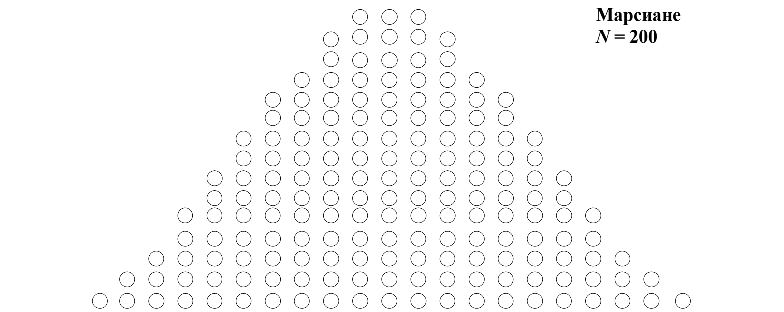

Рис. 2.1. Распределение марсиан по росту. Каждому марсианину соответствует кружок. Обратите внимание, что марсиан среднего роста (около 40 см) больше всего и что высокорослых столько же, сколько коротышек — распределение симметрично.

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 обоих 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*.

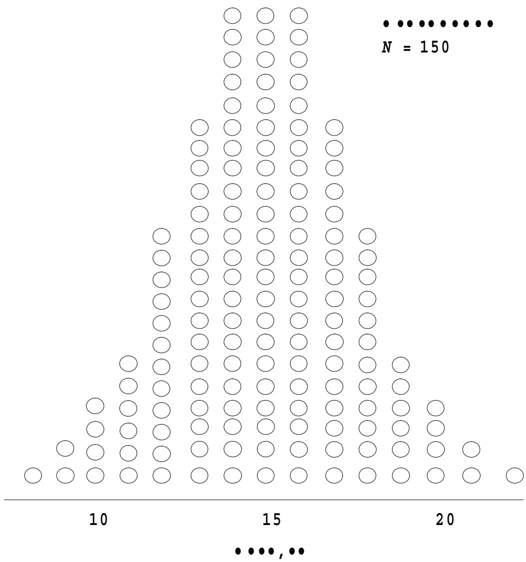

Рис. 2.2. Распределение венерианцев по росту. Венерианцы ниже марсиан, разброс значений меньше. Однако по форме распределения, напоминающей колокол, венерианцы и марсиане схожи друг с другом.

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

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

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:

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

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»):

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

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
Венерианцы	150	15	2,5

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

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

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

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-

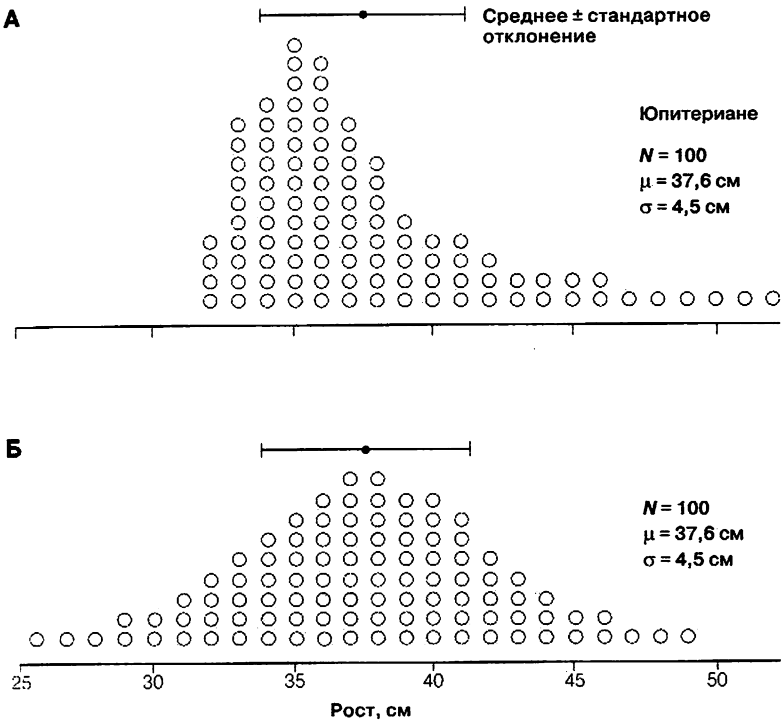

Рис. 2.3. Если распределение асимметрично полагаться на среднее и стандартное отклонение нельзя. **А.** Распределение юпитериан по росту. **Б.** Нормальное распределение с теми же средним и стандартным отклонением, не смотря на тождественность параметров, оно ничуть не похоже на реальное распределение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

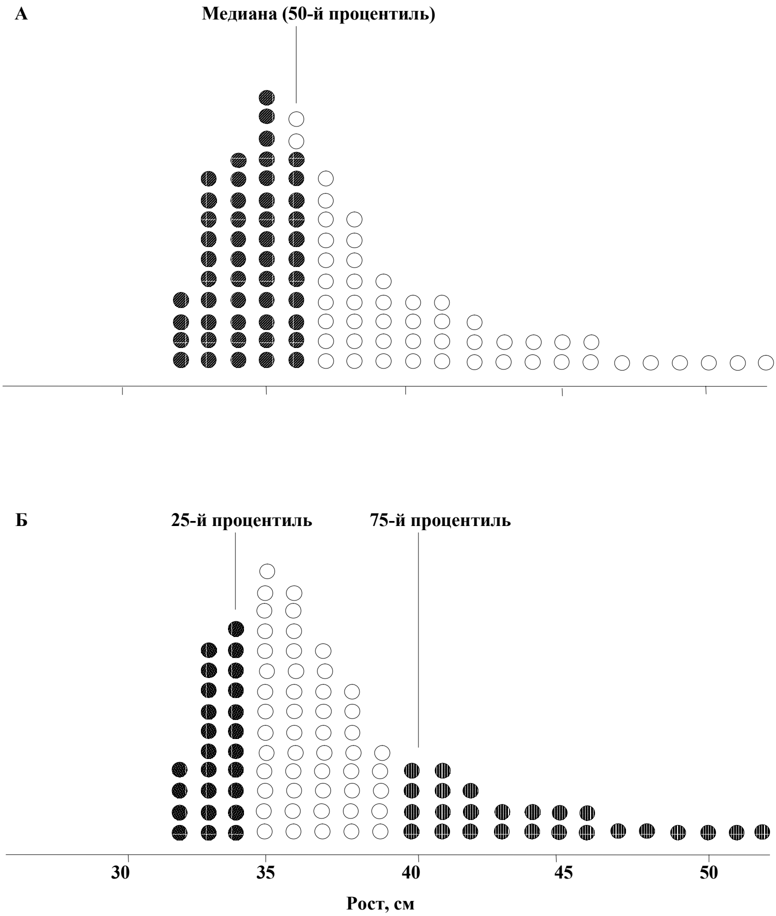

Рис. 2.4. Для описания асимметричного распределения следует использовать медиану и процентили. Медиана — это значение, которое делит распределение пополам. **А.** Медиана роста юпитериан — 36 см. **Б.** 25-й и 75-й процентили отсекают четверть самых низких и четверть самых высоких юпитериан 25-й процентиль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

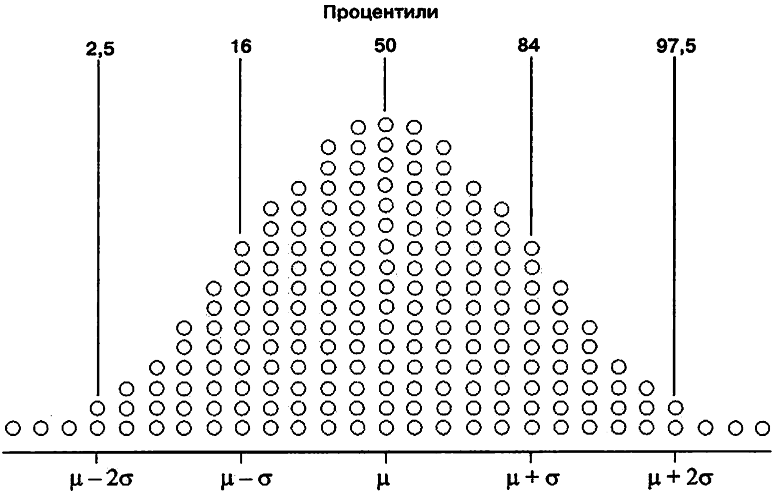

Рис. 2.5. Нормальное распределение, соответствие между числом стандартных отклонений от среднего и процентилями.

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

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

где n – объем выборки.

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

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

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

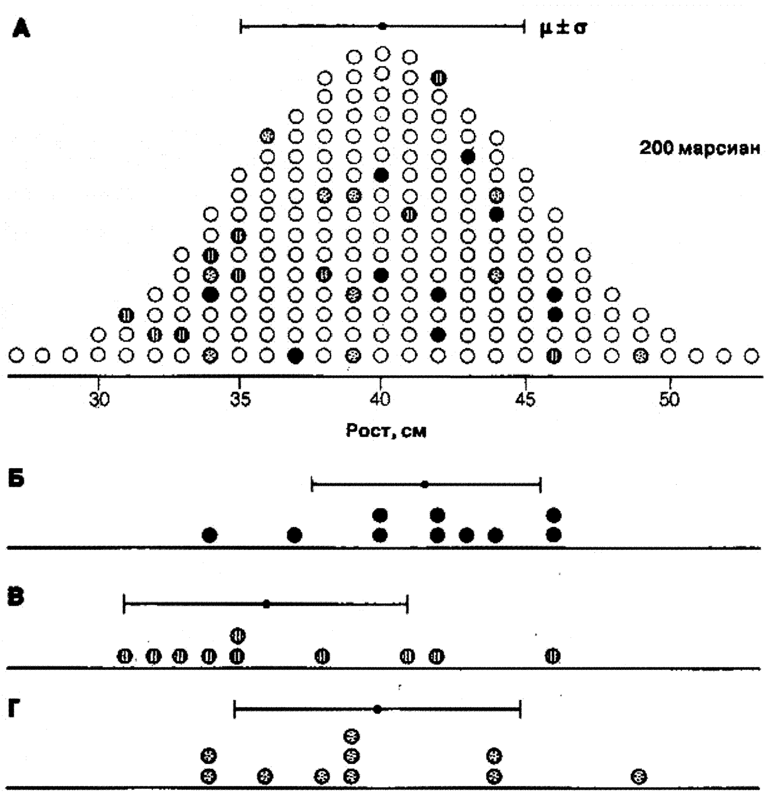

Рис. 2.6. Три случайные выборки из одной совокупности дают три разных оценки среднего и стандартного отклонения.

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

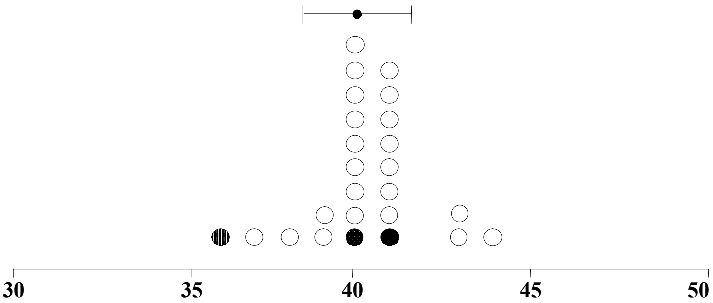

Рис. 2.7. Такое распределение мы получим, выбрав 25 раз по 10 марсиан из совокупности представленной на рис 2 6А, и рассчитав среднее для каждой выборки (средние для трех выборок с рис. 2.6 показаны заполненными кружками). Если построить распределение средних для всех возможных выборок, оно окажется 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 ($\overline{X}$ = 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 кружок, 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 $\bar{X}_{\bar{X}}$. Аналогично, стандартное отклонение обозначим $s_{\bar{X}}$. По формулам для среднего и стандартного отклонения находим $\bar{X}_{\bar{X}}= 40$ см и $s_{\bar{X}}= 1,6$см.

Среднее выборочных средних $\bar{X}_{\bar{X}}$ 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 $s_{\bar{X}}$ стандартному отклонению, σ совокупности из 200 марсиан? Стандартное отклонение для совокупности выборочных средних $s_{\bar{X}}$ равно 1,6 см, а стандартное отклонение самой совокупности — 5 см. Почему $s_{\bar{X}}$ меньше, чем σ? В общих чертах это можно понять, если учесть, что в случайную выборку редко будут попадать одни только коротышки и одни гиганты. Чаше 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 выборки 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.

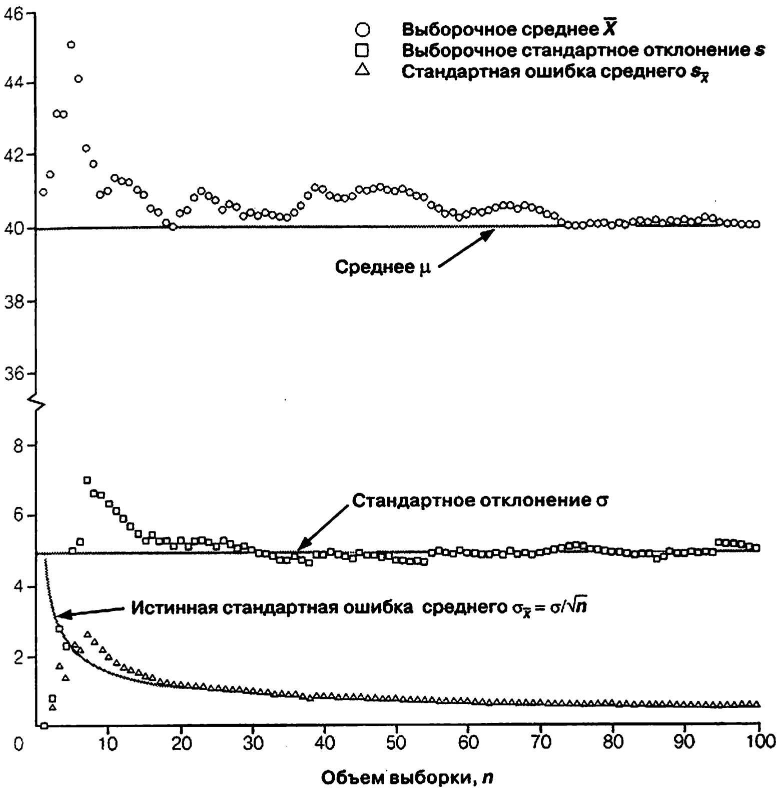

Рис. 2.8. С увеличением объема выборки возрастает точность оценки параметров распределения. Выборочное среднее $\bar{X}$ стремится к среднему в совокупности μ выборочное стандартное отклонение s стремится к стандартному отклонению в совокупности σ, а стандартная ошибка среднего стремится к нулю.

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 отклонения. 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; 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 seventy. *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 человеке 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 $\alpha = 0{,}05$.

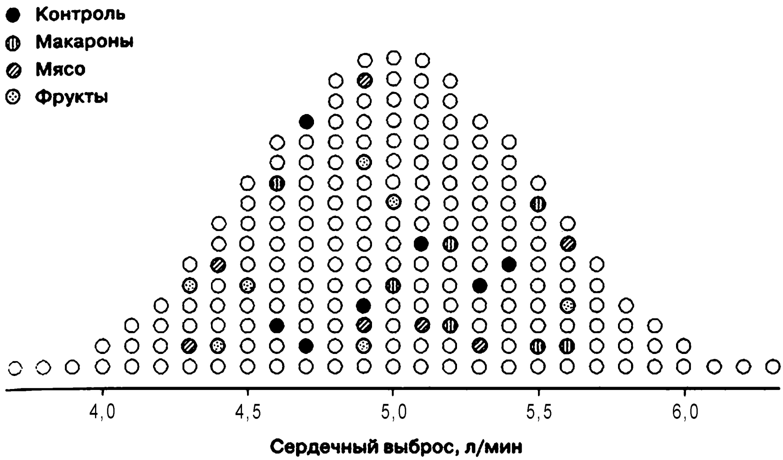

Рис. 3.1. Распределение жителей городка по величине сердечного выброса. Диета не влияет на сердечный выброс, и экспериментальные группы представляют собой просто четыре случайные выборки из нормально распределенной совокупности.

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-

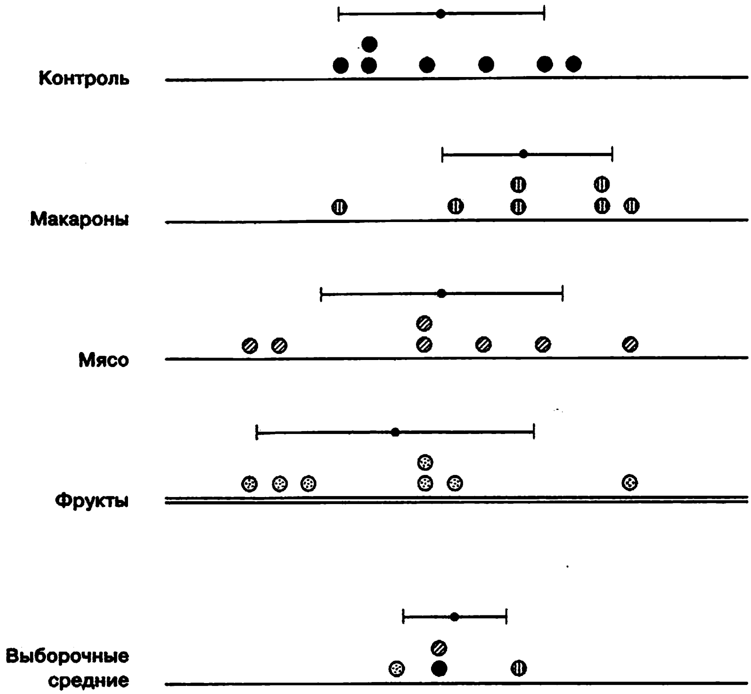

Рис. 3.2. Исследователь не может наблюдать совокупность, все, чем он располагает – это его экспериментальные группы. На этом рисунке данные с рис. 3.1 представлены такими, какими их видит исследователь. Результаты в разных группах несколько различаются. Вызваны эти различия диетой или просто случайностью? Внизу рисунка показаны средние значения 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

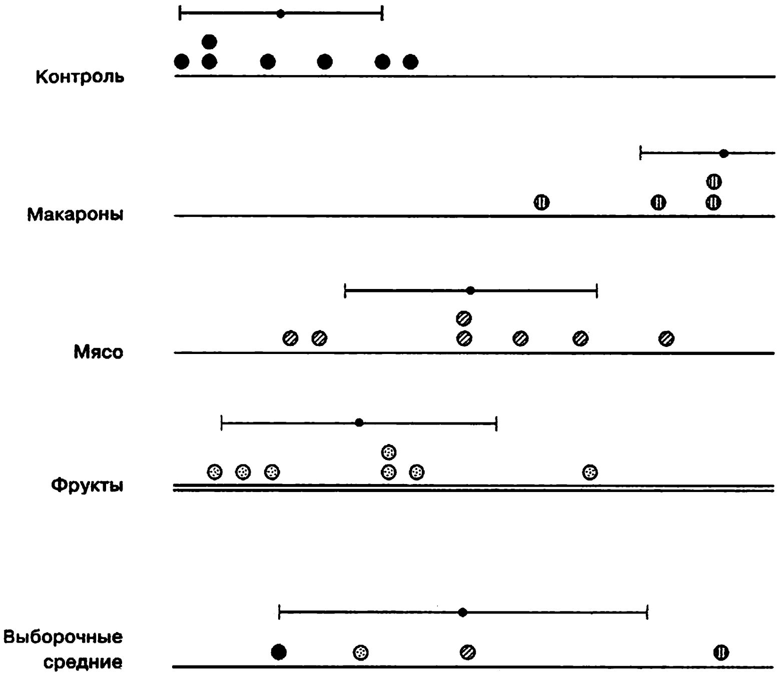

Рис. 3.3. Те же группы что на предыдущих рисунках; теперь они раздвинуты по горизонтальной оси. Вряд ли такие различия можно отнести на счет случайности — влияние диеты налицо! Обратите внимание, что разброс выборочных средних превышает разброс внутри групп. На предыдущем рисунке картина была иной, —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

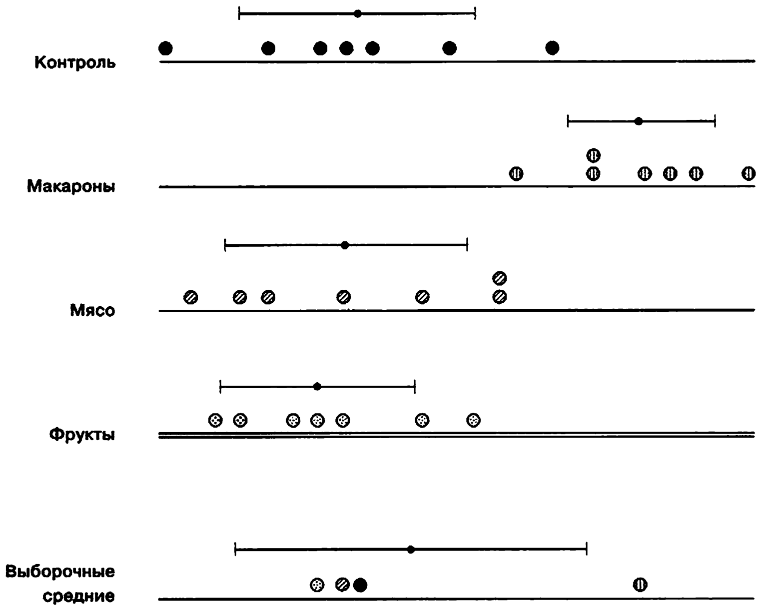

Рис. 3.4. Еще один возможный исход эксперимента с диетой. В трех группах средние примерно равны и только в группе макаронной диеты сердечный выброс явно повысился. Такой результат, как и предыдущий никто не отнесет на счет случайности. И снова разброс выборочных средних превышает разброс внутри групп.

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

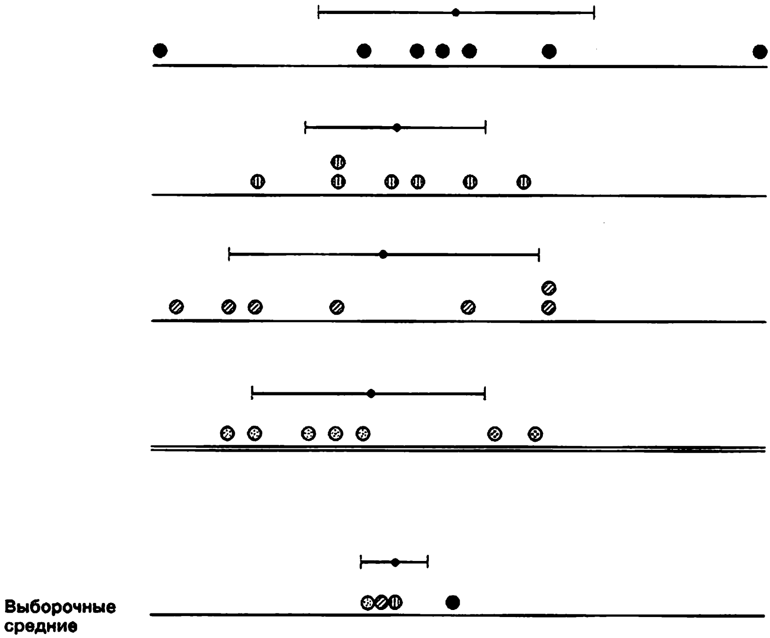

Рис. 3.5. Еще один набор из четырех случайных выборок по семь человек в каждой, извлеченых из совокупности в 200 человек (население городка, где изучали влияние диеты на сердечный выброс).

распределения. Более подробно мы рассмотрим непараметрические 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 $s^2_{\text{вну}}$.

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

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. Напомним, 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 $s_{\text{меж}}^2$.

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

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*:

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

или

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

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. Полученные 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 названии 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

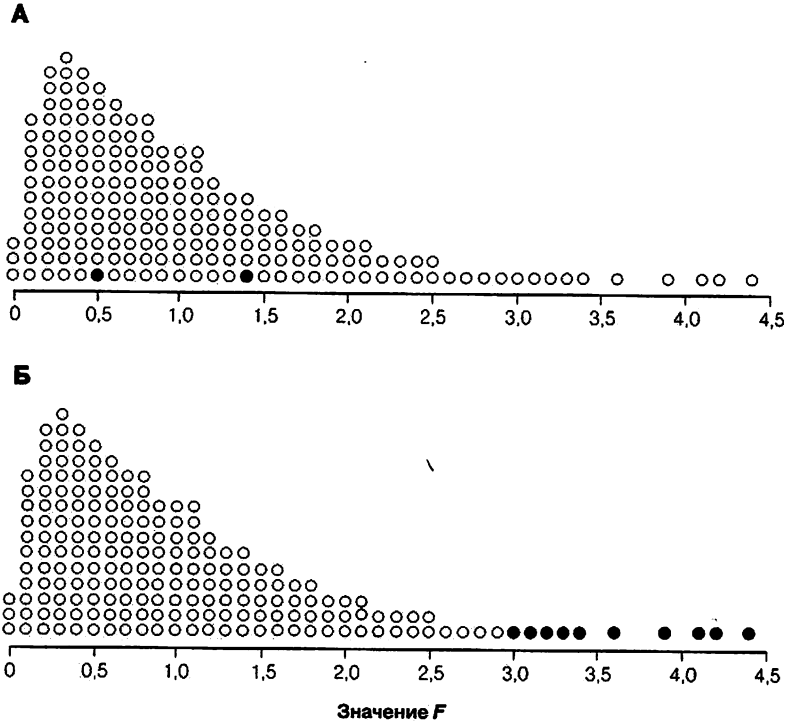

Рис. 3.6. А. Четыре случайные выборки по 7 человек в каждой извлекли из той же совокупности (население городка) 200 раз. Каждый раз рассчитывали значение *F* и наносили его на график. Результаты для выборок с рис. 3.2 и 3.5 помечены черным. **Б.** Десять наибольших значений помечень черньм. Область черных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

каждой из групп минус единица $\nu_{вну} = m(n-1)$. 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, $\nu_{меж}$ и $\nu_{вну}$ (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.

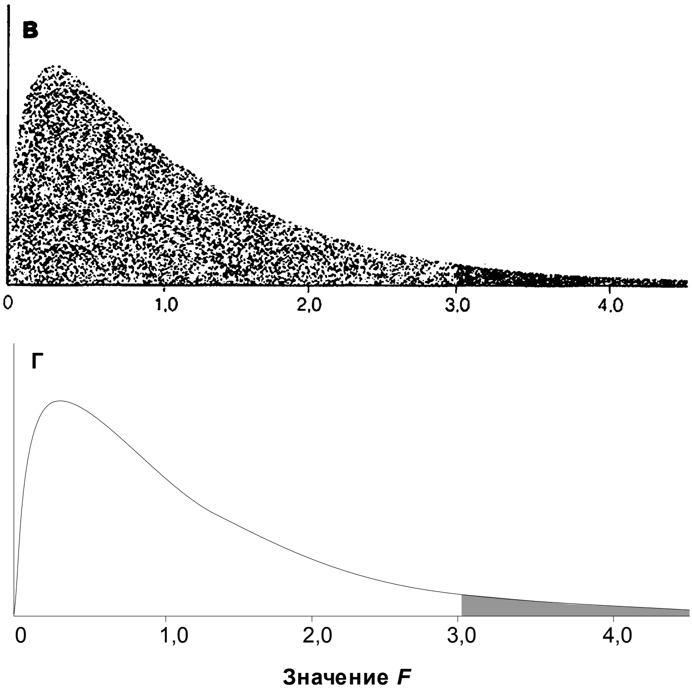

Рис. 3.6. (продолжение). В. Из той же совокупности извлекли все воэможнье наборы из 4 выборок по 7 человек в каждой и построили распределение F. Отдельные значения слились, превратившись в песчинки. 5% песчинок с самыми большими значениями F помечены черным. **Г.** Такое распределение F получится, если извлекать выборки из бесконечной совокупности. Пяти процентам самых высоких значений F соответствует заштрихованная область (ее площадь составляет 5% от общей площади всей 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	199	216	225	230	234	237	239	241	242	243	244	245	246	248	249	250	251	252	253	253	254	254	254
	4052	**4999**	**5404**	**5624**	**5764**	**5859**	**5928**	**5981**	**6022**	**6056**	**6083**	**6107**	**6143**	**6170**	**6209**	**6234**	**6260**	**6286**	**6302**	**6324**	**6334**	**6350**	**6360**	**6366**
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,40	19,41	19,42	19,43	19,45	19,45	19,46	19,47	19,48	19,48	19,49	19,49	19,49	19,50
	98,50	**99,00**	**99,16**	**99,25**	**99,30**	**99,33**	**99,36**	**99,38**	**99,39**	**99,40**	**99,41**	**99,42**	**99,43**	**99,44**	**99,45**	**99,46**	**99,47**	**99,48**	**99,48**	**99,48**	**99,49**	**99,49**	**99,50**	**99,50**
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,76	8,74	8,71	8,69	8,66	8,64	8,62	8,59	8,58	8,56	8,55	8,54	8,53	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**	**26,28**	**26,24**	**26,18**	**26,15**	**26,13**
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,94	5,91	5,87	5,84	5,80	5,77	5,75	5,72	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**	**14,55**	**14,45**	**14,37**	**14,25**	**14,15**	**14,02**	**13,93**	**13,84**	**13,75**	**13,69**	**13,61**	**13,58**	**13,52**	**13,49**	**13,46**
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,70	4,68	4,64	4,60	4,56	4,53	4,50	4,46	4,44	4,42	4,41	4,39	4,37	4,37
	16,26	**13,27**	**12,06**	**11,39**	**10,97**	**10,67**	**10,46**	**10,29**	**10,16**	**10,05**	**9,96**	**9,89**	**9,77**	**9,68**	**9,55**	**9,47**	**9,38**	**9,29**	**9,24**	**9,17**	**9,13**	**9,08**	**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	3,73	3,71	3,69	3,68	3,67
	13,75	**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**	**7,40**	**7,31**	**7,23**	**7,14**	**7,09**	**7,02**	**6,99**	**6,93**	**6,90**	**6,88**
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,60	3,57	3,53	3,49	3,44	3,41	3,38	3,34	3,32	3,29	3,27	3,25	3,24	3,23
	12,25	**9,55**	**8,45**	**7,85**	**7,46**	**7,19**	**6,99**	**6,84**	**6,72**	**6,62**	**6,54**	**6,47**	**6,36**	**6,28**	**6,16**	**6,07**	**5,99**	**5,91**	**5,86**	**5,79**	**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	3,35	3,31	3,28	3,24	3,20	3,15	3,12	3,08	3,04	3,02	2,99	2,97	2,95	2,94	2,93
	11,26	**8,65**	**7,59**	**7,01**	**6,63**	**6,37**	**6,18**	**6,03**	**5,91**	**5,81**	**5,73**	**5,67**	**5,56**	**5,48**	**5,36**	**5,28**	**5,20**	**5,12**	**5,07**	**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	3,14	3,10	3,07	3,03	2,99	2,94	2,90	2,86	2,83	2,80	2,77	2,76	2,73	2,72	2,71
	10,56	**8,02**	**6,99**	**6,42**	**6,06**	**5,80**	**5,61**	**5,47**	**5,35**	**5,26**	**5,18**	**5,11**	**5,01**	**4,92**	**4,81**	**4,73**	**4,65**	**4,57**	**4,52**	**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	2,98	2,94	2,91	2,86	2,83	2,77	2,74	2,70	2,66	2,64	2,60	2,59	2,56	2,55	2,54
	10,04	**7,56**	**6,55**	**5,99**	**5,64**	**5,39**	**5,20**	**5,06**	**4,94**	**4,85**	**4,77**	**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	2,85	2,82	2,79	2,74	2,70	2,65	2,61	2,57	2,53	2,51	2,47	2,46	2,43	2,42	2,41
	9,65	**7,21**	**6,22**	**5,67**	**5,32**	**5,07**	**4,89**	**4,74**	**4,63**	**4,54**	**4,46**	**4,40**	**4,29**	**4,21**	**4,10**	**4,02**	**3,94**	**3,86**	**3,81**	**3,74**	**3,71**	**3,66**	**3,62**	**3,60**
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,72	2,69	2,64	2,60	2,54	2,51	2,47	2,43	2,40	2,37	2,35	2,32	2,31	2,30
	9,33	**6,93**	**5,95**	**5,41**	**5,06**	**4,82**	**4,64**	**4,50**	**4,39**	**4,30**	**4,22**	**4,16**	**4,05**	**3,97**	**3,86**	**3,78**	**3,70**	**3,62**	**3,57**	**3,50**	**3,47**	**3,41**	**3,38**	**3,36**
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,63	2,60	2,55	2,51	2,46	2,42	2,38	2,34	2,31	2,28	2,26	2,23	2,22	2,21
	9,07	**6,70**	**5,74**	**5,21**	**4,86**	**4,62**	**4,44**	**4,30**	**4,19**	**4,10**	**4,02**	**3,96**	**3,86**	**3,78**	**3,66**	**3,59**	**3,51**	**3,43**	**3,38**	**3,31**	**3,27**	**3,22**	**3,19**	**3,17**
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,57	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**	**5,04**	**4,69**	**4,46**	**4,28**	**4,14**	**4,03**	**3,94**	**3,86**	**3,80**	**3,70**	**3,62**	**3,51**	**3,43**	**3,35**	**3,27**	**3,22**	**3,15**	**3,11**	**3,06**	**3,03**	**3,01**
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,51	2,48	2,42	2,38	2,33	2,29	2,25	2,20	2,18	2,14	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**	**3,49**	**3,37**	**3,29**	**3,21**	**3,13**	**3,08**	**3,01**	**2,98**	**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	2,46	2,42	2,37	2,33	2,28	2,24	2,19	2,15	2,12	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**	**3,62**	**3,55**	**3,45**	**3,37**	**3,26**	**3,18**	**3,10**	**3,02**	**2,97**	**2,90**	**2,86**	**2,81**	**2,78**	**2,75**

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

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

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:

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

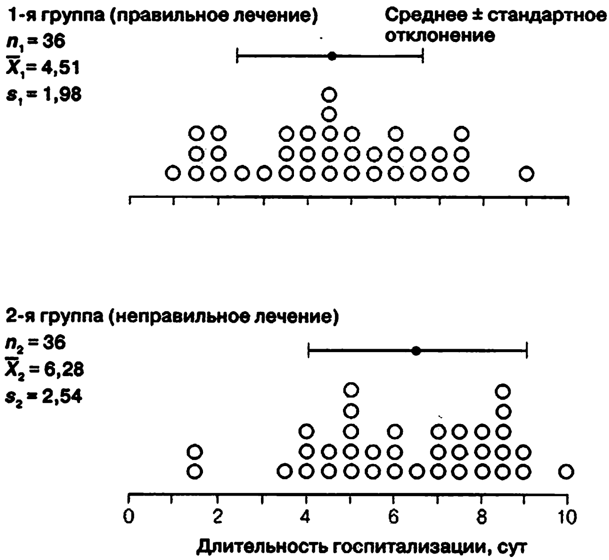

Рис. 3.7. Длительность госпитализации при правильном (1-я руппа) и неправильном лечении. Каждый больгой обозначен кружком; положение кружка соответствует сроку госпитализации. Средняя длительность госпитализации в первой группе меньше, чем во второй. Можно ли отнести это различие на счет случайности?

Теперь вычислим межгрупповую дисперсию.

Среднее двух выборочных средних равно

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

следовательно, стандартное отклонение равно

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

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

и наконец межгрупповая дисперсия равна

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

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 внутри групповое 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 expenence. *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 ни будь 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:

$$\text{АД}_{\text{средн}} = \frac{\text{АД}_{\text{С}} - \text{АД}_{\text{Д}}}{3} + \text{АД}_{\text{Д}},$$

где $\text{АД}_{\text{средн}}$ — среднее артериальное давление, $\text{АД}_{\text{Д}}$ — диастолическое артериальное давление, $\text{АД}_{\text{С}}$ — 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 $\text{АД}_{\text{средн}}$ 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:

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

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

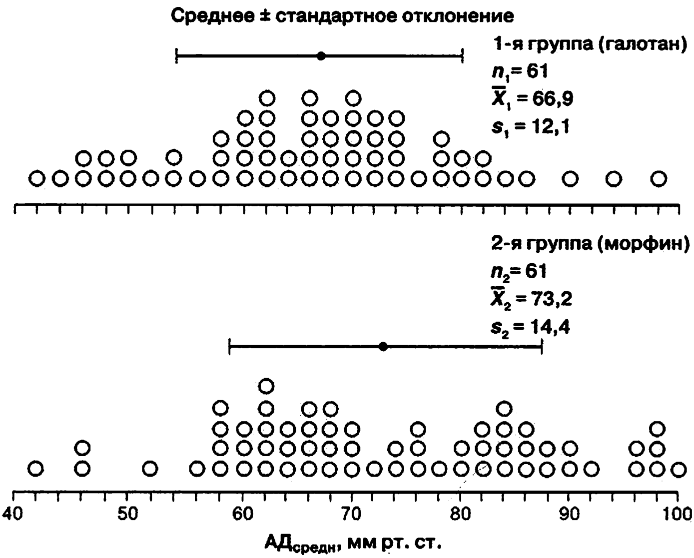

Рис. 3.8. Минимальный уровень АД$_{\text{средн}}$ между началом анестезии и началом операции при галотановой (1-я группа) и морфиновой (2-я группа) анестезии. Можно ли на основании этих данных отвергнуть нулевую гипотезу об отсутствии связи между выбором анестетика и артериальным давлением?

оказывают одинаковое действие на артериальное давление. В этом случае две группы бальных, 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

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

Стандартное отклонение выборочных средних:

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

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

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

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

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 менструации 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 заболевании. 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

---

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

исследование целью, которого было установить, есть та связь между занятиями спортом и частотой менструации. 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 менструации в год. В контрольной группе среднее число менструации 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:

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

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

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

получаем следующую оценку стандартной ошибки:

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

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

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

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

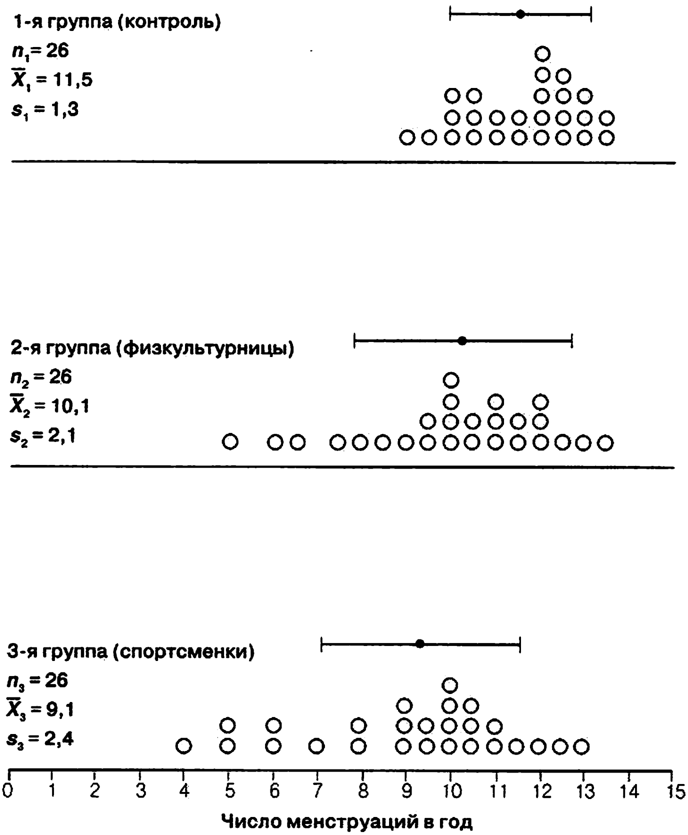

Рис. 3.9. Число менструации в год у женщин которые не занимались ни физкультурой, ни спортом (1-я группа), физкультурниц (2-я группа) и спортсменок (3-я группа). Среднее число менструаций различно. Можно ли отнести эти различия за счет случайности.

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 значимы — 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. MacDonaid. Influence of reproduction 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 средины 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, joggles, 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 (тетрагидроканнабинота 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 внутривенном 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	573	46,4	43,9	65,2
Стандартное отклонение	1,4,3	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 миокарда 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 новорожденных регулируется иначе чем у взрослых. Исследуя эту регуляцию X. 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 различии 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 случаи 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. Phamacol. Ther.,* 15:97—107, 1974.

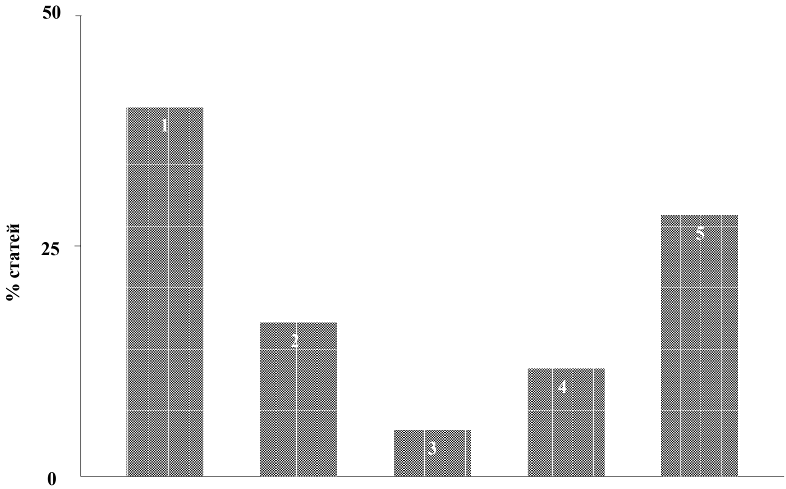

Рис. 4.1. Использование статистических методов в медицинских исследованиях. Рассмотрено 142 статьи опубликованные в 56-м томе журнала Circulation (кроме обзоров, описаний случаев и работ по рентгенологии и патоморфологии). В 39% работ статистические методы не использовались вовсе, в 34% прааильно использовали 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 med call teralure. *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-

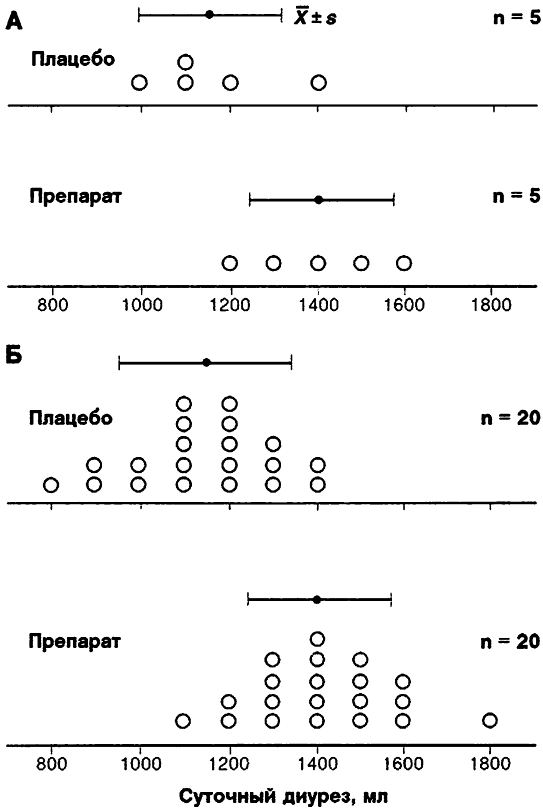

Рис. 4.2. Результаты испытаний предполагаемого диуретика. **А.** Диурез после приема плацебо и препарата. В обеих группах по 5 человек. **Б.** Теперь в обеих группах по 20 человек. Средние и стандартные отклонения остались прежними, однако доверие к результату повысилось.

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.

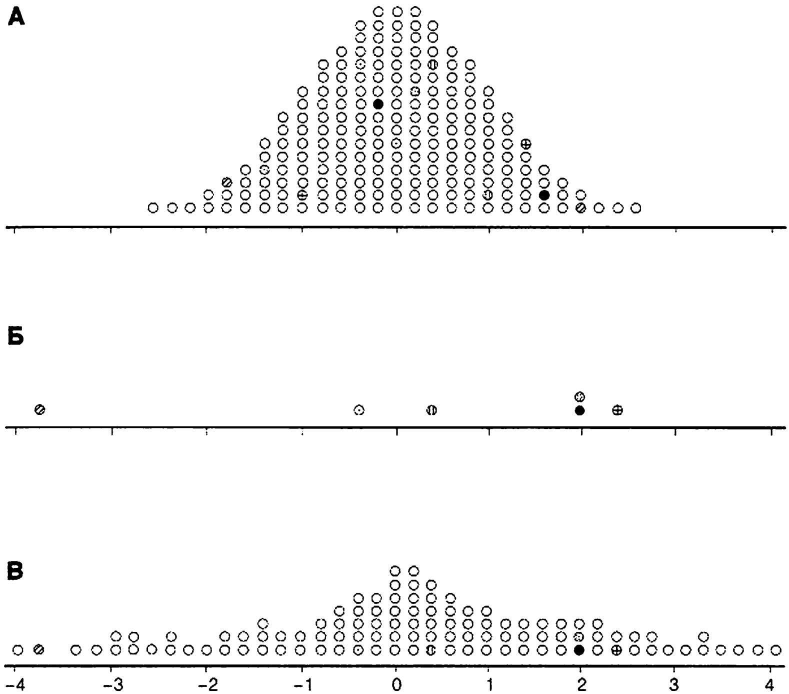

Рис. 4.3 А. Из этой совокупности мы будем извлекать пары и вычислять разности. **Б.** Разности первых 6 пар. **В.** Разности еще ста пар. Разброс разностей больше, чем разброс самих значений.

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 каждой 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 значении равна сумме дисперсии 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

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

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

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

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

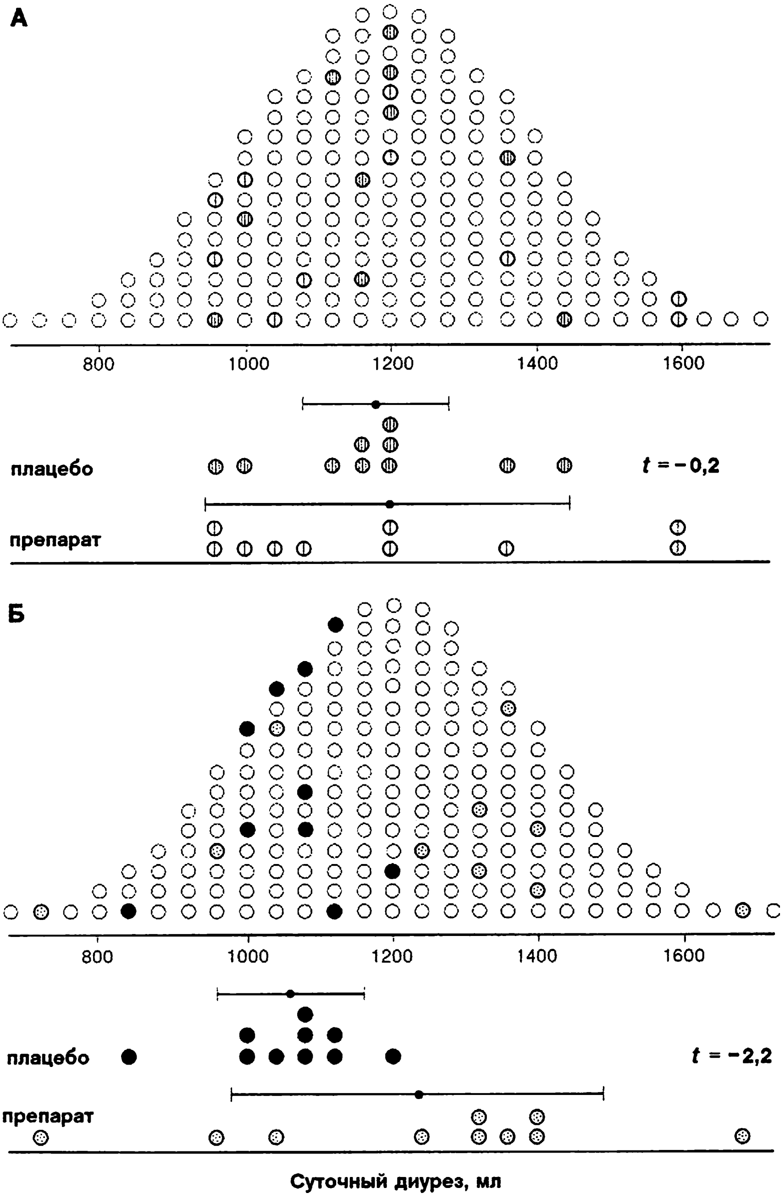
А
800
1000
1200
1400
1600
плацебо
t = −0,2
препарат
Б
800
1000
1200
1400
1600
плацебо
t = −2,2
препарат
Суточный диурез, мл

Рис. 4.4.

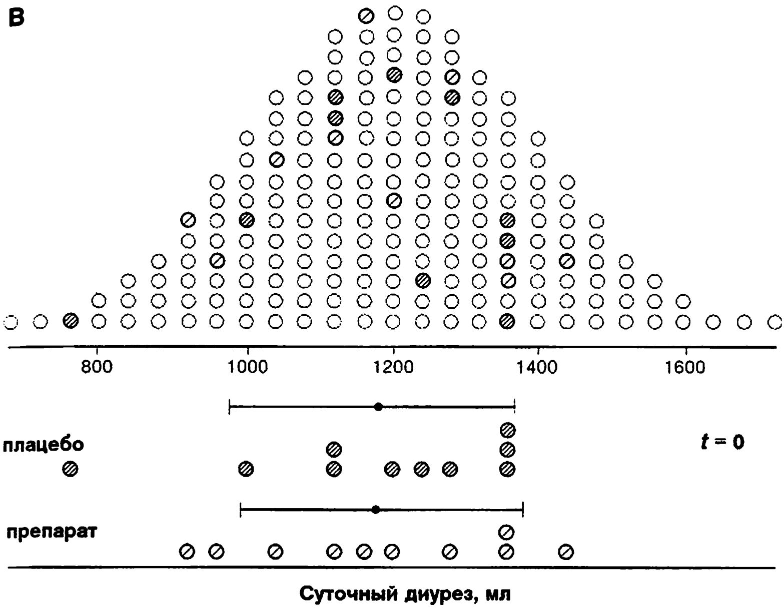

Рис. 4.4. Испытания предполагаемого диуретика. **А.** В действительности препарат не обладает диуретическим действием, поэтому обе группы — просто две случайные выборки из совокупности, показанной в верхней части рисунка. Члены совокупности, которым посчастливилось принять участие в исследовании, помечены 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 бы счел 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

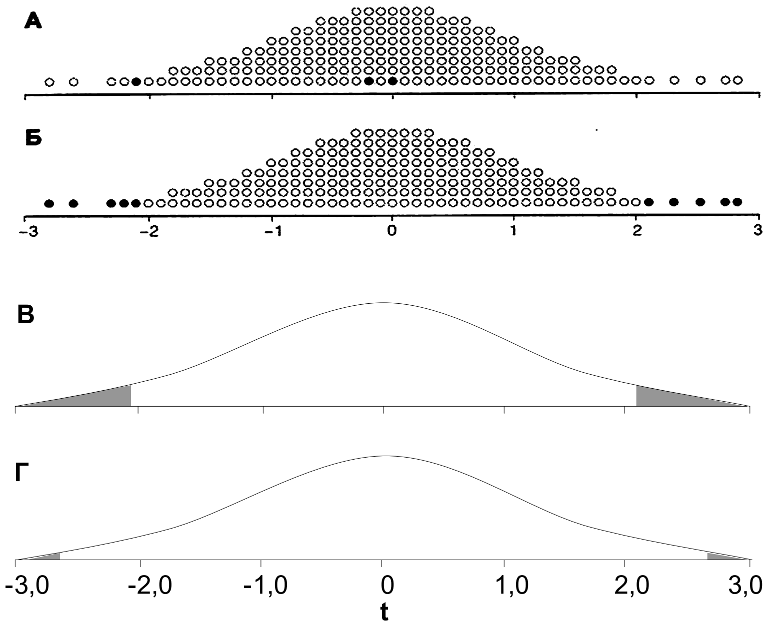

Рис. 4.5. А. Из совокупности, показанной на рис. 4.4, извлекли 200 пар случайных выборок по 10 членов в каждой, для каждой пары рассчитали значение *t* и нанесли его на график. Значения для *t* трех пар выборок с рис. 4.4 помечены черным. Большая часть значений сгруппирована вокруг нуля, однако некоторые 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. Чтоб снизить вероятность ошибочного заключения, мы можем выбрать более высокое критическое значение. Например, чтобы площадь заштрихованной области составляла 1% от обшей 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 различии 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 различии 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	0,5	0,2	0,1	0,05	0,02	0,01	0,005	0,002	0,001
1	1,000	3,078	6,314	12,706	31,821	63,656	127,321	318,289	636,578
2	0,816	1,886	2,920	4,303	6,965	9,925	14,089	22,328	31,600
3	0,765	1,638	2,353	3,182	4,541	5,841	7,453	10,214	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	5,894	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	3,689
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	3,660
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	0,5	0,2	0,1	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	2,847	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	2,838	3,131	3,340
∞	0,675	1,282	1,645	1,960	2,327	2,576	2,808	3,091	3,291

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 отклонении 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

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

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

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

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 $s^2 = \frac{1}{2}\left(1,98^2 + 2,54^2\right) = 5,18.$ 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 галотановои 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 и 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.

$$s_{\text{вну}}^2 = \frac{1}{2}\left(s_1^2 + s_2^2\right).$$

Вторая межгрупповая оценка вычисляется по выборочным средним:

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

следовательно,

$$s_{\bar{X}}^2 = \left(\bar{X}_1 - \bar{X}\right)^2 + \left(\bar{X}_2 - \bar{X}\right)^2,$$

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

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

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

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

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

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

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

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

Таким образом, межгрупповая оценка дисперсии

$$s_{\text{меж}}^2 = ns_{\bar{X}}^2 = \frac{n}{2}\left(\bar{X}_1 - \bar{X}_2\right)^2.$$

F есть отношение межгрупповой оценки к внутригрупповой и равно

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

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

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 критерии 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 различии 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 сравнении сообщают что «$P < 0,05$». Это означает, что вероятность ошибочного заключения о существовании различии 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 также 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 сравнении 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 сравнении 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.

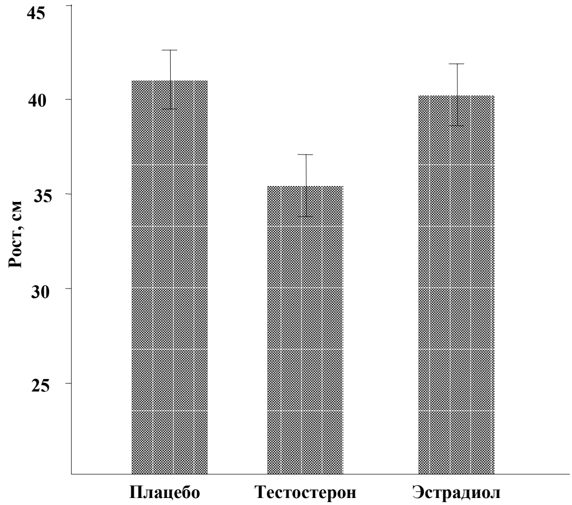

Рис. 4.6. Влияние гормонов человека на рост марсиан. Именно в таком виде результаты исследования увидели бы свет в каком-нибудь медицинском журнале. Высота столбиков соответствует средним, вертикальная черта на верхушке у каждого столбика соответствует интервалу плюс-минус одна стандартная ошибка среднего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

- Если схема эксперимента предполагает большее число групп, воспользуйтесь дисперсионным анализом.
- Если критерии 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 анализ 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 критерии 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 множественного сравнения. Этот подход изложен в B. W. Broun, Jr., M. Hollander. Statistics: a biomedical introduction. Wiley, NewYork, 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 критерии 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

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

где s^2 – объединенная оценка дисперсии совокупности.

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

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

Если объемы выборок одинаковы то

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

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:

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

при сравнении контрольной группы и группы спортсменок:

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

и при сравнении группы физкультурниц и группы спортсменок:

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

Мы провели 3 сравнения, поэтому уровень значимости в каждом 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 менструации ниже, чем в контрольной группе при этом у спортсменок и физкультурниц она не отличается.

КРИТЕРИЙ НЬЮМЕНА-КЕЙЛСА**

При большом числе сравнении поправка Бонферрони делает критерии 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 (в нашем примере это 0,01) либо приблизительно рассчитать нужное критическое значение по соседним. Если нужное нам значение $\alpha_{\text{н}}$ находится между α_1 и α_2, которым соответствуют критические значения t_1 и t_2 то

$$t_{\text{н}} = t_1 + (t_2 - t_1)\frac{(\alpha_{\text{н}} - \alpha_1)}{(\alpha_2 - \alpha_1)},$$

где $t_{\text{н}}$ — критическое значение для уровня значимости $a_{\text{н}}$.

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

где $\bar{X}_A$ и $\bar{X}_B$ — 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-ое с 1-ым, затем m-ое со 2-ым, 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 менструации 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 $\bar{X}_1$, $\bar{X}_2$, $\bar{X}_3$ соответственно. Оценка внутригрупповой дисперсии $s^2_{\text{вну}} = 3{,}95$, число степе-

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

ν	Интервал сравнения l								
	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	3,82	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	5,15
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 n = 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	164,3	185,6	202,2	215,8	227,2	237	245,6
2	14,04	19,02	22,29	24,72	26,63	28,2	29,53	30,68	31,69
3	8,261	10,62	12,17	13,33	14,24	15	15,64	16,2	16,69
4	6,512	8,12	9,173	9,958	10,58	11,1	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	6,96	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	5,27	5,769	6,136	6,428	6,669	6,875	7,055	7,213
11	4,392	5,146	5,621	5,97	6,247	6,476	6,672	6,842	6,992
12	4,32	5,046	5,502	5,836	6,101	6,321	6,507	6,67	6,814
13	4,26	4,964	5,404	5,727	5,981	6,192	6,372	6,528	6,667
14	4,21	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	5,14	5,43	5,659	5,847	6,007	6,147	6,27
18	4,071	4,703	5,094	5,379	5,603	5,788	5,944	6,081	6,201
19	4,046	4,67	5,054	5,334	5,554	5,735	5,889	6,022	6,141
20	4,024	4,639	5,018	5,294	5,51	5,688	5,839	5,97	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	4,2	4,497	4,709	4,872	5,005	5,118	5,214	5,299
∞	3,643	4,12	4,403	4,603	4,757	4,882	4,987	5,078	5,157

H. I. 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 поучилось у нас.

Следовательно, различие статистически значимо.

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

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

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.

КРИТЕРИИ ТЬЮКИ

Критерии 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 критерии Тьюки к только что рассмотренной задаче о влиянии бега на частоту менструации 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 критерии 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 $\alpha' = 0{,}05$

ν	Интервал сравнения l													
	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 l 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.

Критерии Даннета

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

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

Число сравнении 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.

Критерии 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 критерии Ньюмена-Кейлса обнаруживает те различия, которые упускает критерии 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. Чаше 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 оценивает *P*. Возможна и противоположная ошибка — принять неверную нулевую гипотезу то есть не найти действительно существующее различие. Это гак 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 артерии 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?