



Белоголовый А.Е.

Статистический анализ

в медицине и биологии

Москва 2007

задачи,
терминология,
логика,
компьютерные
методы

УДК 57.087.1
ББК 22.172
П375
ISBN 5-7901-0022-8

П375 Платонов А.Е. Статистический анализ в медицине и биологии: задачи, терминология, логика, компьютерные методы. - М.: Издательство РАМН, 2000. - 52 с.

Описаны параметрические и непараметрические методы статистического анализа и их использование в медицинских и биологических исследованиях.

Рассмотрены этап планирования эксперимента (формулирование статистической гипотезы), выбор адекватных статистических процедур, интерпретация полученных результатов и способы представления их для публикации. Теоретические положения дополнены примерами и иллюстративным материалом.

Книга предназначена для научных работников — биологов и медиков, а также может использоваться в учебном процессе.

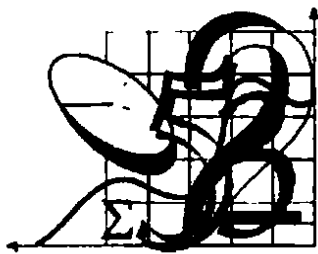
ББК 22.172

© Издательство РАМН, 2000



Содержание

Введение	5
Генеральная совокупность	7
Характеристика распределения	11
Статистические гипотезы и их проверка	17
Дискретные качественные переменные, описываемые номинальными шкалами	18
Переменные, описываемые порядковыми (ранжируемыми) шкалами	21
Анализ связи нескольких переменных	26
Анализ наступления событий («дожития» или «выживаемости») [survival analysis]	32
Статистические понятия, связанные с таблицами 2x2	34
Опасность множественных сравнений	37
Заключение	42
Краткий список литературы	50



Введение

Биомедицинская статистика — это инструмент для анализа экспериментальных данных и клинических наблюдений, а также язык, с помощью которого исследователь сообщает читателю полученные им результаты. Неадекватный инструментарий или нечеткое использование языка автором неизбежно делает сообщение неинформативным.

Цель данной книги — дать краткую сводку наиболее часто встречающихся и достаточно простых задач и методов статистического анализа, а также предостеречь читателя от распространенных ошибок. Более отдаленная цель — создание информационной среды, в которой общепонятный и общепринятый язык статистического анализа станет одним из элементов «доказательной» (evidence-based) медицины и биологии.

Первый и решительный совет — проводите статистическую обработку данных только на персональном компьютере (ПК), качество анализа повысится в десятки раз. Забудьте о вычислениях в столбик и расчерчивании таблиц. Выберите специализированный программный пакет для статистического анализа. Учтите, что вы будете работать с ним (или его новыми версиями) ближайшие 5 лет. Переходить от программы к программе — терять время. Программа должна быть достаточно мощной и уметь многое, а не только сравнивать средние значения и считать коэффициенты корреляции. Для обладателей современ-

ных ПК можно порекомендовать программы SPSS или «Statistica for Windows»; на менее мощных ПК под DOS будут хорошо работать различные, в том числе ранние, версии «Statgraphics» [1, 2]. В принципе можно освоить начала статистического анализа по описаниям данных программ, представленным разработчиками, или разделу «Справка/Help» непосредственно в программе. Список публикаций, которые читатель может использовать, в том числе с разбором большого числа практических задач, прилагается в конце книги [1-6]. Однако следует иметь в виду, что чтение некоторых из них, например [1, 6], требует определенных математических познаний и, главное, математической культуры мышления. Автор данной книги пытался сделать ее понятной для читателей с базовым медико-биологическим образованием. К сожалению, многие популярные руководства по статистике и биометрике безнадежно устарели: в них написано все правильно, но не о том. Если в книге на пяти страницах приведены десятки формул для расчета среднего или критерия χ^2 — отложите ее. Вам фактически не надо знать ни одной формулы, но следует четко понимать другое:

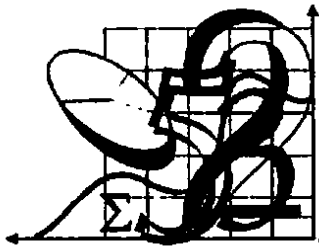
?

Как формулируется тот вопрос, на который вы хотите ответить с помощью статистического анализа?

Какие статистические процедуры наиболее адекватны для ответа на данный вопрос?

Как следует интерпретировать результаты?

К сожалению, в статистике русскоязычная терминология не достаточно устоялась. Ниже автор будет зачастую давать несколько терминов (выделенных курсивом) для описания одного и того же статистического объекта, а также приводить их англоязычный аналог. Последнее, во-первых, поможет в работе с нерусифицированными компьютерными программами, а во-вторых, позволит опереться на более стандартизованную англоязычную лексику.



Генеральная совокупность

Можно выделить два основных типа статистических задач:

- 1) кратко описать большой массив данных;
- 2) проверить некую статистическую гипотезу.

В идеале массив данных должен описывать всю *генеральную совокупность* описываемых явлений (например, продолжительность жизни всех людей на Земле). На практике мы чаще всего имеем дело с некоторой *выборкой* из генеральной совокупности. Любая выборка описывает генеральную совокупность с неизбежными погрешностями и искажениями, природу и степень выраженности которых необходимо понимать.

Каждый *элемент* [case] генеральной совокупности может характеризоваться несколькими признаками (человек — ростом, массой тела и т.п.). Множество *значений* [values] конкретного признака генеральной совокупности есть, на математическом языке, *переменная* [variable] (в статистике иногда называемая также *вариационным рядом*).

Переменные могут быть *непрерывными* [continuous] (температура больного) или *дискретными* [discrete] (число инъекций); частный случай дискретных переменных — *дихотомическая* [dichotomous] переменная, принимающая одно из двух значений (выжил или умер). Переменные описываются с помощью *шкал* [scale]. Шкалы бывают *номинальные/качественные* [nominal/categorical], *порядковые*

или *ранжируемые* [ordinal] и *интервальные* [interval], иногда называемые *количественными*. Номинальной шкалой описываются переменные, значения которых, в рамках конкретной задачи, не могут быть ранжированы (например, районы города или этническое происхождение испытуемого). Однако не исключено, что в другой задаче «районы города» могут быть ранжированы (расставлены в порядке возрастания), например, по степени загрязненности воздуха, и описываться в этом случае порядковой шкалой. Особенность порядковых шкал — отсутствие количественной меры расстояния между величинами на шкале (можно допустить, что тяжелое течение болезни «хуже», чем среднетяжелое, а очень тяжелое течение — «еще хуже»; однако нельзя сказать, во сколько раз хуже). Для интервальных шкал подобная количественная мера однозначно определена (например, температура). Важно четко понимать, каким именно типом переменных и типом шкалы описывается изучаемое явление. Для различных переменных и шкал применяются разные методы статистического анализа!

!

Генеральные совокупности характеризуются *распределением* значений исследуемой переменной (признака), то есть абсолютной или относительной *частотой встречаемости* [frequency distribution] ее конкретных значений. Соответственно, выборка характеризуется *выборочным (эмпирическим) распределением* значений признака. В идеале свойства выборки должны максимально близко и без искажений отражать свойства генеральной совокупности, а эмпирическое распределение приближаться к распределению в генеральной совокупности. Однако как добиться этого практически, если свойства генеральной совокупности нам априори неизвестны? Чаще всего выборку организуют случайным образом, а исследуемую переменную величину рассматривают, соответственно, как *случайную* величину. Увы, полностью случайной, в математическом смысле, выборки в биологии и медицине добиться невозможно, в частности по этическим или техническим причинам. Организация эксперимента или контролируемого наблюдения — отдельная область науки,

частично пересекающаяся со статистикой, частично — имеющая собственный предмет и задачи и заслуживающая отдельного обсуждения [4].

Итак, если вам известно большое число (скажем, больше 30) индивидуальных значений исследуемой переменной (по интервальной шкале), постройте с помощью ПК ее распределение. Вероятнее всего, результат будет близок к одному из четырех распределений: *равномерному, колоколообразному, асимметричному или полимодальному* (рис. 1). Даже такой упрощенный подход даст некоторую информацию: полимодальное распределение может быть обусловлено действием нескольких скрытых факторов, которые потом полезно будет попытаться выяс-

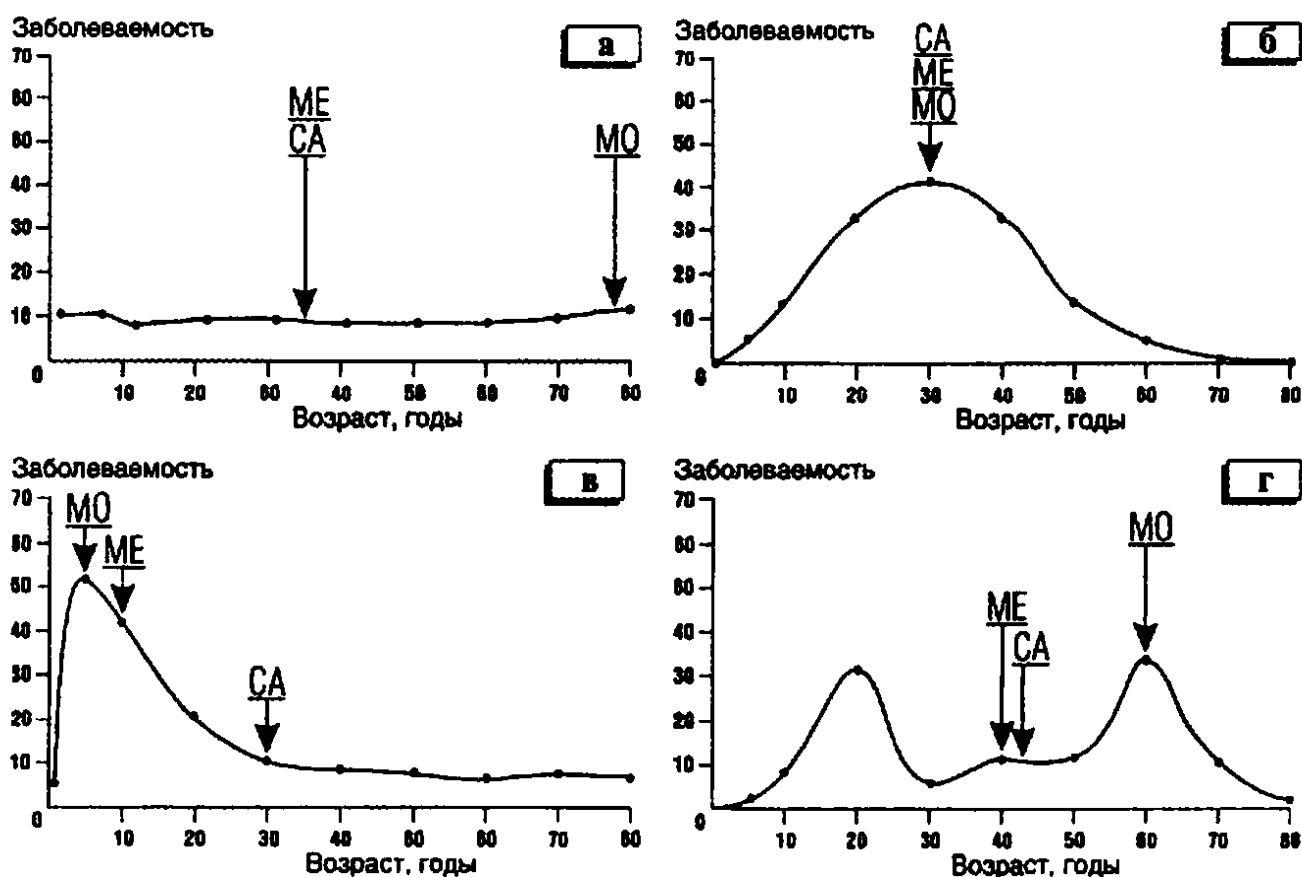


Рис. 1. Пример распределений различной формы и их статистические характеристики.

Зависимость заболеваемости от возраста при четырех условных патологиях.

а — равномерное, б — колоколообразное, в — асимметричное, г — полимодальное распределения.

Для каждого распределения указаны положение среднего арифметического (СА), медианы (МЕ) и моды (МО).

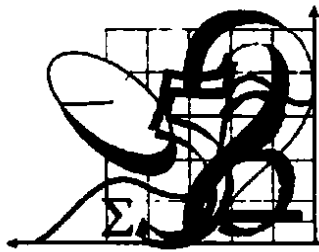
нить, равномерное распределение указывает на малое влияние данной переменной на исследуемый процесс.

Важно отметить, что так называемые *параметрические* методы статистического анализа, например *t* критерий Стьюдента, которым учат в институтах и описанию которых посвящена большая часть популярных книг по биометрике и медицинской статистике, применимы **только** для анализа переменных, характеризующихся *нормальным (гауссовым)* или «квазинормальным» распределением колоколообразного вида. **Для распределений, не являющихся нормальными, параметрические методы неприменимы!**

!

Некритическое использование параметрических методов является распространеннейшей ошибкой, встречающейся в существенной части биомедицинских публикаций. В принципе, в статистике существуют приемы проверки соответствия эмпирического распределения нормальному (или теоретическим распределениям, описываемым иными формулами, — например, пуассоновскому).

Практически в большинстве случаев разумно полностью отказаться от параметрических методов и пользоваться *непараметрическими (ранговыми)* методами статистического анализа, применимыми для **любых распределений**. С использованием компьютерного программного обеспечения их применение не вызывает никаких затруднений.



Характеристика распределения

Все, что мы на настоящий момент выяснили о свойствах генеральной совокупности, содержится в эмпирическом распределении и, следовательно, данные о нем и надо приводить в публикациях о результатах исключительной важности или новизны. Однако обычно соображения краткости и наглядности диктуют необходимость применения сжатых способов описания распределения. В первую очередь автора и читателя интересует характеристика *центральной тенденции* распределения, т.е. его «наиболее типичного» значения. Чаще всего используют три характеристики: среднее арифметическое (M), медиану и моду.

Медиана — это значение, делящее распределение пополам, так что справа и слева от него находится равное число значений. Если ранжировать ряд значений (вариационный ряд) по возрастанию, то в случае нечетного числа значений в ряду медиана — это центральное значение, а в случае четного числа значений в ряду медиана — это среднее арифметическое двух центральных значений.

Мода — значение с наивысшей частотой встречаемости.

Какую из характеристик «правильнее» применять? Это определяется как видом эмпирического распределения, так и аналитической задачей, которая стоит перед исследователем.

Для симметричного «почти нормального» распределения среднее арифметическое, медиана и мода совпадают (рис. 1, б). Обычно указывают среднее арифметическое.



Внимание! В биомедицинских публикациях вместе со средним арифметическим очень часто указывают так называемую *стандартную ошибку среднего* [standard error of mean], обозначаемую обычно m и вычисляемую по формуле s/\sqrt{n} , где s — выборочное стандартное отклонение, а n — число измерений (или наблюдений). Этого делать не надо. Стандартная ошибка среднего стремится к нулю при увеличении числа измерений, что радует автора публикации «точностью» приводимых данных $M \pm m$. И зря. Использование стандартной ошибки среднего пришло в статистику из физики, где при измерении параметров одинаковых объектов (например, заряда электрона) вариабельность получаемых значений определяется только случайными погрешностями. Увеличивая n , мы все точнее знаем реальное значение измеряемого параметра. В биомедицине тоже можно обследовать 10 000 пациентов и сообщить (пример условный), что среднее количество CD4⁺-лимфоцитов у больных СПИДом составляет $41.384 \pm 0.001\%$ от нормального. Кому нужна такая точность, если реальное количество CD4⁺-лимфоцитов у больных СПИДом может варьировать от 0 до 100%? (Автор не утверждает, что не нужно увеличивать число измерений. При увеличении числа измерений улучшается качество проверки статистических гипотез). Что нужно выяснить исследователю (и знать читателю), так это то, в какой диапазон укладывается большинство значений определяемого параметра, т.е. охарактеризовать ширину распределения.

Ширину нормального распределения принято характеризовать *стандартным отклонением* (синоним — *среднее квадратичное отклонение*) [standard deviation], точнее, выборочным стандартным отклонением, поскольку обычно мы работаем с выборочными распределениями, и обозначать его s .

$$s = \sqrt{\frac{\sum_{i=1}^n (\bar{x} - x_i)^2}{n-1}},$$

где n — число случаев, \bar{x} — среднее арифметическое, x_i — значение в i -ом случае.

В отличие от m , s не уменьшается при увеличении n , но определяется более точно. Результаты представляют как $M \pm s$; в диапазон $M \pm s$ укладываются около 70% значений нормального распределения.

К сожалению, стандартное отклонение плохо приспособлено для описания распределений, отличных от нормального, в особенности асимметричных распределений. Для возрастной зависимости, приведенной на рис. 1, \bar{x} , средний возраст заболевания $M \pm s$ пришлось бы записать как 5 ± 15 лет, в то время как очевидно, что отрицательных возрастов не существует.

Из непараметрических характеристик ширины распределения простейшей является его *размах* или *амплитуда* [range] — разница максимального и минимального значений, которые принимает переменная. Однако размах не описывает «кучности» или «крутизны» распределения, т.е. степени группировки значений вокруг центрального.

Наиболее адекватная непараметрическая характеристика ширины — это так называемые *квантили* [quantile] или *перцентили* [percentile]. Разобьем весь диапазон значений, которые принимает переменная, на десять интервалов так, чтобы в первый диапазон попали наименьшие значения (10% от общего количества), в следующий диапазон 10% значений побольше и так далее. В последний диапазон попадут 10% наибольших значений. Границы между интервалами называются, соответственно, 10%, 20% ... 90% квантилями. Вне интервала от 10% до 90% квантиля оказывается 20% наиболее отклоняющихся значений.

Часто весь диапазон значений разбивают только на четыре интервала, получая, соответственно, 25%, 50% и 75% квантили; 25% и 75% квантили называют также *нижним* [low quartile] и *верхним квантилями* [high quartile]. 50% квантиль — это, как нетрудно заметить, медиана. Внутри *интерквартильного интервала* лежат 50% наиболее типичных (близких к центральному) значений.

Итоги измерений (наблюдений) могут быть представлены в виде тройки цифр нижний квантиль—медиана—

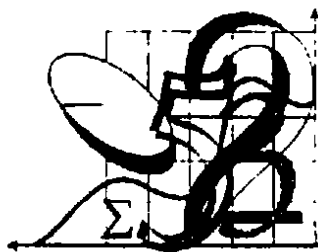
верхний квартиль, дающих представление о центральной тенденции, ширине и асимметрии распределения результатов. К сожалению, этот простой и безотказный подход пока не получил достаточного распространения в русскоязычной литературе.

Внимание! Оценка ширины распределения в виде $M \pm s$ или нижний квартиль—медиана—верхний квартиль позволяет выбрать правильный формат представления данных. Если, как это обычно бывает в биомедицине, s составляет более 10% от M , или интерквартильный интервал более 10% от медианы, величина s (или интерквартильный интервал) записывается с точностью не более 2 значащих цифр, а величина M (или медиана) — с точностью до значащей цифры того же порядка. Не следует писать: концентрация вещества в растворе X равна 333333 ± 33333 мкг/л, а в растворе Y — 222222 ± 22222 мкг/л (даже если прибор позволяет мерить концентрацию с точностью до 1 мкг/л). Во-первых, это затрудняет зрительное сравнение результатов. Во-вторых, это очевидный признак того, что автор публикации плохо знаком с основами статистического анализа, вызывающий естественную настороженность при оценке качества всех результатов. Следует, округляя, писать: концентрация в растворе X равна 333000 ± 33000 мкг/л, а в растворе Y — 222000 ± 22000 мкг/л.

Оценка доверительного интервала. Наиболее типичный случай оценки доверительного интервала — представление результатов измерения некоторой количественной величины в виде «среднее арифметическое $\pm 95\%$ доверительный интервал» [95% confidence interval]. На самом деле, не только для среднего, но и для любой, точно неизвестной случайной величины можно попытаться оценить область значений, в которой эта величина находится с вероятностью, не меньшей заданной (например 95%). Проблема заключается в том, что вычисление доверительного интервала требует точного знания или верного допущения о характере распределения данной величины и, следовательно, не всегда возможно. Один из немногих слу-



чаев в медико-биологической статистике, когда вычисление доверительного интервала не только допустимо, но и информативно — это расчет относительного риска.



Статистические гипотезы и их проверка

В статистике, как и в жизни, важные утверждения редко удается доказать окончательно и неоспоримо. Можно только выдвинуть утверждение, справедливое с некоторой степенью достоверности. Статистическая гипотеза — это формально строго сформулированное предположение об истинном распределении, которое мы хотим проверить по имеющимся эмпирическим данным. Гипотеза, которую исследователь предполагает отклонить, обычно называется *нулевой гипотезой*. (Так, если ожидается различие результатов в контрольной и опытной группах, то формулировка нулевой гипотезы звучит так: значения переменной в контрольной и опытной группах неотличимы, т.е. являются выборками из одной и той же генеральной совокупности). *Альтернативная гипотеза* логически отрицает нулевую.

С помощью адекватного статистического критерия оценим вероятность (обозначаемую обычно p) случайно получить наши фактические результаты в предположении, что нулевая гипотеза верна. Если эта вероятность достаточно низка, ниже некоторого условного порогового значения, то нулевая гипотеза отклоняется. Пороговое значение для принятия отклоняющего решения называется *уровнем значимости* [significance level] и обозначается α . В биомедицинской статистике обычно выбирают уровень значимости α , равный 0.05 или 0.01. Чем меньше выбрано значение α , тем ниже вероятность

ошибки первого рода, то есть ошибочного отклонения верной нулевой гипотезы. (Однако не следует забывать, что при этом возрастает вероятность *ошибки второго рода*, т.е. ошибочного принятия ложной нулевой гипотезы.) При публикации полезно указывать реальное значение p (например, $p=0.0007$), т.е. указать минимальный уровень значимости, при котором можно отвергнуть гипотезу, а не просто писать $p<0.05$, поскольку точное значение p позволяет оценить степень отклонения результатов от нулевой гипотезы и степень их достоверности.

При сравнении двух выборок могут выдвигаться *направленные* и *ненаправленные* гипотезы. Ненаправленная альтернативная гипотеза предполагает, что значения переменной в первой выборке отличны от значений во второй (или отличны от некоторого фиксированного числа). Направленная альтернативная гипотеза предполагает, что значения переменной в первой выборке больше значений во второй (или фиксированного числа). Направленные и ненаправленные гипотезы проверяются, соответственно, с помощью *односторонних* [one-tail] и *двусторонних* [two-tails] критериев.

Дискретные качественные переменные, описываемые номинальными шкалами

В простейшем случае дихотомической переменной нулевая гипотеза о частоте встречаемости ее значений (со-

ПРИМЕР

Допустим, нулевая гипотеза предполагает обнаружить среди 100 заболевших 50 мужчин и 50 женщин ($P=0.5$). Наблюдаемая частота событий — 70 мужчин из 100 пациентов (испытаний). Биномиальный критерий покажет, на каком уровне значимости можно отвергнуть нулевую гипотезу.

бытий) формулируется так: реальная вероятность события A равна P (а вероятность альтернативного события $\text{не-}A$ равна, соответственно, $Q=1-P$). Для проверки необходимо знать наблюдаемые частоты событий A и $\text{не-}A$, обозначаемые как p и q , и число испытаний.

Проверка проводится с помощью *биномиального критерия* [binomial test].

С помощью *критерия серий* [runs] можно проверить, насколько случайна последовательность выпадения значений дихотомической переменной.

Если дискретная переменная может принимать несколько значений (многозначна), то задача состоит в сравнении наблюдаемых частот различных значений (событий) и их теоретически ожидаемых частот. Нулевая гипотеза заключается в том, что наблюдаемые и ожидаемые частоты совпадают. Проверка проводится с помощью критерий χ^2 (*хи-квадрат*) [chi-square]. Результатом статистической обработки, которую, напомним, рекомендуется проводить с помощью ПК, будет величина χ^2 (которая тем больше, чем больше различие наблюдаемых и ожидаемых частот) и уровень значимости, с которым можно отвергнуть нулевую гипотезу. Поскольку дихотомические переменные — это частный случай дискретных, для их анализа можно также применять критерий χ^2 .

Если каждый элемент генеральной совокупности описывается значениями двух (или более) номинальных или порядковых переменных (признаков), часто статистический интерес вызывает построение *таблицы сопряженности признаков* [contingency table]. Если первая переменная может принимать m различных значений, а вторая — n значений, таблица сопряженности признаков представляет собой матрицу размером $m \times n$, в клетки которой заносятся частоты встречающихся комбинаций признаков.

В наиболее простом и часто применяемом виде анализа таблиц сопряженности нулевой гипотезой является *независимость* переменных, т.е. отсутствие связи между

ПРИМЕР

В последовательности 1 1 1 1 1 0 0 0 0 нули и единицы группируются, в последовательности 1 0 1 0 1 0 1 0 — чередуются, обе последовательности неслучайны.

ПРИМЕР

Данные по некоторой группе можно занести в следующую таблицу:

Число лиц	Американцы	Итальянцы	Китайцы	Всего
Мужчины	15	5	25	45
Женщины	15	25	5	45
Всего	30	30	30	90

признаками. В вышеприведенной таблице это означает, что ожидаемое отношение мужчины:женщины среди американцев (итальянцев, китайцев) такое же, как в группе в целом 1:1=45:45. Гипотеза проверяется с помощью критерия χ^2 (Пирсона—Фишера), сравнивающего наблюдаемые и ожидаемые частоты и дающего уровень значимости, с которым можно отвергнуть нулевую гипотезу. Особенности критерия χ^2 требуют для адекватного применения наличия не менее 5 элементов в каждой клетке и общего числа элементов (случаев) не менее 30.

Если обе переменные дихотомические, получается матрица размерности 2×2 , для анализа которой может применяться критерий χ^2 (с поправкой Yates «на группировку») или критерий Фишера (точной вероятности). Таблицы 2×2 имеют исключительное значение, например, в эпидемиологии, где с их помощью проводятся поиск и оценка так называемых факторов риска [risk factors], вычисляется относительный риск [relative risk] и коэффициент асимметрии [odds ratio]. Таблицы 2×2 используются также для расчета специфичности [specificity] и чувствительности [sensitivity] лабораторных тестов и прогностических показателей. Эти важные показатели будут рассмотрены в особом разделе.

Стандартная, но важнейшая статистическая задача — сравнение значений переменной (или нескольких однотипных переменных) в нескольких группах (или подгруппах), выбранных из генеральной совокупности согласно некоему условию. Подобные выборки могут быть независимыми [independent samples] или связанными (сопряженными, парными) [related samples].

Простейший пример независимых выборок — значение некоторого показателя (например, артериального давления) в случайным образом сформированных контрольной и экспериментальной группах.

Простейший пример связанных выборок — значение показателя (например, артериального давления) у одних и тех же пациентов до и после некоторого воздействия (или воздействий). Сравнение связанных и несвязанных выборок производится с помощью разных критериев!

!

Оценка сходства/различия двух связанных выборок значений дихотомических переменных производится с помощью критерия Макнимара [McNemar]. Для сравнения нескольких выборок применяется Q критерий Кокрена [Cochran's Q]. Нулевая гипотеза в этих случаях — две (или несколько) выборок происходят из одной генеральной совокупности, т.е. изменение значений изучаемой дихотомической переменной до и после воздействия (воздействий) не имеет определенного направления. Альтернативная гипотеза — изменение значений изучаемой переменной до и после воздействия не случайно.

ПРИМЕР

Так, например, в следующей таблице нули чаще сменяются на единицу, чем наоборот. Критерий, как обычно, определяет уровень значимости, с которым можно отвергнуть нулевую гипотезу

Значение переменной до воздействия	0	0	0	0	0	1	1	1	1	1
Значение переменной после воздействия	0	1	1	1	1	0	1	1	1	1

Переменные, описываемые порядковыми (ранжируемыми) или интервальными шкалами

Сравнение независимых выборок. Для сравнения двух независимых выборок (значения переменной разбиваются на 2 подгруппы в соответствии со значением некоего дополнительного фактора) можно рекомендовать либо широко известный U критерий Манна—Уитни [Mann—Whitney U] либо критерий серий Вальда—Вольфовица [Wald—Wolfowitz runs].

Логика обоих критериев проста: значения обеих выборок смешиваются в один ряд, ранжируемый по возрастанию; если нулевая гипотеза верна, то значения, принадлежащие каждой выборке, случайным образом распределены вдоль этого ранжированного ряда (рис. 2). Для критерия Манна—Уитни в качестве нулевой принимается гипотеза о совпадении средних (медианных) значений двух выборок. Более строгий критерий Вальда—Вольфовица проверяет гипотезу о совпадении распределения значений в двух выборках.

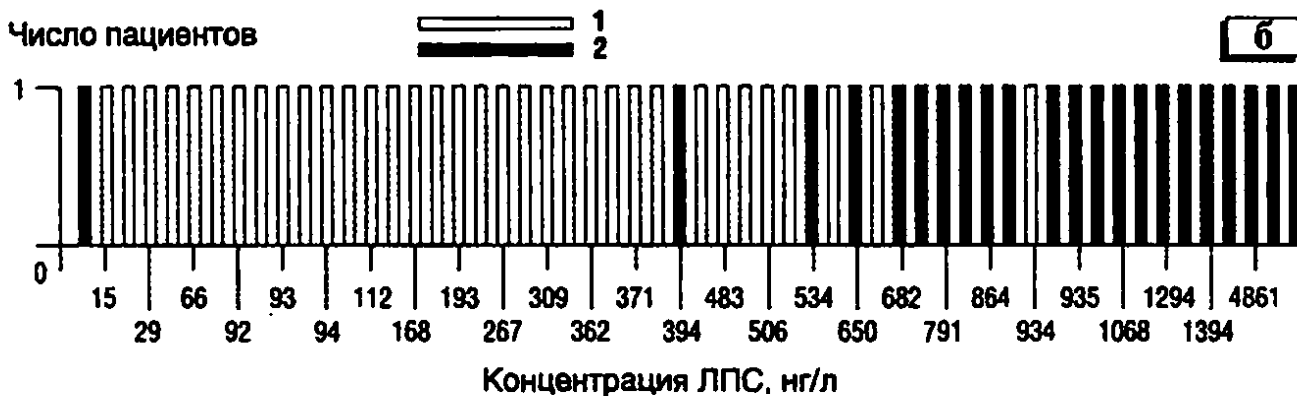
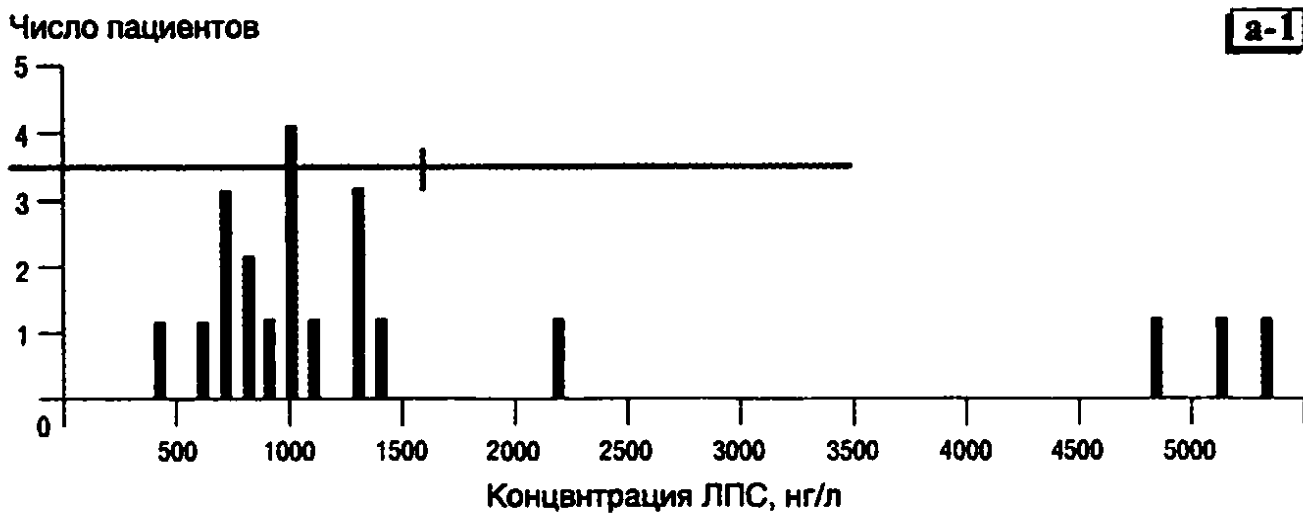


Рис. 2. Логика критериев сравнения двух независимых переменных.

Переменная 1 — концентрация менингококкового эндотоксина-липолисахарида (ЛПС) в крови больных генерализованными формами менингококковой инфекции без развития инфекционно-токсического шока. Концентрация измерена при поступлении в стационар с помощью лимулюс-теста.

Переменная 2 — то же, но в крови больных, у которых развился инфекционно-токсический шок. В обе группы включены только пациенты с благоприятным исходом заболевания.
а — среднее арифметическое значение (вертикальная засечка) ± стандартное отклонение (горизонтальная линия).

Обращает на себя внимание очевидное неравенство стандартных отклонений переменных 1 и 2, а также отличие распределений от нормального (асимметричность, наличие соответствующих высоким концентрациям ЛПС «выскакивающих» значений, которые не являются ошибкой измерения и, следовательно, не могут быть исключены из анализа).

Применение t критерия Стьюдента, основанного на сравнении средних арифметических и стандартных отклонений переменных и предполагающего равенство стандартных отклонений, приводит к выводу, что переменные 1 и 2 различны с уровнем значимости различия ниже 0.001. Этот вывод неправилен, поскольку специальный тест выявляет статистически значимое неравенство стандартных отклонений (или дисперсий). Вариант t критерия, учитывающий неравенство стандартных отклонений, показывает, что переменные 1 и 2 различны с уровнем значимости 0.004. Этот вывод ненадежен, поскольку распределения отличны от нормального.

б — в непараметрическом тесте Манна—Уитни обе группы измерений смешиваются и ранжируются по возрастанию значений. Значения переменной 2 группируются в правой части (средний ранг 43), а переменной 1 — в левой (средний ранг 21). Критерий Манна—Уитни указывает, что в данном случае переменные 1 и 2 достоверно различны с уровнем значимости ниже 0.001.

(Данные из материалов публикации [7].)

Соответственно, гипотеза о совпадении значений нескольких независимых выборок проверяется с помощью медианного критерия [median] или критерия Краскел—Уоллиса [Kruskal—Wallis one way ANOVA].

Внимание! В русскоязычной литературе критерий Манна—Уитни упорно называют критерием Вилкоксона—Манна—Уитни или даже просто критерием Вилкоксона, что создает путаницу с критерием Вилкоксона для связанных выборок.

Сравнение связанных выборок. Для анализа различия между связанными выборками, т.е. сравнения различных переменных одной природы при одинаковом числе случаев, применяются критерий Вилкоксона [Wilcoxon] (для 2 выборок) критерий Фридмана [Friedman two way ANOVA] или критерий Кендалла [Kendall's W] (для нескольких выборок). Критерий знаков [signs] является более популярным, но менее мощным аналогом критерия Вилкоксона, поскольку последний учитывает не только направление, но и степень выраженности (ранг) различий между связанными вы-



борками (чтобы подчеркнуть это обстоятельство, в русскоязычной литературе он называется критерием знаковых ранговых сумм Вилкоксона).

ПРИМЕР

Так, например, для таблицы

Значение переменной до воздействия	0	1	2	3	4	5	6	7	8	9
Значение переменной после воздействия	-1	0	1	2	4	10	20	30	40	50

критерий знаков не обнаружит достоверного эффекта (возрастание или убывание переменной встречается с одинаковой частотой), а критерий Вилкоксона отразит большую выраженность возрастания (см. также реальный пример на рис. 3).



анализа независимых выборок критериев, предназначенных для анализа связанных выборок, заведомо неправильно.

2. Все критерии, предназначенные для переменных, описываемых порядковыми шкалами, можно применять и для анализа переменных, описываемых интервальными шкалами, которые являются частным случаем порядковых. Поскольку все вышеупомянутые критерии являются непараметрическими, т.е. не зависят от вида распределения и мало чувствительны к экспериментальным ошибкам и случайным аномальным значениям, мы настойчиво рекомендуем их применение в случае как порядковых, так и интервальных шкал.

3. Для анализа количественных переменных, измеряемых по интервальным шкалам, допустимо, в случае нормального распределения значений переменных, применять и параметрические методы, а именно t критерий Стьюдента (для сравнения двух выборок), классический дисперсионный анализ ANOVA (для нескольких независимых выборок) или Repeated Measures ANOVA (для не-

Следует подчеркнуть три обстоятельства:

1. Применение для анализа связанных выборок критериев, предназначенных для анализа несвязанных выборок, как правило, не приводит к ошибке первого рода, но вероятность ошибки второго рода (принятие неверной нулевой гипотезы) возрастает: реально существующие тенденции к изменению в результате некоторого воздействия могут быть недооценены. Применение для ана-

скольких связанных выборок). При использовании t критерия Стьюдента следует обращать внимание на то, какая его разновидность применяется: предполагающая или нет равенство дисперсий в выборках; для независимых выборок [independent samples t test] (рис. 2) или связанных выборок [paired samples t test] (рис. 3). В любом случае надо проверять и доказывать нормальность распределений, что создает лишние и зачастую непреодолимые препятствия. Единственным, но весьма слабым доводом в пользу применения t критерия Стьюдента (вместо критериев Манна—Уитни и Вилкоксона) является более широкое знакомство с ним исследователей.

Доля живых менингококков
через 1 ч инкубации

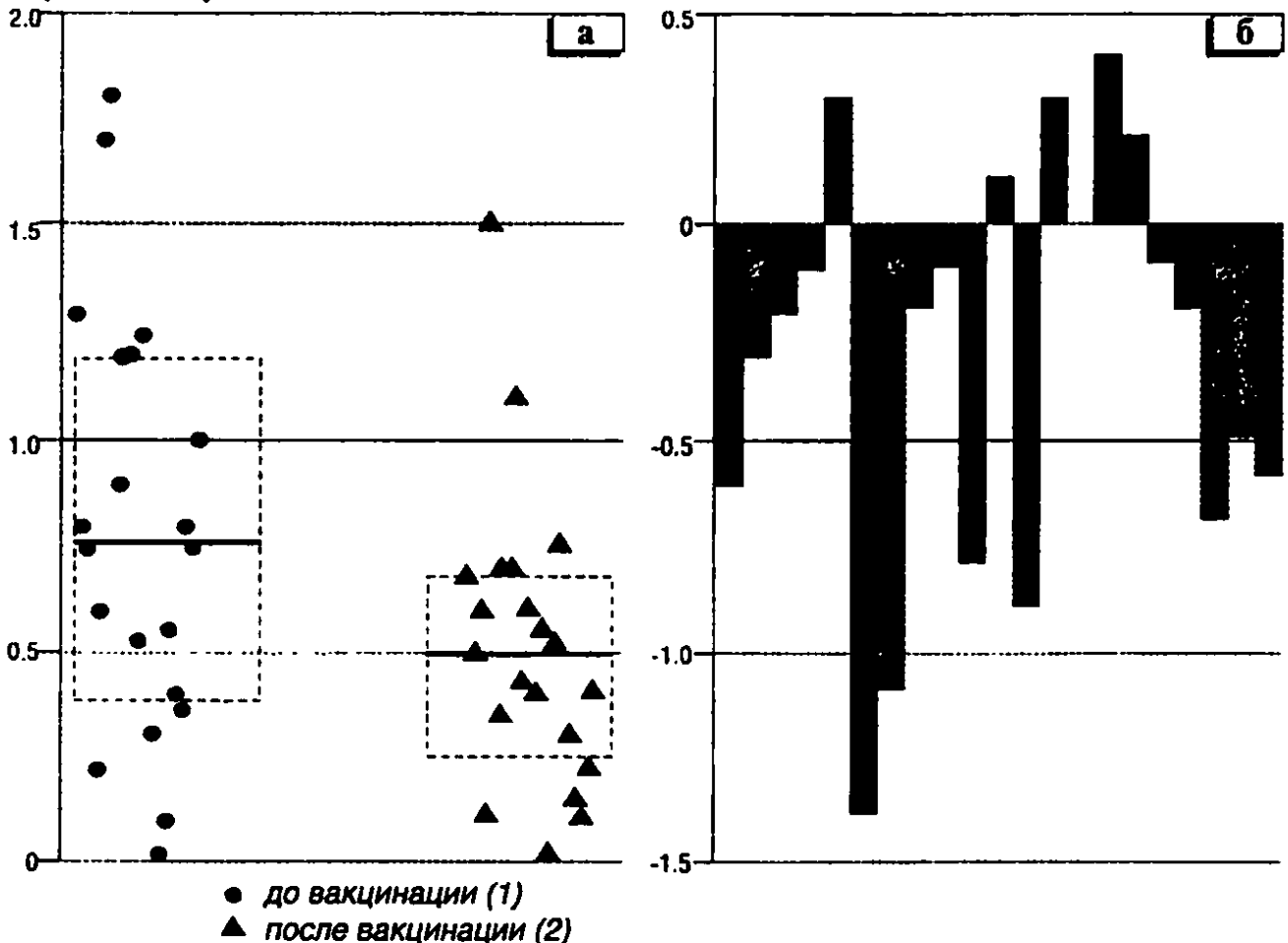


Рис. 3. Логика критериев сравнения двух связанных переменных.

Переменная 1 — бактерицидная активность смеси человеческих нейтрофилов и образцов сыворотки крови 21 пациента с дефицитом терминального компонента комплемента по отношению к менингококкам серогруппы W135. Связанная (парная) переменная 2 — то же, но после вакцинации тех же пациентов менингококковой вакциной.

Постановка опыта: смесь живых менингококков и нейтрофилов инкубируется в индивидуальном образце сыворотки заданное время, после чего определяется доля живых менингококков, количество которых до инкубации принимается за единицу.

а — диапазоны между нижним и верхним квартилями изображены в виде «ящичков», медианные значения переменных 1 и 2 — в виде центральной горизонтальной линии.

б — разница между парными значениями переменных изображена в виде столбиков (отрицательные значения соответствуют возрастанию бактерицидного действия нейтрофилов после вакцинации).

Критерий знаков учитывает, что после вакцинации в 16 случаях произошло увеличение бактерицидной активности (доля живых менингококков ниже при инкубации с нейтрофилами в образце сыворотки, взятой после вакцинации, по сравнению с образцом сыворотки, взятой до вакцинации) и в 5 случаях — уменьшение бактерицидной активности. Критерий знаков позволяет отвергнуть нулевую гипотезу (об отсутствии эффекта вакцинации) на уровне значимости 0.027.

Критерий Вилкоксона принимает во внимание также величину положительных и отрицательных изменений и показывает, что нулевую гипотезу можно отвергнуть с большей степенью достоверности, на уровне значимости 0.013.

Критерий Стьюдента для связанных переменных основан на анализе распределения разницы попарных значений переменных 1 и 2 при допущении, что это распределение нормальное. Согласно t критерию Стьюдента, нулевую гипотезу можно отвергнуть на уровне значимости 0.007; это несколько завышенная оценка достоверности различия, обусловленная наличием нескольких больших (больше 1 по абсолютной величине) попарных разниц.

Примечание. Использование для анализа связанных переменных критериев, предназначенных для анализа независимых переменных, обычно ведет к недооценке достоверности различия. Согласно критерию Манна—Уитни, в данном случае переменные 1 и 2 достоверно различны с уровнем значимости 0.03.

(Данные из материалов публикации [8]).

Анализ связи нескольких переменных

Стандартный способ выявления взаимосвязи нескольких переменных, измеряемых по порядковой или интервальной шкале, — подсчет *коэффициента корреляции*. Для интервальных переменных/шкал, значения которых подчиняются нормальному распределению, можно применять широко известный *коэффициент корреляции Пирсона* [Pearson]. Поскольку этот коэффициент, как и всякий параметрический показатель, весьма подвержен влиянию значений, резко отклоняющихся от среднего, подобные значения разумно исключить из анализа.

Как всегда, правильнее пользоваться ранговыми методами — вычислять коэффициент корреляции Кендалла [Kendall's tau-b] (для порядковых переменных/шкал) или, лучше, коэффициент корреляции Спирмена [Spearman] — непараметрический аналог коэффициента Пирсона для интервальных и порядковых переменных, не подчиняю-

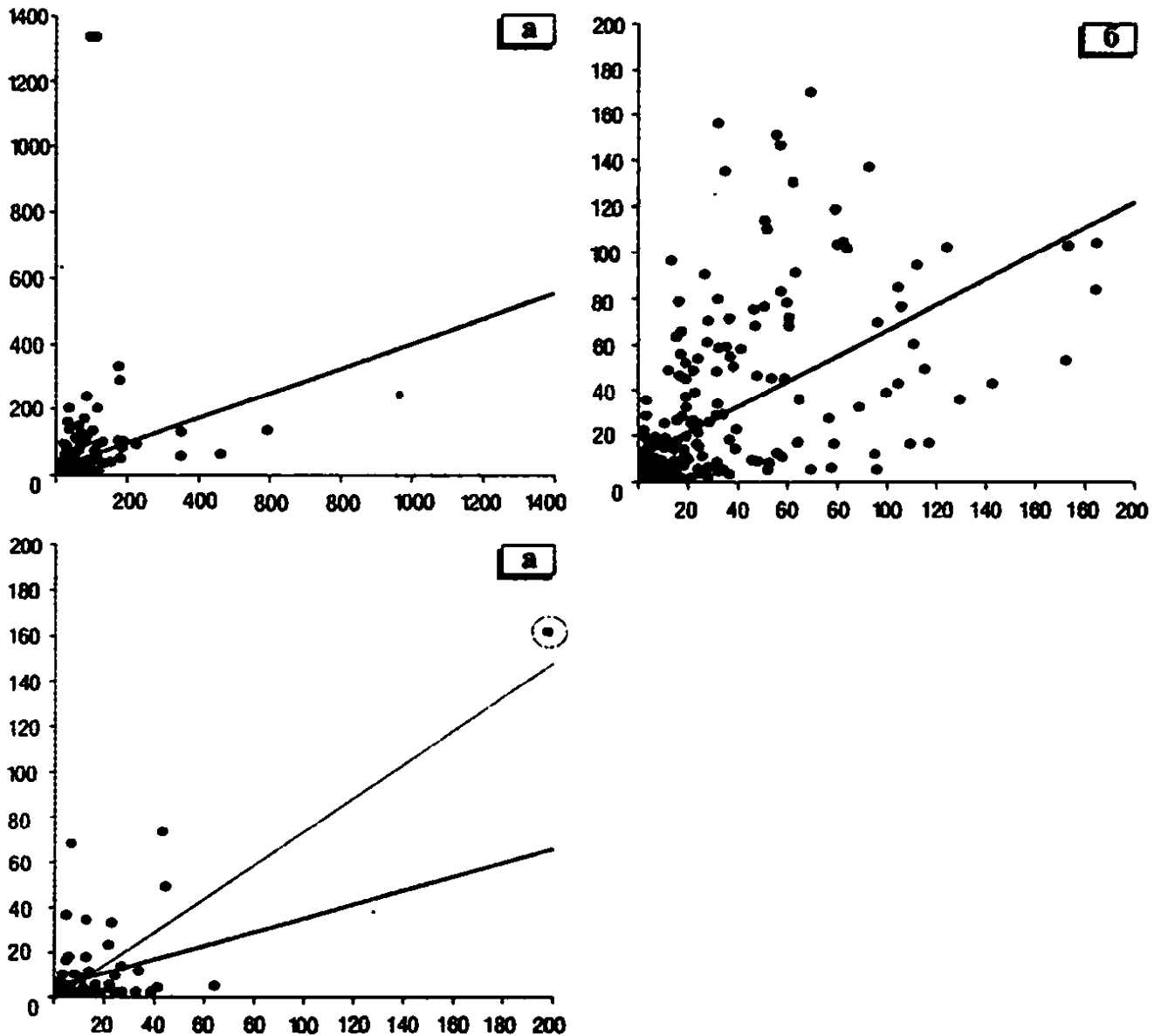


Рис. 4. Логика вычисления и интерпретации коэффициентов корреляции.

Корреляция и линейная регрессия общей концентрации антител против полисахарида менингококков серогруппы А (абсцисса, мкг/мл) и серогруппы С (ордината, мкг/мл) в 257 (а), 244 (б) и 65 (в) образцах сыворотки крови испытуемых.

Расчет коэффициента корреляции по Спирмену дает величину $R=0.69$ (ранговый непараметрический метод); коэффициент корреляции по Кендаллу равен 0.50 (знаковый непараметрический метод), коэффициент корреляции по Пирсону (параметрический метод) равен 0.26 (а). Заниженная оценка степени корреляции, получаемая при

расчетах по Пирсону, обусловлена наличием образцов, в которых концентрация антител очень велика, но мало коррелирована. Непараметрические методы менее чувствительны к подобным обстоятельствам. Все коэффициенты корреляции значимо отличны от нуля ($p < 0.001$, односторонний критерий) в силу большого числа наблюдений.

Если ограничить анализ образцами сыворотки, в которых концентрация специфических антител меньше 200 мкг/мл (б), то коэффициенты корреляции по Спирмену, Кендаллу и Пирсону равны, соответственно, 0.65, 0.47 и 0.57. Не вполне законная операция ограничения проведена для того, чтобы проиллюстрировать зависимость коэффициента корреляции по Пирсону от степени отклонения распределения исследуемых величин от нормального распределения. В этой связи мы настойчиво рекомендуем расчет коэффициента корреляции по Спирмену.

Расчет уравнений линейной регрессии $y = b + ax$, где y — концентрация антител против полисахарида менингококков серогруппы С, а x — концентрация антител против полисахарида менингококков серогруппы А, дает для данных, представленных на рис. 4а, $b = 27$ (95% доверительный интервал 10-44), $a = 0.38$ (95% доверительный интервал 0.20-0.55), в то время как для данных, представленных на рис. 4б, $b = 9.5$ (95% доверительный интервал 4.4-14), $a = 0.55$ (95% доверительный интервал 0.45-0.64). Соответствующие взаимосвязи изображены в виде линий на рис. 4а и 4б. Таким образом, такие параметрические методы аппроксимации, как линейная регрессия, также подвержены влиянию искажений, обусловленных «ненормальным» распределением исследуемых величин.

Этот пример демонстрирует, что высокая степень корреляции отражает лишь их взаимосвязь, но не обязательно зависимость одной переменной от другой. Наши испытуемые были привиты поливалентной вакциной, содержащей полисахариды менингококков серогрупп А и С. В образцах, взятых в течение первого года после вакцинации, наблюдаются высокие концентрации антител как против полисахарида серогруппы А, так и против полисахарида серогруппы С. В образцах, взятых до вакцинации или 2-3 года спустя, концентрации специфических антител низкие. Несмотря на корреляцию, нельзя заключить, что продукция антител против полисахарида А обуславливает продукцию антител против полисахарида С.

Действительно, в 65 образцах, взятых до вакцинации (рис. 4в), уровень корреляции ниже: коэффициенты корреляции по Спирмену, Кендаллу и Пирсону равны, соответственно, 0.33, 0.23 и 0.84. Высокое значение коэффициента по Пирсону является грубым артефактом, обусловленным единственной «выпадающей» точкой (выделена кружком). После удаления этого образца из анализа коэффициенты корреляции по Спирмену, Кендаллу и Пирсону равны, соответственно, 0.30, 0.20 и 0.26. На рисунке также показана линейная регрессия с учетом этого образца (тонкая линия) и без него (толстая линия).

(Данные из материалов публикации [9].)

щихся нормальному распределению (рис. 4). Коэффициент Пирсона равен единице (или минус единице) тогда и только тогда, когда две переменные (x и y) связаны линейной зависимостью ($y=b+ax$). Коэффициент Спирмена (или Кендалла) равен 1, если две переменные связаны правилом: большему значению переменной x всегда соответствует большее значение переменной y . Чем ниже коэффициент корреляции, тем сильнее отклонение от этих правил. Иногда вместо коэффициента корреляции приводят значения *квадрата коэффициента корреляции*. Не стоит забывать, что при малом числе наблюдений даже большое значение коэффициента корреляции может быть статистически недостоверным. Статистические программы могут оценить уровень значимости, с которым можно отвергнуть ненаправленную нулевую гипотезу (коэффициент корреляции равен нулю) или направленную гипотезу (коэффициент корреляции не меньше или не больше нуля).

Внимание! Отсутствие непосредственной корреляции двух переменных не обязательно означает отсутствия взаимосвязи между ними: коэффициент корреляции переменных x и y , жестко связанных по правилу $y=\sin(ax)$, равен нулю (при достаточно широком диапазоне изменений x). Не менее важно то, что наличие корреляции двух переменных не означает их причинно-следственной связи (рис. 4). В некоем тесте можно выявить высокую корреляцию массы тела человека и знания им статистики. Однако, возможно, эта взаимосвязь обусловлена корреляцией с возрастом как массы тела человека, так и знания им статистики. Для поиска причинно-следственных отношений необходим продуманный дизайн эксперимента или наблюдений и применение более изощренных методов статистического анализа и математического моделирования.

Если связь одной количественной переменной y (условно называемой «зависимой») с другой количественной переменной x (условно называемой «независимой») приблизительно линейна, т.е. предположительно описывается уравнением вида $y=b+ax$, то интерес представляет



не только значение коэффициента корреляции, но и величина коэффициентов a и b . Процедура их вычисления называется *линейной регрессией* [linear regression] и основана на минимизации отклонений реальных значений x и y от теоретической кривой. Часто минимизация проводится методом наименьших квадратов. Многие зависимости «стандартного» вида — квадратическая, логарифмическая и т.п. — сводятся к линейным путем соответствующих математических преобразований. Например, зависимость вида $y = b \exp(ax)$ сводится к виду $\ln(y) = \ln(b) + ax$, после чего вычисляются коэффициенты, входящие в исходное уравнение [curve estimation]. Возможна также оценка (подбор) значений параметров в уравнениях нестандартного (задаваемого исследователем) вида $y = f(x)$ с нелинейной зависимостью от параметров [nonlinear regression], но эта сравнительно сложная процедура требует более глубоких знаний статистики и математики.

В общем случае линейная регрессия позволяет выявить связь одной зависимой количественной переменной y с несколькими независимыми количественными переменными $x_1, x_2, x_3 \dots$ и так далее до x_n , т.е. связь вида $y = b + a_1 x_1 + a_2 x_2 + a_3 x_3 + \dots + a_n x_n$, и вычислить величину коэффициентов a и b . При этом процедура *шагового отбора* [stepwise selection] позволяет по определенному критерию отбросить несущественные независимые переменные (слабо связанные с зависимой переменной) и оставить только существенные независимые переменные.

!

Внимание! В случае нелинейной зависимости или зависимости от порядковых переменных применение процедуры пошагового отбора незаконно и приводит к ошибочным заключениям.

В статистическом анализе важна и часто встречается ситуация, когда зависимая переменная не количественная, а дихотомическая качественная, принимающая значения ДА=1 или НЕТ=0. Пример такой переменной — исход заболевания (или испытания на лабораторных животных), при этом ДА=летальный исход, НЕТ=выздоров-

ление. Зависимость риска летального исхода от значений того или иного интересующего нас фактора (независимой переменной) обычно не линейна, но имеет сигмоидную форму, отражающую резкое увеличение риска после превышения фактором некоторого порога. В статистике принято полагать, что связь подобной переменной с другими количественными или качественными переменными (факторами) наиболее адекватно выявляется после так называемого *логистического преобразования*:

$$\ln(P/1-P) = C + \sum \beta_i X_i,$$

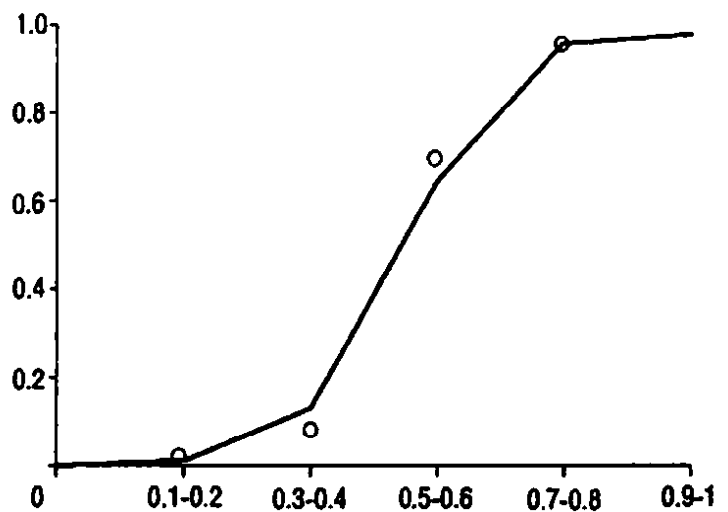
где $\sum \beta_i X_i$ — сумма значений определенных факторов X_i , умноженных на соответствующий коэффициент β_i , C — константа, P — вероятность (или доля) летальных исходов.

Подбор коэффициентов осуществляется с помощью процедуры логистической [logistic] регрессии (рис. 5).

Особого упоминания требуют три конкретные, но существенные медико-биологические проблемы.

Рис. 5. Логистическое преобразование и процедура логистической регрессии. Взаимосвязь оценки тяжести состояния при поступлении больного по прогностической шкале PtLC и исхода заболевания при менингококковой инфекции.

По оси абсцисс — значения (от 0 до 1) по нормированной прогностической шкале PtLC; по оси ординат — летальность в соответствующих подгруппах



больных, имевших прогностические оценки, принадлежащие интервалам значений, указанным на оси абсцисс.

Кружки — реальная летальность в подгруппах больных; линия — вероятность летального исхода в данных подгруппах, высчитанная с помощью логистического преобразования по формуле: $\ln(P/1-P) = C + \sum \beta_i X_i$ или, что то же самое, $P = \exp(C + \sum \beta_i X_i) / (\exp(C + \sum \beta_i X_i) + 1)$, где $\sum \beta_i X_i$ — сумма значений определенных прогностических факторов X_i , умноженных на соответствующий коэффициент β_i , C — константа, P — вероятность (или доля) летальных исходов.

Анализ наступления событий («дожития» или «выживаемости») [survival analysis]

Вычисление и сравнение (в различных подгруппах) времени наступления некоторого «события», определение вероятности наступления «события» через заданный интервал времени — простая задача, решаемая вышеупомянутыми непараметрическими методами, если это «событие» происходит с каждым наблюдаемым объектом.

Скажем, можно взять 50 мужчин и 50 женщин в Москве и, наблюдая их всех от рождения до смерти, определить среднюю продолжительность жизни в Москве и сравнить достоверность ее различия в зависимости от пола. Однако весьма распространена ситуация, когда данное событие не регистрируется у части наблюдаемых объектов и такие наблюдения необходимо рассматривать как цензурированные [censored events]. Простейшая причина этого — ограниченное время наблюдения. В предыдущем примере мы могли наблюдать мужчин и женщин в течение 80 лет и не знать точно, сколько проживут те, кто дожил до 80. Другая причина — выпадение части объектов из исследования по независимым от экспериментатора причинам, в предыдущем примере, допустим, в результате переезда в другой город.

Для статистического анализа подобных ситуаций разработана процедура Survival Analysis, включающая статистические критерии для оценки достоверности различия времени наступления события в разных группах, а именно критерии Каплана-Мейера [Kaplan-Meier], Гехана [Gehan] и др. (рис. 6). Для оценки связи времени наступления события с набором независимых переменных, часть которых может сама зависеть от времени, применяется рег-

рессионный анализ Кокса [Cox regression, Cox with time dependent covariates].

Рис. 6. Логика «анализа выживаемости».

Эффект вакцинации лиц с дефицитом терминального компонента комплемента (ДТКК) тетравалентной полисахаридной вакциной на их заболеваемость менингококковой инфекцией.

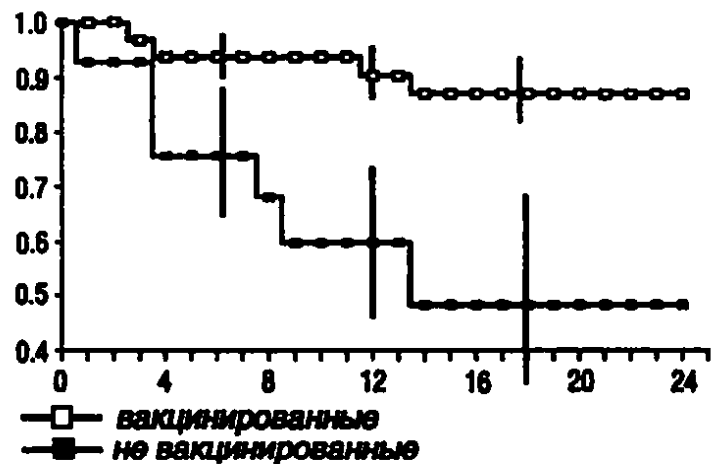
По оси абсцисс — время наблюдения (кварталы), прошедшее после вакцинации лиц с ДТКК или после их включения в контрольную невакцинированную группу.

По оси ординат — общая доля лиц в конкретной группе, которые за данное время наблюдения не заболели менингококковой инфекцией.

Медицинские предпосылки исследования и интерпретация его результатов: лица с ДТКК в силу отсутствия бактерицидной активности плазмы крови чрезвычайно подвержены менингококковой инфекции (заболеваемость в тысячи раз выше, чем у лиц с нормальной системой комплемента), в том числе способны заболевать неоднократно (до 5-6 раз в жизни). Для изучения возможности частичного предотвращения заболеваемости проведена вакцинация менингококковой вакциной.

Как в группе вакцинированных, так и в группе невакцинированных пациентов наблюдаются новые случаи заболевания, графически отображаемые как ступеньки на зависимости «общей доли незаболевших» от времени наблюдения. Высота ступенек различна, во-первых, потому, что группа вакцинированных больше (31 человек), чем группа невакцинированных (14 человек), во-вторых, потому, что заболевшие пациенты выбывают из анализа (цензурируются) и одному заболевшему из 14 (начало наблюдения) соответствует меньшая ступенька, чем одному заболевшему из 9 (4-й год наблюдения). К концу периода наблюдения заболевают далеко не все пациенты и это является предпосылкой к использованию процедуры «анализа выживаемости», которая оценивает средний риск нового случая заболевания как 0.1 на человека в год в группе невакцинированных и 0.03 на человека в год в группе вакцинированных. Использование соответствующих критериев (Вилкоксона—Гехана и др.) позволяет на уровне значимости 0.05 отвергнуть нулевую гипотезу о равенстве заболеваемости в группах (на графике показана оценка стандартной ошибки для значений «общей доли незаболевших» спустя 6, 12 и 18 кварталов после вакцинации). Таким образом, анализ демонстрирует наличие протективного эффекта вакцинации лиц с ДТКК менингококковой вакциной.

(Данные из материалов публикации [11].)



Статистические понятия, связанные с таблицами 2×2

Таблицы 2×2 часто появляются в следующих типах исследований (см. таблицу).

Результат «нового» диагности- ческого теста	Результат «общепризнанного» диагностического теста	
	положительный	отрицательный
Положительный	Число истинно положительных результатов нового теста (ИП)	Число ложноположительных результатов нового теста (ЛП)
Отрицательный	Число ложноотрицательных результатов нового теста (ЛО)	Число истинно отрицательных результатов нового теста (ИО)

Исходя из этой таблицы, можно вычислить *специфичность* [specificity] «нового» теста относительно «стандартного» теста, равную $100 \times \text{ИО} / (\text{ИО} + \text{ЛП})$, и его *чувствительность* [sensitivity], равную $100 \times \text{ИП} / (\text{ИП} + \text{ЛО})$. Очевидно, что специфичность тем больше (лучше), чем меньше число ложноположительных результатов; чувствительность тем больше (лучше), чем меньше число ложноотрицательных результатов. Довольно часто, хотя и не всегда, при повышении специфичности чувствительность теста понижается и наоборот. Иногда «новый» тест характеризуют также долей истинных результатов, равной $100 \times (\text{ИО} + \text{ИП}) / (\text{ИО} + \text{ИП} + \text{ЛО} + \text{ЛП})$.

!

Внимание! Специфичность и чувствительность теста не есть, как иногда думают, его абсолютные и неизменные характеристики. Они непосредственно зависят от материала, на котором испытывается «новый» тест. Включив, например, в этот материал много заведомо «отрицательных» образцов, мы тем самым увеличиваем число ИО результатов и, тем самым, абсолютное значение специфичности. Кратко говоря, структура материала (экс-

периментальных образцов), на которых испытывается тест, должна быть подобна той, к которой тест будет применяться на практике.

Внимание! Не следует путать понятия чувствительности и *предела или точности измерения* теста [detection limit]. Если тест способен выявить присутствие 1 пМ/мл вещества в растворе — это именно точность измерения, а не чувствительность теста.



Аналогичные таблицы возникают при анализе прогностических факторов. Представим себе, что перед нами стоит задача предсказать неблагоприятный исход заболевания по величине некоего лабораторного показателя. Для оценки прогностических факторов значение любого клинико-лабораторного показателя дихотомически можно отнести к одной из двух групп — к группе значений, предсказывающих конкретный исход (в нашем случае, допустим, летальный), т.е. к группе «положительных» (по статистической терминологии) значений, либо к группе значений, не предсказывающих (отрицающих) данный исход, т.е. к группе «отрицательных» значений. Границу (порог) между двумя группами, т.е. критерий отнесения данного значения показателя к «положительным» или «отрицательным», выберем так, чтобы максимизировать число верных прогнозов и «разность рисков». После этого определим для показателя: число истинно-положительных прогнозов (ИП) — число больных, у которых согласно значению данного параметра мог быть предсказан летальный исход и которые действительно умерли; число ложноположительных (ЛП) прогнозов (согласно значению данного параметра предсказывается летальный исход, но больной выжил); число ложноотрицательных (ЛО) прогнозов (согласно значению данного параметра предсказывается выздоровление, но больной умирает); число истинно отрицательных (ИО) прогнозов (больной должен был бы выздороветь и выздоравливает). Сведем данные в таблицу (см. таблицу).

Используя эти данные, можно вычислить такие важные характеристики прогностического фактора, как специфичность, равную $100 \times \text{ИО} / (\text{ИО} + \text{ЛП})$, чувстви-

Исход заболевания	Прогноз (на основании значения некоторого фактора или их комбинации)	
	неблагоприятный	благоприятный
Неблагоприятный	Число истинно положительных прогнозов (ИП)	Число ложно- отрицательных прогнозов (ЛО)
Благоприятный	Число ложно- положительных прогнозов (ЛП)	Число истинно отрицательных прогнозов (ИО)

ность, равную $100 \times \text{ИП} / (\text{ИП} + \text{ЛО})$, и долю правильных прогнозов: $100 \times (\text{ИО} + \text{ИП}) / (\text{ИО} + \text{ИП} + \text{ЛО} + \text{ЛП})$.

«Разность рисков» (РР) [factor fatality difference] определяется как разность между долей (в %) летальных исходов в группе больных с «положительным» и в группе больных с «отрицательным» значением фактора: $\text{РР} = 100 \times \{ \text{ИП} / (\text{ИП} + \text{ЛП}) - \text{ЛО} / (\text{ИО} + \text{ЛО}) \}$. Чем выше «разность рисков» (летальности) между этими двумя группами, тем выше прогностическая ценность данного конкретного показателя.

Относительный риск (ОР) [relative risk] прогнозируемого исхода в группе фактор-положительных пациентов вычисляется по формуле: $\text{ОР} = \{ \text{ИП} / (\text{ИП} + \text{ЛП}) \} / \{ \text{ЛО} / (\text{ИО} + \text{ЛО}) \}$. Чем больше рассматриваемый фактор влияет на прогнозируемый исход, тем больше ОР отличается от единицы. Современные пакеты статистических программ обладают возможностью не только рассчитать величину ОР, но и определить для него 95% доверительный интервал. Если исследуемый фактор обладает реальной прогностической ценностью, то доверительный интервал не включает 1. Относительный риск иного исхода в группе фактор-положительных пациентов вычисляется, соответственно, по формуле: $\text{ОР} = \{ \text{ЛП} / (\text{ИП} + \text{ЛП}) \} / \{ \text{ИО} / (\text{ИО} + \text{ЛО}) \}$.

Степень сопряженности таблицы 2×2 , в данном случае степень связи прогноза и исхода, характеризуют также с помощью понятия *коэффициента асимметрии* [odds ratio], равного $(\text{ИП} \times \text{ИО}) / (\text{ЛП} \times \text{ЛО})$. Если связь полностью от-

сутствует, коэффициент асимметрии равен единице, если имеется — значимо отличается от единицы.

Опасность множественных сравнений

В медико-биологических исследованиях существует ловушка, о существовании которой не все знают. Представим себе, что мы взяли сотню-другую людей, истинно случайным образом, например, с помощью генератора случайных чисел, разбили их на две группы — опытную и контрольную, и стали задавать произвольные, но многочисленные вопросы или проводить разнообразные тесты и измерения. К нашему удивлению, мы достаточно скоро, через 20-50 вопросов, обнаружим, что надежные критерии «статистически значимо», на уровне значимости 0.05, показывают, что группы отличаются по какому-либо показателю, например по любви к мороженому или артериальному давлению. Но ведь группы были образованы случайно! В чем ошибка?

Ошибка в том, что *одномерные* критерии сравнения, обсуждаемые и рекомендуемые выше, предназначены для оценки различия групп по одному, конкретному и заранее выбранному показателю, а не для ответа на вопрос: «А нет ли между группами хоть какого-нибудь различия»?

Разработка *многомерных* статистических критериев, предназначенных, в том числе, для множественных сопоставлений свойств групп, только начинается [1]. Существующие методы заведомо неполны и, в силу своей сложности, не получили широкого распространения в медико-биологических кругах.

С другой стороны, в комплексном и дорогостоящем исследовании влияния нового препарата на исход и последствия инфаркта миокарда трудно удержаться от измерения множества физиологических параметров, касающихся сердечно-сосудистой системы, системы свертывания крови и т.п. И нередко, как в примере с мороженым, 19 показателей остаются неизменными, а по 20-му «после назначения препарата наблюдается статистичес-

ки значимое улучшение, что вселяет определенные надежды».

Что делать? Однозначного рецепта пока нет, но есть определенные эмпирические правила.

1. В случае множественных сопоставлений уровень значимости, на котором отвергается нулевая гипотеза, следует задавать более жестко, например 0.01 или даже 0.001.

2. Полезно сравнивать не контрольную группу как целое с опытной группой как целым, а сначала случайным образом выбранную половину контрольной группы с половиной опытной группы. Если эффект препарата, обнаруживающийся при сравнении первых половин (назовем их «отработочными» подгруппами), воспроизводится и при сравнении вторых половин контрольной и опытной групп («тестовых» подгрупп) — это некоторое свидетельство в пользу его реального существования. Если нет — это серьезное указание на то, что эффект является артефактом неадекватного статистического анализа.

3. Результаты, полученные в ходе множественных сопоставлений, имеют в науке «отложенный статус» и должны быть окончательно подтверждены дополнительными независимыми исследованиями и их метаанализом.

На этом мы остановимся, хотя статистический анализ, разумеется, включает множество других, более изощренных методов, подробнее рассмотренных в работах [1, 2, 6] и заслуживающих хотя бы беглого перечисления.

Взаимосвязь (одной) количественной «зависимой» переменной с несколькими номинальными переменными (факторами) может быть выявлена с помощью моделей ANOVA (*дисперсионного анализа*). В том случае, если зависимая переменная зависит не от непосредственно наблюдаемых независимых переменных, а от их некоторых, заранее неизвестных комбинаций, применяют методы *факторного анализа* [factor analysis].

Классификация «случаев», т.е. наборов значений неких переменных, относящихся к одному объекту, часто проводится методом *кластерного анализа*, т.е. разбиения множества «случаев» на заранее заданное число более гомогенных подгрупп-кластеров [clusters]. Возможно так-

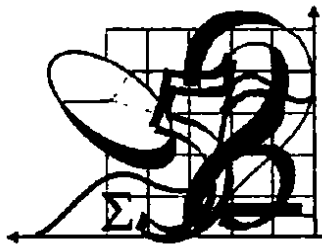
же разбиение множества «случаев» на кластеры, когда «разумное» число более гомогенных подгрупп заранее не известно. Это так называемый *иерархический* [hierarchical] кластерный анализ.

Внимание! Результат классификации может зависеть от способа задания «расстояний», т.е. меры различия, между «случаями» и принципа группировки, которые должны быть выбраны адекватно исследуемой проблеме.

Отнесение каждого «случая» к одной из заранее заданных подгрупп (категорий) на основе значений набора некоторых переменных, сопровождающееся отбором «значимых» независимых переменных и оценкой качества классификации — задача *дискриминантного анализа* [discriminant analysis].

Отдельный комплекс подходов существует для анализа *временных рядов* [time series], т.е. зависимости некоторой численной переменной от другой временной переменной. Интерес представляет выявление *тренда* [trends] и учет *известной периодичности* зависимой переменной (например, сезонности) или поиск неизвестной периодичности (*спектральный анализ*, построение *периодограмм*); выявление *автокорреляций* и *кросскорреляций* ее значений, т.е. взаимосвязи значений меняющейся во времени переменной между собой или со значениями другой переменной; прогноз поведения временных рядов на основе моделей «со скользящим средним»; *авторегрессионные модели*, симуляционное моделирование.





* * *

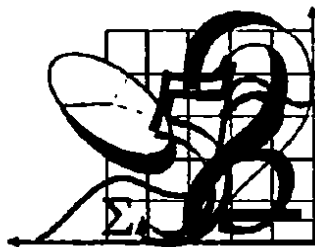
Может показаться, что в основном в нашей публикации обсуждаются тривиальные вещи. Это не так. Возьмем всего один конкретный номер одного авторитетного отечественного биомедицинского журнала. Экспериментальных статей — 29. Из них: не нуждаются в применении статистического анализа в силу предмета исследований — 3, нуждаются — 26. Поставленные авторами статистические задачи: сравнение значений переменных в разных экспериментальных группах (21 статья), корреляционный анализ (3), частотный анализ (4), прогнозирование (1).

Характеристики распределения указаны в виде «что-то±что-то» — 19 статей. Из них: указано, что первое «что-то» — среднее арифметическое, а второе «что-то» — стандартное отклонение в одной статье; первое «что-то» обозначено M , второе «что-то» — m (15 статей), вероятно, следовало бы предположить, что M — среднее арифметическое, m — стандартная ошибка среднего (однако в одном случае это явно не так: высчитанное через m среднее±стандартное отклонение равнялось бы 83 ± 180 , что маловероятно); статистический смысл «что-то±что-то» не указан (3 статьи). Из 19 статей, содержащих «что-то±что-то», данные представлены в неверном формате (типа 5.865 ± 1.194 , 187.05 ± 15.02 , 29179 ± 2827) — в 8 статьях.

Из 26 исследований, нуждавшихся в статистическом анализе: статистические методы не применялись (2 ста-

тьи); применение статистических методов упомянуто, но никаких следов этого в тексте не содержится (3 статьи); применялся некий неописанный критерий, вероятно, t критерий Стьюдента, — в 6 статьях; применялся t критерий Стьюдента — в 10 статьях; применялся критерий Фишера — в 2 статьях; применялся метод ANOVA в различных модификациях — в 2 статьях; вычисляли коэффициент корреляции — в 3 статьях (из них вид коэффициента корреляции не указан нигде, в одной статье вводится загадочное понятие «суммарного коэффициента корреляции» достигающего значения 1.514!); вычисляли уравнение линейной регрессии — в одной статье (при этом записывая уравнение в неверном формате $Y=193.08+24.849 \times X$); применялся непараметрический критерий Манна—Уитни — в 2 статьях; применялся непараметрический критерий χ^2 и метод Каплана—Мейера [Survival Analysis] в одной статье; вычисляли «долю точных прогнозов» — в одной статье. Из 20 статей, использующих параметрические методы анализа, в первую очередь t критерий Стьюдента, правомерность их применения (нормальный тип распределения переменных) не обсуждается и не подтверждается ни в одной. Применение компьютерных программ статистической обработки упоминается в 4 статьях. Таким образом, даже беглый взгляд обнаруживает бедный спектр применяемых статистических методов и неудовлетворительный уровень анализа более чем в половине статей. В сущности, только в 3-5 публикациях используемый набор и уровень применения статистических методов адекватен предмету и задачам исследования.

Несомненно, желательно повышение уровня преподавания биомедицинской статистики в высших учебных заведениях, а также подготовка или перевод монографий и справочных пособий, рассматривающих проблемы биомедицинской статистики с современных позиций, для нужд исследователей. Можно было бы только приветствовать более тесное сотрудничество экспериментаторов с профессиональными статистиками.



Заключение

Рассмотрим последовательность этапов, из которых складывается профессиональный статистический анализ в биологии и медицине.

1. До начала исследования сформулируйте, хотя бы для себя, его задачу (задачи) в виде вопроса, допускающего прямую экспериментальную проверку и последующий статистический анализ. «Исследовать влияние лекарства А (или вещества Б) на особенности течения заболевания Х (или процесса Y)» — это не вопрос. Вопрос — «Как после применения лекарства А (или вещества Б) изменяются конкретные, допускающие формальную оценку и имеющие медико-биологический смысл параметры заболевания Х (или процесса Y)?».

2. Определите схему экспериментов или наблюдений, предположительно позволяющих дать ответ на поставленный вопрос: группы исследуемых объектов и ситуаций, регистрируемые параметры, применяемые методы. В дальнейшем эти параметры станут анализируемыми переменными, а каждой группе (или ситуации) будет поставлено в соответствие значение некоей качественной, классифицирующей переменной. Неверно организованный эксперимент не спасет никакой статистический анализ.

3. Проведите эксперимент (наблюдение), сохраните результаты, лучше всего в виде базы данных, созданной, например, с помощью программы Microsoft Access и ра-

циональным образом включающей анализируемые и классифицирующие переменные.

4. Определите тип стоящей перед вами статистической задачи: описание массива данных (распределения) и/или проверка статистических утверждений (нуль-гипотез), сформулируйте нуль-гипотезы, которые вы намерены проверить.

5. При описании массива данных оцените медиану и среднее арифметическое распределение. Если они расходятся более чем на 10-20%, то распределение, вероятнее всего, не нормальное (не гауссовское) и применение параметрических методов его описания и анализа (например, критериев Стьюдента) неправомерно. Однако, если медиана и среднее арифметическое близки, это еще не означает, что распределение нормальное. В целом, автор рекомендует познакомиться с непараметрическими методами и их широко использовать.

6. Характеристики распределения могут быть представлены в виде $A (L, H)$, где A — медиана, L — нижний квантиль, H — верхний квантиль; в виде $A (min, max)$, где A — медиана, min — наименьшее значение, max — наибольшее значение (этот способ не рекомендуется, если есть резко выпадающие из распределения значения); $M \pm s$, где M — среднее арифметическое, s — стандартное отклонение (этот способ рекомендуется использовать только для нормальных распределений).

Не пишите 17.087 ± 6.913 ! Пишите 17.1 ± 6.9 или даже 17 ± 7 . Если точные количественные значения параметров распределения менее важны, чем тенденции их изменений в различных ситуациях, рассмотрите возможность представления результатов в графической форме (рис. 7, 8).

7. Выберите соответствующий критерий для проверки вашей нуль-гипотезы. Не забывайте о различии между качественными, порядковыми и количественными (интервальными) переменными, а также между связанными и независимыми переменными.

8. Используйте выбранный критерий. Для этого лучше всего перенести (экспортировать) ваши данные в профессиональный статистический пакет программ, такой как

Statistica®, SPSS, Statgraphics (в принципе, вы изначально можете хранить данные в файле данных этих программ, но базы общего назначения, такие как Microsoft Access, имеют расширенный и более удобный интерфейс для разнообразных запросов, форм ввода, отчетов).

9. Обратите внимание на уровень значимости, с которым вы можете отвергнуть нулевую гипотезу. Имеет смысл записать (и приводить в публикации) либо точную величину уровня значимости, например 0.012, или, если так кажется привычнее, его верхнюю оценку по типу $p < 0.1$, или $p < 0.05$, или $p < 0.01$ и т.д. Имейте в виду, что существует негласное соглашение о том, что запись $p < 0.05$ означает на самом деле $0.01 < p < 0.05$ и так далее, т.е. уровень значимости 0.008 правильнее указывать, как $p < 0.01$, а не как $p < 0.05$.

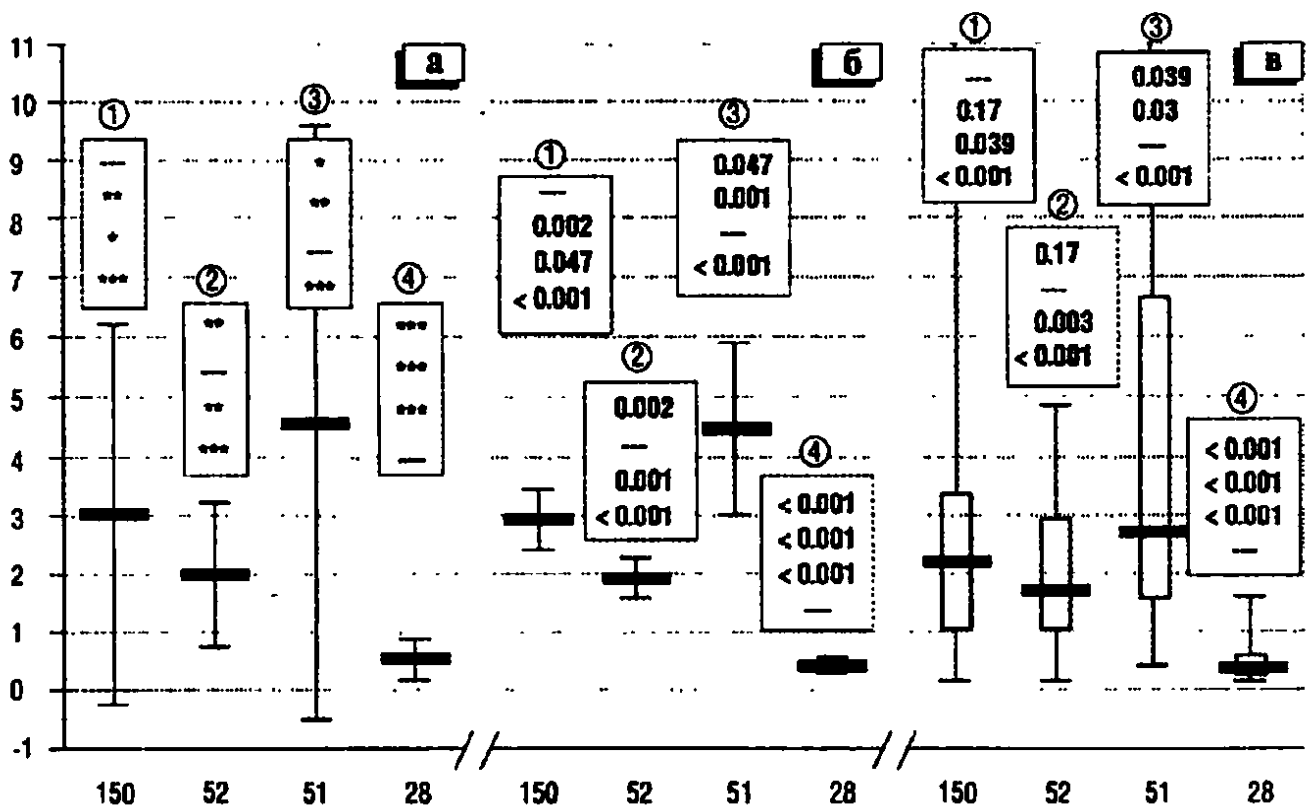


Рис. 7. Варианты графического представления результатов.

① — бактериальный менингит, вызванный *N. meningitidis*; ② — бактериальный менингит, вызванный *H. influenzae* типа b; ③ — бактериальный менингит, вызванный *S. pneumoniae*; ④ — серозный энтеровирусный менингит.

Цифры под рисунком — число обследованных в каждой группе.

По оси ординат — концентрация белка в спинномозговой жидкости (СМЖ), пункция при поступлении в стационар.

(Данные из материалов работы Тютюник Е.Н., Венгерова Ю.Я., Платонова А.Е., готовящейся публикации)

а — среднее арифметическое значение (средняя горизонтальная засечка) \pm стандартное отклонение (вертикальные отрезки от среднего арифметического до верхней и нижней горизонтальной засечек). Обращают на себя внимание неравенство стандартных отклонений (разброс значений концентрации белка выше для пневмококкового и менингококкового менингитов), а также косвенные признаки отличия распределений от нормального (средняя концентрация минус стандартное отклонение принимает отрицательные значения для пневмококкового и менингококкового менингитов, что явно не имеет биологического смысла).

б — среднее арифметическое значение (средняя горизонтальная засечка) $\pm 95\%$ доверительный интервал для среднего значения (вертикальные отрезки от среднего арифметического до верхней и нижней горизонтальной засечек). В отличие от стандартного отклонения доверительный интервал уменьшается при увеличении числа наблюдений и характеризует не распределение значений, а точность определения среднего. При данном числе пациентов в группах доверительный интервал меньше стандартного отклонения, т.е. среднее значение концентрации белка можно определить достаточно точно.

в — представление в виде «ящика с усами» [box-and-whisker plot]. Диапазоны между нижним и верхним квартилем изображены в виде «ящиков», медианные значения в виде центральной горизонтальной линии этого ящика; «усы» тянутся от «ящика» вверх и вниз, достигая, соответственно, максимального и минимального значений концентрации белка в группе. Очевидна асимметричность, т.е. «ненормальность» распределений, которые указывают на очень большие концентрации белка в СМЖ некоторых пациентов. Поэтому среднее арифметическое значение больше значения медианы и является неудачной, завышенной оценкой среднего, а стандартное отклонение является неудачной оценкой ширины распределения, завышенной по сравнению с интерквартильным интервалом.

Для каждой группы в виде столбика цифр в квадратике внутри рисунка показан уровень значимости различия концентрации белка в СМЖ в данной группе с концентрацией белка в СМЖ в группе с менингококковым менингитом (первая строка), или в группе с гемофильным менингитом (вторая строка), или в группе с пневмококковым менингитом (третья строка), или в группе с энтеровирусным менингитом (четвертая строка). Пустые строки обозначены прочерком. (Примечание. Для упрощения рисунка уровень значимости можно кодировать звездочками: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, как на рис. 7а).

Уровни значимости различий, приведенные на рис. 7а и 7б, вычислены по критерию Стьюдента для независимых переменных с поправкой на неравность дисперсий. Уровни значимости различий, приведенные на рис. 7в, вычислены по непараметрическому кри-

терию Манна—Уитни. Обратите внимание, что уровни значимости, высчитанные по критерию Стьюдента, нередко количественно отличны от уровней значимости по критерию Манна—Уитни. В одном случае разница принципиальна: критерий Стьюдента приводит к ошибочному заключению о различии концентрации белка в СМЖ при менингитах, вызванных *N. meningitidis*, и менингитах, вызванных *H. influenzae*.

Текстовое представление той же информации выглядело бы приблизительно так:

Концентрация белка в СМЖ, г/л	Менингококковый менингит
Среднее арифметическое ± стандартное отклонение	3.0 ± 3.3
Среднее арифметическое ±95% доверительный интервал	3.0 ± 0.6
Минимум — нижний квартиль — <u>медиана</u> — верхний квартиль — максимум	0.06 - 1.0 - <u>2.2</u> - 3.3 - 12
Уровень значимости отличия от концентрации белка при менингите, вызванном <i>H. influenzae</i> : по критерию Манна—Уитни/по критерию Стьюдента	0.17 / 0.002
Уровень значимости отличия от концентрации белка при пневмококковом менингите: по критерию Манна—Уитни/по критерию Стьюдента	0.039 / 0.047
Уровень значимости отличия от концентрации белка при энтеровирусном менингите: по критерию Манна—Уитни/по критерию Стьюдента	< 0.001 / < 0.001

Конечно, для реальной публикации разумно выбрать один, наиболее удобный способ представления данных и один, наиболее адекватный критерий сравнения.

10. Не забывайте о статистическом смысле уровня значимости. Допустим, вы утверждаете, что значения некоторого параметра в двух группах различаются с уровнем значимости 0.1. Несколько упрощая, это можно переформулировать так: если бы эти две группы на самом деле были одинаковы, то, повторив ваш эксперимент 10 раз (при этом в каждом одиночном эксперименте определяются 2 распределения), в 1 случае из 10 вы обнаружили бы наблюдаемые различия, обусловленные при этом чисто случайными причинами. Уровень значимости 0.05 соответствует тому, что неверное заключение о различии было бы получено в одном из 20 воображаемых повторных экспериментов и т.д. Поэтому то, какой

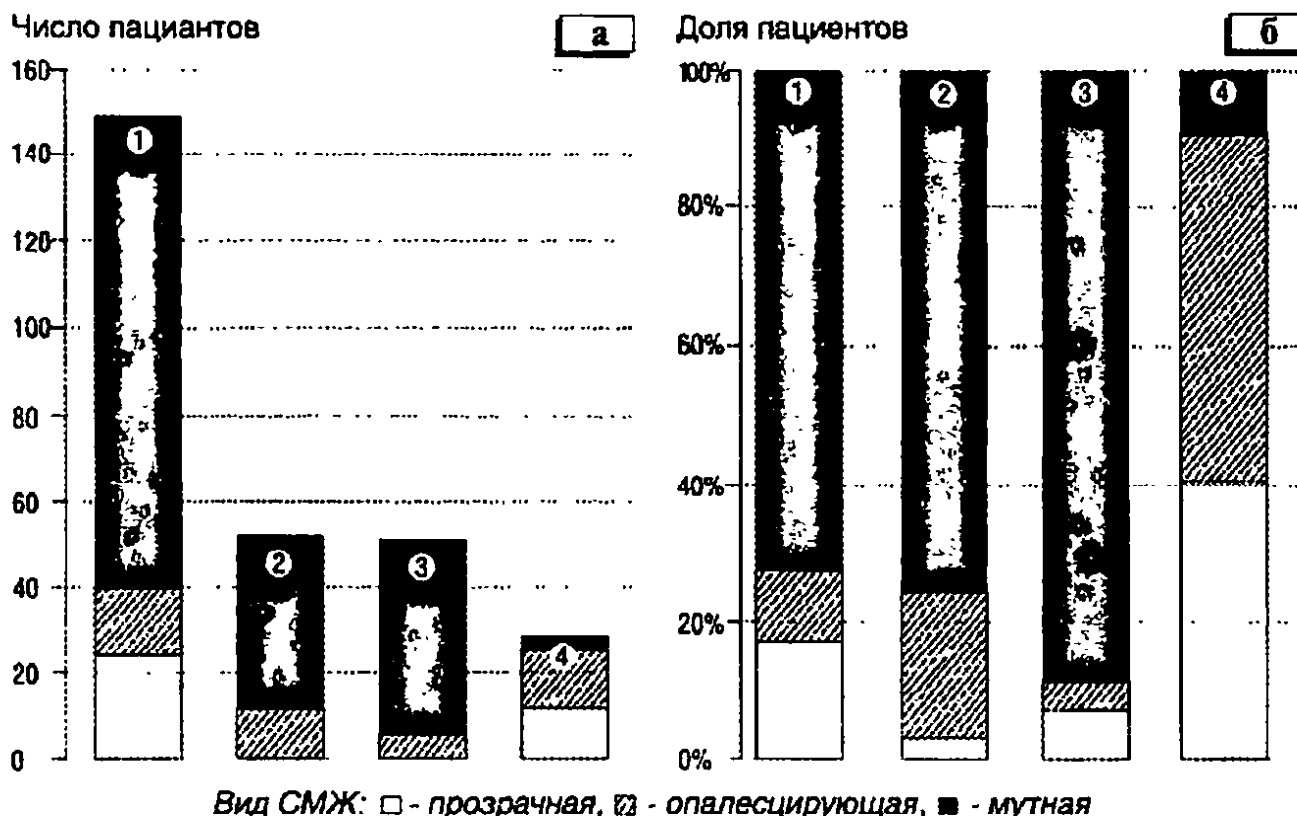


Рис. 8. Варианты графического представления результатов.

① — бактериальный менингит, вызванный *N. meningitidis*; ② — бактериальный менингит, вызванный *H. influenzae* типа b; ③ — бактериальный менингит, вызванный *S. pneumoniae*; ④ — серозный энтеровирусный менингит.

Вариант, представленный на рис. 8б, нагляднее показывает наиболее типичный вид спинномозговой жидкости (СМЖ) при разных патологиях: мутная СМЖ при бактериальном менингите, прозрачная или опалесцирующая СМЖ при вирусном менингите. Вариант, представленный на рис. 8а, несет дополнительную информацию о величине массива наблюдений.

Заболевание	Вид спинномозговой жидкости			
	прозрачный	опалесцирующий	мутный	всего
Менингококковый менингит	24 (16%)	16 (11%)	109 (73%)	149
Менингит, вызванный <i>H. influenzae</i> *	1 (2%)	11 (21%)	40 (77%)	52
Пневмококковый менингит**	3 (6%)	2 (4%)	46 (90%)	51
Энтеровирусный менингит***	11 (39%)	14 (50%)	3 (11%)	28

Текстовое представление той же информации выглядело бы приблизительно так:
 Достоверность наблюдаемых различий можно проверить по критерию χ^2 и представить как

примечание к таблице: * $p < 0.01$ для отличия от менингококкового менингита; ** $p < 0.05$ для отличия от менингококкового менингита или менингита, вызванного *H. influenzae*; *** $p < 0.001$ для отличия от менингококкового менингита, или пневмококкового менингита, или менингита, вызванного *H. influenzae*.

уровень значимости свидетельствует о наличии «статистически достоверного различия», есть на самом деле предмет условного соглашения. Обычно 5% вероятность ошибочного заключения ($p = 0.05$) считается достаточно малой и позволяющей считать утверждение «достоверным».

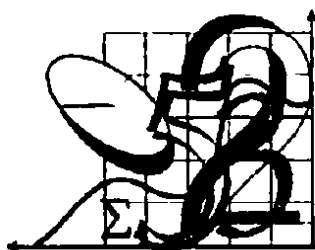
11. Не переоценивайте медико-биологический смысл статистически достоверных утверждений, например, утверждений о различии распределений. Изучив 10 000 больных с болезнью X и 10 000 больных с болезнью Y, можно обнаружить, что значение некоего параметра A равно 128 ± 35 при болезни X и 134 ± 36 при болезни Y и эта разница достоверна. Однако эти два распределения перекрываются почти до полной неотличимости, и на основании измерений параметра A нельзя дифференцировать заболевания X и Y у конкретного больного, рассуждать о различии патогенеза данных заболеваний или назначать разное лечение.

12. Представляя материал для публикации, дайте его описание, позволяющее читателю понять его полностью и без искажений: расшифруйте все обозначения; укажите смысл приведенных статистических характеристик (не просто « $X \pm x$ », но « $X \pm x$, где X — среднее арифметическое, а x — стандартное отклонение») и количество объектов в каждой группе; перечислите статистические процедуры и критерии, примененные для проверки конкретных утверждений; упомяните использованные компьютерные программы.

Стандартные статистические методы, к которым относятся большинство обсуждаемых в данной книге, не требуют ссылки на описывающие их источники литературы. Если вы подозреваете, что примененные вами статистические процедуры могут быть неизвестны большинству читателей, дайте их краткое описание в тексте ва-

шей публикации и ссылку на соответствующие оригинальные работы или монографии.

В целом – руководствуйтесь рациональной житейской логикой. Статистика – это инструмент, который должен помочь читателю самостоятельно оценить, что вы делали, зачем делали, какие получили результаты, какие на основе результатов сделаны утверждения и какова им цена.



Краткий список литературы

1. *Тюрин Ю.Н., Макаров А.А.* Статистический анализ данных на компьютере. - М.: Инфра М, 1998.
2. *Боровиков В.П., Боровиков И.П.* Statistica® - Статистический анализ и обработка данных в среде Windows®. - М.: ИИД "Филин", 1997.
3. *Руштон Р.* Справочник по непараметрической статистике. - М.: Финансы и статистика, 1982.
4. *Флетчер Р., Флетчер С., Вагнер Э.* Клиническая эпидемиология. Основы доказательной медицины. - М.: Медиа Сфера, 1998.
5. *Гланц С.* Медико-биологическая статистика. - М.: Практика, 1999.
6. *Справочник по прикладной статистике.* В 2 тт. / Под ред. Э.Ллойда, У.Ледермана, Ю.Н.Тюрина. - М.: Финансы и статистика, 1989, 1990.

Публикации, использованные для построения и обсуждения иллюстраций

1. *Платонов А.Е., Вершинина И.В., Кейти Х. и др.* Бактерицидное действие нейтрофилов человека на менингококки *in vitro*. // Журн. микробиол. - 1999. - № 3. - С. 46-52.
2. *Платонов А.Е., Троцанский Д.В., Белобородов В.Б. и др.* Связь тяжести течения менингококковой инфекции с уровнями эндотоксина и компонента в крови больных. // Клин. мед. - 1999. - № 2. - С. 32-37.
3. *Платонов А.Е., Троцанский Д.В., Белобородов В.Б. и др.* Прогнозирование исхода менингококковой инфекции: методология и практика // Клин. мед. - 1999. - № 4. - С. 32-38.

4. *Platonov A.E., Beloborodov V.B., Pavlova L.I. et al.* Vaccination of patients deficient in a late complement component with tetravalent meningococcal capsular polysaccharide vaccine // *Clin. Exp. Immunol.* - 1995. - Vol. 100. - P. 32-39.
5. *Platonov A.E., Vershinina I.V., Dankert J. et al.* Long-term follow-up of late complement component deficient patients vaccinated with meningococcal polysaccharide vaccine: Antibody persistence and efficacy of vaccination // *Pathogenic Neisseria* / Eds. W.D. Zollinger *et al.* - Baltimore, 1996. - P. 235-236.