Medical Statistics at a Glance

Aviva Petrie

Head of Biostatistics Unit and Senior Lecturer
Eastman Dental Institute
University College London
256 Grays Inn Road
London WC1X 8LD and
Honorary Lecturer in Medical Statistics
Medical Statistics Unit
London School of Hygiene and Tropical Medicine
Keppel Street
London WC1E 7HT

Caroline Sabin

Professor of Medical Statistics and Epidemiology
Department of Primary Care and Population Sciences
Royal Free and University College Medical School
Rowland Hill Street
London NW3 2PF

Second edition

Blackwell
Publishing

Авива Петри, Кэролайн Сэбин

Наглядная МЕДИЦИНСКАЯ СТАТИСТИКА

Перевод с английского
под редакцией В.П. Леонова

2-е издание, переработанное и дополненное

Москва
Издательская группа «ГЭОТАР-Медиа»
2009

УДК 615.036.2=111=03.161.1(075.8)
ББК 51.1я73
П30

Рекомендовано Учебно-методическим объединением по медицинскому и фармацевтическому образованию вузов России в качестве учебного пособия для студентов медицинских вузов

Рецензенты:

Власов Василий Викторович — д-р мед. наук, проф.
Комарова Марина Валериевна — канд. биол. наук, доцент кафедры радиотехники и медицинских диагностических систем Самарского государственного аэрокосмического университета

Петри А., Сэбин К.

П30 Наглядная медицинская статистика / А. Петри, К. Сэбин ; пер. с англ. под ред. В.П. Леонова. — 2-е изд., перераб. и доп. — М. : ГЭОТАР-Медиа, 2009. — 168 с. : ил.

ISBN 978-5-9704-0914-5

Цель второго издания книги, как и первого, — донести до читателя основные понятия и принципы медицинской статистики, которые довольно широко применяют зарубежные медики и биологи. Книга содержит не только необходимую теоретическую часть, но и в доступной форме даёт практическое описание того, как можно применять статистические методы в реальных клинических исследованиях. Для освоения изложенного материала читателю вполне достаточно знаний по математике в объёме школьного курса.

Учебное пособие предназначено для студентов и аспирантов медицинских вузов, биологических факультетов университетов, врачей, исследователей-клиницистов и всех тех, кто интересуется применением статистики в медицине и биологии.

УДК 615.036.2=111=03.161.1(075.8)
ББК 51.1я73

ISBN 978-5-9704-0914-5

Содержание

Предисловие редактора к изданию на русском языке

Статистика — часть математики. Название этой отрасли знания происходит от греческого слова «матейн» (*mathein*) — учиться, познавать. Так что с полным основанием всех, кто в древнем мире изучал основы медицины, можно было называть математиками. Более того, древние греки считали науку, познание (*mathema*) и математику (*mathematike*) синонимами. Если основной метод математики — абстрагирование, то в медицине, наоборот, лечение конкретного пациента требует учёта частностей, деталей, поскольку «лечат не болезнь, а больного». Именно здесь наиболее чётко проявляется «водораздел» между медициной и математикой. Не случайно многие, кто поступает в медицинские вузы, не в ладах с математикой. И ничего плохого в этом нет. Это лишь свидетельствует о разном складе ума у людей и о востребованности разных систем мышления в обществе. В предисловии авторы говорят, что «Наглядная медицинская статистика» «...предназначена для медиков-интернов, медицинских исследователей, аспирантов биомедицинских дисциплин и персонала фармацевтической промышленности». Общеизвестно, что статистика — прекрасный инструмент, используемый при обобщении множества наблюдений. Ясно и то, что студенты II–III курса медицинского вуза накопить такие наблюдения ещё не в состоянии, поэтому-то не случайно авторы делают акцент именно на этой аудитории.

В 2006 г. при подготовке тематического номера «Международного журнала медицинской практики» посвящённого обучению медиков статистике, мы проанализировали анкету, содержавшую 26 вопросов [91]. Анализ результатов анкетирования показал, что статистический инструментарий прежде всего необходим исследователям, занятым научной работой. Современная статистическая наука и технология по своей сложности и трудности не уступает сложности и трудности разнообразных медицинских специальностей, поэтому учёному в области медицины и биологии необходимо освоить статистические технологии на уровне, достаточном для того, чтобы адекватно понимать использованный метод и в дальнейшем интерпретировать полученные результаты. Такой же точки зрения придерживаются и авторы этой книги, давая лишь основные идеи, несложные формулы и примеры применения метода. В этом основное достоинство книги, которая требует от читателя, чтобы он знал математику лишь на уровне школьного курса.

Как зарубежный, так и отечественный опыт показывает, что можно хорошо овладеть этим инструментом, если упражняться в его использовании на собственных наблюдениях. В справедливости такого утверждения автор этих строк неоднократно убеждался при подготовке и проведении выездных семинаров по биометрике. Слушатели таких семинаров обучаются на результатах предварительно выполненного нами статистического анализа их данных, что позволяет им выработать на семинаре навык адекватной интерпретации результатов такого анализа. Осенью 2007 г., занимаясь подготовкой такого семинара в Красноярске, я впервые столкнулся с тем, что организаторы семинара за полгода его подготовки не смогли собрать собственные данные для предварительного анализа. В результате профессор Ш., занимавшаяся организацией семинара, предложила обучать слушателей на чужих примерах. Очевидно, что эффективность такого семинара будет сродни результатам обучения гинеколога с использованием стоматологического кресла. Не случайно слушатели этого семинара, состоявшегося в конце января — начале февраля 2008 г., в анкетах, приведённых на сайте БИОМЕТРИКА, единодушно отметили необходимость использовать собственные данные для успешного обучения, поэтому читателю, работающему с этой книгой, следует, изучая тот или иной метод, сразу же пытаться применить его к собственным данным. Те же, у кого таких данных не будет под рукой, могут найти близкие к их специальности в библиотеке RusDASL (российская библиотека данных для изучающих биостатистику) [136], где в формате EXCEL представлены материалы из различных областей медицины и биологии.

В последние 50 лет статистику в медицине часто использовали весьма некорректно, иногда лишь как средство «онаучивания» полученных результатов исследования. Так, в 1955 г. в книге А.Я. Боярского «Статистические методы в экспериментальных медицинских исследованиях» говорилось следующее: «Уже беглое ознакомление с состоянием дела показывает, что статистическая обработка экспериментальных данных является наиболее слабым местом во многих исследованиях. ... Трудно требовать от медика, чтобы он,

наряду со знаниями в своей области, был в то же время достаточно компетентен, скажем, в радиотехнике для конструирования аппаратуры, улавливающей биотоки, или в статистике для нахождения наиболее правильных методов статистической обработки своих экспериментальных данных. И подобно тому, как медику, несомненно, приходится обращаться за содействием к радиотехнику, для правильной статистической обработки экспериментальных данных нередко приходится обращаться к специалисту-статистику. Так или иначе, но бесспорным фактом являются и недостаточная вооружённость медиков статистическими знаниями, и недостаточно высокий научный уровень статистической методики в большинстве их экспериментальных работ».

Проведённый нами анализ огромного количества статей, книг и диссертаций по медицине и биологии показал, что и спустя 50 лет мало что изменилось. По-прежнему более половины публикаций содержит многочисленные ошибки использования статистики, которые делают сомнительной и бездоказательной значительную часть последующих умозаключений авторов. С анализом наиболее типичных ошибок таких публикаций читатели могут познакомиться в литературе [88]. Причины устойчивости этого явления разнообразны [92]. Во-первых, большинство авторов — не профессиональные исследователи и в силу этого они не владеют необходимыми знаниями в области статистики. Во-вторых, нередко деформирована сама цель публикации. Она подчас воспринимается как средство априорного доказательства «научности» результатов, невзирая при этом на надёжность и доказательность выводов. Этому же способствует и то, что возможные негативные последствия опубликования сомнительных результатов отдалены от автора во времени и пространстве. Играет свою роль и то, что доступ широкой читательской аудитории как к самим публикациям, так и к исходным авторским данным ограничен. Не меньший вклад в это вносит и незаинтересованность редакций, выпускающих периодические издания, ректоров медицинских вузов, диссертационных советов, ВАК РФ и дирекций НИИ в повышении качества журнальных публикаций и диссертаций. Очевидно, что для изменения такого положения необходимы последовательные, но в то же время и радикальные меры.

Как известно, с середины 2006 г. диссертанты обязаны публиковать в Интернете авторефераты своих диссертаций. А в скором времени будут публиковать до защиты и всю диссертацию. Именно об этом заявил 31 октября 2007 г. первый вице-премьер Д.А. Медведев на встрече с членами ВАК и ректорами вузов, поддержав наши давние предложения, на что председатель ВАК декан биологического факультета МГУ академик Михаил Кирпичников ответил, что «в ближайшее время мы будем готовы говорить о публикации полностью диссертаций». Очевидно, что в связи со всем сказанным актуальность этого и других подобных изданий, знакомящих медиков с основами биостатистики, будет только возрастать. Этому же способствуют и возросшая доступность многочисленных статистических пакетов для исследователей, а также возможность дистанционно обучаться работе на них (http://www.biometrica.tomsk.ru/edu_1.htm). Востребованность настоящего издания обусловлена и тем, что усиливается интерес к концепции доказательной медицины, в которой статистика — один из основных инструментов.

Как и следовало ожидать, второе издание книги содержит некоторые дополнения, о которых авторы сами говорят в предисловии. Можно предполагать, что новые главы появились как результат освоения авторами новых методов, которые они посчитали актуальными и полезными. Так как в обоих изданиях отсутствуют списки рекомендуемой литературы по биостатистике, мы дополнили второе издание весьма обширной библиографией. Более полный перечень книг приведён по адресам: http://www.biometrica.tomsk.ru/razdel_2_1.htm — http://www.biometrica.tomsk.ru/razdel_2_12.htm.

Поскольку отечественная терминология в этой области ещё только складывается, возможно, некоторые термины могут показаться читателям непривычными. В перевод этого издания мы добавили некоторые примечания, конкретизирующие ряд деталей. Несомненное достоинство издания «Study of Case» — обучение на примерах, наличие обширного глоссария и детального предметного указателя. Всё это позволяет утверждать, что книга найдёт своего читателя и позволит повысить уровень знаний о статистике многим специалистам в области медицины.

Канд. тех. наук, доцент факультета информатики
Томского государственного университета,
редактор сайта БИОМЕТРИКА,
http://www.biometrica.tomsk.ru

В.П. Леонов

Предисловие

«Наглядная медицинская статистика» (*Medical Statistics at a Glance*) предназначена для медиков-интернов, медицинских исследователей, аспирантов биомедицинских дисциплин и персонала фармацевтической промышленности. Все эти лица в профессиональной практике неизбежно столкнутся с количественными результатами (своими собственными или чужими), которые потребуют критической оценки и интерпретации, а некоторые из них, конечно, будут вынуждены пройти экзамен по этой наводящей ужас статистике! Надлежащее понимание статистических концепций и методологии неоценимо для этих целей. Будучи прагматиками, мы хотели бы зажечь читателя энтузиазмом в области статистики. Наша цель — дать студенту и исследователю, как и клиницистам, которые сталкиваются со статистическими концепциями в медицинской литературе, книгу, которая логична, легко читается, исчерпывающа и имеет практическое применение.

Мы полагаем, что «Наглядная медицинская статистика» будет особенно полезна как дополнительный материал к лекциям по статистике и как руководство с соответствующими ссылками. Вместе с другими книгами серии «At a Glance» мы ведём читателя через ряд самостоятельных двух- и трёхстраничных тем, каждая из которых охватывает отдельный аспект медицинской статистики. Из опыта обучения мы выяснили и приняли во внимание трудности, с которыми столкнулись наши студенты при изучении медицинской статистики. По этой причине мы предпочли ограничить теоретическое содержание книги уровнем, достаточным для понимания включённых в неё процедур, но всё-таки не затмевающим практические аспекты их выполнения.

Медицинская статистика — предмет, имеющий широкий диапазон и включающий множество разделов. Мы даём элементарное введение в основополагающие концепции медицинской статистики и руководство по наиболее часто применяемым статистическим процедурам. Эпидемиология тесно связана с медицинской статистикой, поэтому обсуждаются некоторые основные вопросы, относящиеся к разработке исследования и его интерпретации. Включены также темы, которые могут быть полезны читателю только изредка, но которые, тем не менее, необходимы во многих областях медицинских исследований: например, доказательная медицина, систематические обзоры и метаанализ, анализ временных рядов, анализ выживаемости и байесовские методы. Мы объяснили принципы, лежащие в основе этих тем, так, чтобы читатель смог понять и интерпретировать результаты, когда они встречаются в литературе.

Порядок первых 30 глав этого издания не отличается от первого издания. Большинство из них остались неизменёнными и в новом издании, некоторые же содержат незначительные отличия, представляющие собой последние результаты и взаимные ссылки, или преобразованы с учётом нового материала. Главные поправки касаются сравнительно сложных форм регрессионного анализа, которые сейчас используются гораздо шире, чем тогда, когда был написан первый вариант книги, частично потому, что соответствующее программное обеспечение стало доступнее и эффективнее, чем прежде. Мы изменили главу, посвящённую бинарным откликам (исходам) и логистической регрессии, включили новую главу об интенсивности и пуассоновской регрессии и значительно расширили первоначальную тему по статистическому моделированию, так что теперь она включает три главы: «Общие линейные модели», «Объясняющие переменные в статистических моделях» и «Актуальные вопросы в статистическом моделировании». Мы также изменили главу 41, которая описывает различные подходы к анализу сгруппированных данных, и добавили главу 42, выделив различные методы регрессии, которые можно использовать для анализа этого типа данных. Первое издание содержало краткое описание анализа временных рядов, которое мы решили не включать во второе издание, поскольку чувствовали, что оно, вероятно, было весьма ограниченным, чтобы его практически использовать, а его расширение отодвинуло бы сроки издания. Из-за этого пропуска и некоторых дополнений нумерация глав, начиная с главы 31, в первом и во втором издании различается. Большинство из них по сравнению с первым изданием либо слегка изменены, либо остались неизменёнными.

Описание каждой статистической технологии сопровождается примером, иллюстрирующим, как её использовать. Данные для примеров мы взяли из совместных исследований, в которых участвовали мы или наши коллеги, а в отдельных случаях — из ранее опубликованных статей. Где возможно, мы использовали один и тот же набор данных, чтобы приблизить к реальности технологию анализа, которая редко ограничивается единственным методом или подходом. Хотя предполагается, что формулы должны подтверждать логику подхода, помогая объяснению и пониманию, мы старались не приводить детально сложные вычисления: большинство читателей имеют доступ к компьютеру и вряд ли станут выполнять любые, даже самые простые вычисления вручную.

Мы полагаем, что для читателя особенно важно умение интерпретировать результаты вычислений на компьютере, поэтому предпочли там, где это возможно, иллюстрировать результаты фрагментами компьютерных вычислений. В отдельных случаях, когда с такой интерпретацией могут возникнуть трудности, мы включали (приложение В) и аннотированные компьютерные распечатки (листинги), содержащие анализ набора данных. Есть много статистических пакетов общего применения; для того чтобы показать читателю, как эти распечатки могут видоизменяться, мы не ограничились подробным разбором какого-либо одного конкретного пакета, а использовали три хорошо известных статистических пакета: SAS, SPSS и STATA.

В тексте книги есть обширные перекрёстные ссылки, чтобы помочь читателю связать различные процедуры. Основной набор статистических таблиц содержится в приложении А (Neave H.R. Elementary Statistical Tables Routledge, 1981, и Diem K. Documenta Geigy Scientific Tables. — 7th edn. — Oxford: Blackwell Publishing, 1970), которые среди прочих дают более полные версии, если читателю потребуется точнее рассчитать результаты вручную. Словарь терминов (приложение Г) содержит легкодоступные объяснения часто применяемой терминологии.

Известно, что одна из трудностей, с которыми сталкиваются нестатистики, — выбор подходящего метода, поэтому мы создали две схемы, которые можно использовать и для того, чтобы выбрать метод в данной ситуации, так и для лёгкого поиска этого метода. Они воспроизведены на внутренней стороне обложки.

Читателю могут быть полезными диалоговые упражнения на веб-сайте (http://www.medstatsaag.com/). Он содержит полный набор ссылок (некоторые из них непосредственно связаны с базой данных Medline), что позволяет расширить круг ссылок, указанных в книге, и обеспечить себя дополнительной полезной информацией для примеров. Читателям, желающим более подробно изучить специфические области медицинской статистики, мы рекомендуем следующие книги.

- Altman D.G. Practical Statistics for Medical Research. — London: Chapman and Hall, 1991.
- Armitage P., Berry G., Matthews J.F.N. Statistical Methods in Medical Research. — 4th edn. — Oxford: Blackwell Science, 2001.
- Pocock S.J. Clinical Trials: A Practical Approach. — Chichester: Wiley, 1983.

Мы чрезвычайно благодарны Марку Гилторпу и Джонатану Стерну за ценные комментарии и предложения, касающиеся ряда аспектов во втором издании, а также Ричарду Моррису, Фионе Ламп, Шаку Хаджату и Абулу Басару за обсуждение первого издания. Мы благодарим всех, кто помогал нам искать данные для примеров. Естественно, мы несём полную ответственность за любые ошибки в тексте или примерах. Мы также благодарим Майка, Джеральда, Нину, Эндрю и Карен, которые терпеливо, хладнокровно переносили нашу увлечённость, работой над первым изданием и пережили вместе с нами все испытания и несчастья, сопровождавшие издание второго.

Авива Петри,
Каролина Сэбин,
Лондон

Список сокращений

ВКК — внутриклассовый коэффициент корреляции
ДИ — доверительный интервал
ДМ — доказательная медицина
ИБС — ишемическая болезнь сердца
ИМ — инфаркт миокарда
МНК — метод наименьших квадратов
ОЛМ — обобщённая линейная модель
ОМП — оценка максимального правдоподобия
ОНК — оценка наименьших квадратов
ООУ — обобщённые оценки уравнения
ОР— относительный риск
ППК — площадь под кривой
ПРК — первичное послеродовое кровотечение
РКИ — рандомизированные контролируемые испытания
СМП — статистика отношения правдоподобия
СПИД — синдром приобретённого иммунодефицита человека
УДВВ — уровень давления в воротной вене
ЦМВ — цитомегаловирусная инфекция
ЧБНЛ — число больных, которое вам необходимо лечить
ШФА — шкала физической активности
CABG — операция шунтирования коронарной артерии
CMV — цитомегаловирус
df — «степени свободы»
HAART — высокоактивное антиретровирусное лечение
HHV-8 — вирус герпеса человека
HRT — терапия замены гормона в постменопаузе
FEV_1 — объём форсированного выдоха в 1 с
IVF — экстракорпоральное оплодотворение
LR — отношение правдоподобия
ROC — кривая операционной характеристики диагностического теста
SD — стандартное отклонение
SEM — стандартная ошибка среднего

1 Типы данных

ДАННЫЕ И СТАТИСТИКА[1]

Цель большинства исследований состоит в сборе данных, которые впоследствии помогают получить информацию относительно какой-либо области исследования. Данные всегда основаны на наблюдениях одной или нескольких переменных; термин переменная означает количественный показатель, способный изменяться. Например, мы можем собрать основную клиническую и демографическую информацию о пациентах со специфической болезнью. Интересующими нас переменными могут быть пол, возраст и рост больного.

Обычно мы получаем данные из выборки индивидуумов, представляющих популяцию — группу индивидуумов, которая представляет для нас интерес. Цель состоит в том, чтобы сгруппировать эти данные и извлечь полезную информацию. Статистика использует различные методы, например, сбор данных, их обобщение, анализ и подведение итогов, основанных на полученных сведениях.

Существуют различные формы данных. Прежде чем решить, какой статистический метод окажется наиболее подходящим для конкретного случая, мы должны знать, к какому типу данных относится каждая переменная. Все переменные, результирующие показатели, можно разделить на два типа: категориальный (качественный) или числовой (количественный) (рис. 1-1).

Рис. 1-1. Различные типы переменных (схема).

КАТЕГОРИАЛЬНЫЕ (КАЧЕСТВЕННЫЕ) ДАННЫЕ

Встречаются тогда, когда индивидуум может принадлежать только к одной из множества категорий переменных.

- Номинальные данные. Категории не упорядочиваются, а просто имеют названия. Например, группа крови (A, B, AB и O) и семейное положение (замужем, вдова, не замужем и т.д.). В этом случае нет оснований полагать, что быть замужем лучше (или хуже), чем быть не замужем!
- Ординальные (ранговые, порядковые) данные. В некоторых случаях категории (градации, уровни) упорядочиваются. Например, стадии болезни (заболевание в запущенной стадии, средняя, лёгкая форма болезни или её отсутствие) и степень боли (сильная, умеренная, слабая, отсутствие боли).

Категориальная (качественная) переменная — бинарная или дихотомическая переменная, когда имеются только две возможные категории. Например, «да/нет», «умер/жив» или «пациент болен/пациент здоров».

ЧИСЛОВЫЕ (КОЛИЧЕСТВЕННЫЕ) ДАННЫЕ

Встречаются, когда переменная имеет некоторую числовую величину (значение). Мы можем подразделить числовые данные на два типа.

- Дискретные данные. Переменная может принимать только определённые числовые значения. Часто ведётся подсчёт количества событий, например, количество посещений врача в год или количество заболеваний человека за последние пять лет.
- Непрерывные данные. Нет никаких ограничений в отношении данных, которые переменная может принимать, например, масса тела или рост.

РАЗЛИЧИЕ МЕЖДУ ТИПАМИ ДАННЫХ

В зависимости от того, оказываются ли данные категориальными или числовыми, используют различные статистические методы. Казалось бы, различия между категориальными и числовыми данными и так понятны, но в некоторых случаях отличие становится не совсем ясным. Например, когда у нас есть переменная с множеством установленных категорий (боль может иметь семь категорий), могут возникнуть трудности, как отличить её от дискретной числовой переменной. Различие между дискретными и непрерывными числовыми данными может быть даже менее понятным, хотя в общем на результатах большинства исследований это не отразится. Возраст — пример переменной, которую часто трактуют как дискретную, хотя её нужно отнести к непрерывным. Обычно мы ссылаемся на «возраст в последний день рождения», нежели на «возраст по состоянию на сегодняшний день», и поэтому женщине, которая сообщает, что ей 30, может быть 30 лет и несколько дней или почти 31 год.

Не торопитесь вначале записывать числовые данные как категориальные, поскольку часто теряется важная информация. Гораздо проще преобразовать числовые данные в категориальные данные сразу же, как только они будут собраны.

[1] Детальное изложение особенностей различных типов данных [127]. Прим. переводчика.

ПРОИЗВОДНЫЕ (ВТОРИЧНЫЕ) ДАННЫЕ

Существует множество других типов данных в области медицины. Они могут включать в себя следующее.

- Проценты. Могут возникать при рассмотрении вопроса относительно улучшения состояния больного во время лечения, например состояние больного (объём форсированного выдоха в 1 с, FEV_1) может улучшиться на 24% после лечения новым препаратом. В этом случае имеет значение степень улучшения, а не абсолютные данные о состоянии пациента.
- Пропорции или отношения. Иногда встречается два варианта пропорций или отношений. Например, индекс массы тела (индекс Кетле): отношение массы тела индивидуума (кг) к квадрату роста, таким образом проводится оценка, превышает ли масса норму или, наоборот, недостаточна.
- Интенсивность. Относительная частота заболеваний, где количество заболеваний делят на общее число лет, в течение которых вели наблюдения за пациентами в этом исследовании (см. главу 31), общепринята при эпидемиологическом исследовании (см. главу 12).
- Метки, оценки. Произвольные значения, или метки, используют в том случае, когда невозможно измерить количество. Например, ряд вопросов на ответы относительно качества жизни можно суммировать для того, чтобы дать полную оценку относительно качества жизни у каждого индивидуума[1].

Все эти переменные можно рассматривать в большинстве исследований как непрерывные переменные. В данном случае переменную можно будет сконструировать, если использовать более чем одну величину (например, числитель и знаменатель для процента), важно при этом регистрировать все используемые значения. Например, состояние больного улучшилось на 10% после лечения, данное улучшение может иметь различную клиническую значимость в зависимости от того, в каком состоянии находился больной до лечения.

ЦЕНЗУРИРОВАННЫЕ ДАННЫЕ

Мы можем рассмотреть цензурированные данные на следующих примерах.

- Если мы проводим лабораторные измерения, используя прибор, который может обнаружить значения только выше некоторого предельного уровня, тогда любая величина ниже этого уровня не будет обнаружена. Например, вирус, уровень обнаружения которого ниже предела, часто рассматривается как «необнаруженный», при том что на самом деле он может находиться в образце.
- Мы можем столкнуться с цензурированными данными, например, когда некоторые больные из группы исследуемых отстраняются от испытания до окончания исследований. Этот тип данных подробно обсуждается в главе 44.

[1] Такие числовые метки могут быть получены при анализе наборов качественных и количественных признаков. Многие из них — результаты применения ряда многомерных методов, таких, как дискриминантный анализ, корреспондентский анализ и т.д., не рассмотренных в этой книге. Автор [49] даже ввёл новый термин «патометрия» для обозначения технологии получения количественных меток, характеризующих степень патологии. Прим. переводчика.

2 Ввод данных

При проведении какого-либо исследования вам почти всегда необходимо будет вводить данные в компьютерный пакет прикладных программ. Компьютер — неоценимая вещь, с его помощью можно проверить правильность данных, ускорить сбор данных и анализа, гораздо проще проверять ошибки, проводить графические подсчёты данных и новые переменные. Стоит потратить некоторое время на планирование ввода данных — и на последней стадии это сэкономит ваше время и усилия.

ФОРМАТЫ ДЛЯ ВВОДА ДАННЫХ

Существует несколько способов ввода данных и сохранения их в компьютере. Большинство статистических пакетов позволяют сразу же вводить данные. Однако существуют ограничения, а именно: вы не сможете переносить данные из одного пакета в другой. Простейшая альтернатива — сохранять данные либо в электронной таблице, либо в пакете баз данных. К сожалению, их статистические процедуры часто ограничены, и обычно возникает необходимость вводить данные в статистический пакет, чтобы провести исследования.

Наиболее гибкий подход состоит в том, чтобы сохранять ваши данные как ASCII (American Standard Code for Information Interchange — стандартный код информационного обмена США) или в текстовом файле. Данные в ASCII-формате могут читаться большинством пакетов. Формат ASCII состоит просто из текста, который вы можете читать с компьютера. Обычно каждая переменная в файле отделяется от следующей каким-нибудь разделителем, часто пробелом или запятой. Такой формат известен как свободный формат.

Самый простой способ ввода данных в ASCII-формате — печатать данные прямо в этом формате, используя текстовый редактор либо иной редакторский пакет. В качестве альтернативы данные, находящиеся в пакете электронных таблиц (Excel), могут быть сохранены в текстовом формате[1].

Используя любой подход, при исследовании принято, чтобы каждая строка данных соответствовала отдельному индивидууму и каждая колонка соответствовала переменной, хотя может возникнуть необходимость в продолжении последовательных рядов, в случае если на каждого индивидуума собрано большое количество переменных.

ПЛАНИРОВАНИЕ ВВОДА ДАННЫХ

При сборе данных вам необходимо будет использовать форму или анкету для их фиксации. Если они хорошо разработаны, то могут сократить работу, которую необходимо выполнить при вводе данных. В общем, эти формы/анкеты включают ряд «ячеек», в которые заносятся данные — обычно имеется отдельная «ячейка» для каждого возможного числового ответа.

[1] Одна из наиболее часто встречающихся ошибок начинающих исследователей заключается в том, что они вводят данные в таблицы EXCEL в текстовом формате и затем при импорте этих таблиц в какой-либо статистический пакет, например пакет STATISTICA или SPSS, испытывают сложности при попытке анализа этих данных как числовых. Прим. переводчика.

КАТЕГОРИАЛЬНЫЕ ДАННЫЕ

С нечисловыми данными могут возникнуть проблемы при занесении их в некоторые статистические пакеты, поэтому вам необходимо назначить числовые коды категориальным данным, прежде чем вводить данные в компьютер. Например, вы можете выбрать следующие коды: 1, 2, 3 и 4 категориям «нет боли», «лёгкая боль», «средняя боль» и «сильная боль» соответственно. Эти коды могут быть добавлены к формам при сборе данных. Для бинарных данных, например, ответов «да/нет», очень удобно установить код 1 (например, для «да») и 0 (для «нет»).

- Переменные с единственным альтернативным вариантом ответа. Существует только один возможный ответ на вопрос. Например, на вопрос «Умер ли пациент?» невозможно ответить и «да», и «нет».
- Переменные с несколькими альтернативами ответа. Возможен более чем один ответ. Например, на вопрос: «Каковы симптомы болезни у пациента?» — можно перечислить несколько симптомов. Существует два способа обработки этих данных в зависимости от того, какую из двух следующих ситуаций применить.
  - ✧ Существует несколько возможных симптомов и многие из них присутствуют у пациента. Можно создать ряд различных бинарных переменных, все зависит от того, ответит ли больной «да» или «нет» относительно возможных симптомов. Например, был ли у него кашель, болело ли горло.
  - ✧ Существует огромное количество возможных симптомов, но у пациента могут быть только некоторые из них. Можно создать ряд различных номинальных переменных, каждая из которых позволит определить наличие того или иного симптома у больного. Например, какой симптом возник первым, какой — вторым и т.д. Вы заранее должны определить максимальное количество симптомов, которые, вы полагаете, могут быть у больного.

ЧИСЛОВЫЕ ДАННЫЕ

Числовые данные должны быть введены с той же самой точностью, с которой были проведены измерения, и единица измерения должна быть едина для всех наблюдений данной переменной. Например, масса должна быть записана в килограммах или в граммах, но не попеременно то в килограммах, то в граммах.

МНОЖЕСТВЕННЫЕ ФОРМЫ НА ОДНОГО БОЛЬНОГО

Иногда информацию собирают на одного и того же больного более чем в одном случае (наблюдении). Важно отметить, что должен существовать уникальный идентификатор (например, порядковый номер), принадлежащий только одному человеку в данном наблюдении, который предоставит вам возможность соединить все данные, собранные на одного человека при исследовании.

ПРОБЛЕМЫ С ДАТАМИ И ПЕРИОДАМИ

Даты и периоды необходимо вводить последовательно, например: либо день/месяц/год, либо месяц/день/год, но всегда в одном и том же порядке. Важно установить, какой формат может читаться в данном статистическом пакете.

КОДИРОВАНИЕ ОТСУТСТВУЮЩИХ (ПРОПУЩЕННЫХ) ДАННЫХ

Вам следует определить, что вы будете делать с отсутствующими данными, прежде чем вводить информацию. В большинстве случаев вы будете вынуждены использовать какой-нибудь символ для недостающих данных. Статистические пакеты предлагают для этого различные способы. Некоторые пакеты используют специальные символы (например, точку или звездочку) для обозначения пропущенных данных, принимая это во внимание во время анализа, другие требуют от вас ввести свой собственный код для обозначения отсутствующих данных (обычно используют знаки 9, 999 или −9999). Выбранное значение должно быть одно для всех переменных, и его невозможно использовать для другой переменной. Например, при вводе категориальной переменной с четырьмя категориями (коды 1, 2, 3 и 4) вы можете выбрать цифру 9 для недостающих данных. Однако для переменной «возраст ребёнка» будет необходимо выбрать другой код, например, «−9». Более подробно отсутствующие данные рассмотрены в главе 3.

ПРИМЕР

Номинальные переменные – не упорядоченные категории

Дискретная переменная – может принимать значения в интервале

Многоуровневая переменная – использовано создание 4 бинарных (дихотомических) переменных

Ошибки в ответах – некоторые заполнены в килограммах, другие – в фунтах и унциях

Непрерывная переменная

Номинальная переменная

Ординальная переменная

				Вмешательства, требуемые в течение беременности					Масса тела ребёнка			Дата			
№ пациента	Кровотечение	Пол ребёнка	Длительность беременности (недель)	Ингаляции	Внутримышечная инъекция	Внутривенная инъекция	Эпидуральное	Apgar Метка	кг	фунты	унции	Дата рождения	Возраст матери на момент рождения ребёнка	Группа крови	Частота кровотечений из дёсен
47	3	3	.	.	.	.	.	.	.	.	.	08/08/74	.	3	6
33	3	.	41	0	1	0	1	.	.	6	13	11/08/52	27,26	1	4
34	3	1	39	1	0	0	0	.	.	7	14	04/02/53	22,12	1	1
43	3	1	41	1	1	0	0	.	.	8	0	26/02/54	27,51	3	33
23	3	2	.	0	0	0	0	10/1-10/	11,19	.	.	29/12/65	36,58	1	3
49	3	3	.	.	.	.	.	.	.	.	.	09/08/57	.	1	5
51	3	3	.	.	.	.	.	.	.	.	.	21/06/51	.	3	5
20	2	41	0	1	0	0	.	.	7	12	15/08/96	25,61	3	3	.
64	4	.	.	1	1	0	0	.	.	.	.	10/11/51	24,61	3	2
27	3	1	14	1	0	0	0	ok	.	8	8	02/12/71	22,45	1	1
38	3	2	38	1	0	0	0	9/1-9/5	, .	6	10	12/11/61	31,60	1	1
50	3	2	40	0	0	0	0	.	.	5	11	06/02/68	18,75	1	6
54	4	1	41	0	1	0	0	.	.	7	4	17/10/59	24,62	3	2
7	1	1	40	0	0	0	1	.	.	6	5	17/12/65	20,35	2	6
9	1	2	38	0	1	0	0	.	.	5	4	12/12/96	28,49	3	3
17	1	4	.	.	.	.	.	.	.	.	.	15/05/71	26,81	1	5
53	3	2	40	0	0	1	0	.	.	8	7	07/03/41	31,04	1	3
56	4	2	40	0	0	0	0	.	3,5	.	0	16/11/57	37,86	3	3
58	4	1	40	0	1	0	1	.	.	8	0	17/063/47	22,32	3	Y
14	1	1	38	0	0	0	1	.	.	7	12	04/05/61	19,12	4	2

1=Гемофилия A
2=Гемофилия B
3=Болезнь Willebrand's
4=FXI-дефицт

1=Мальчик
2=Девочка
3=Аборт
4=Продолжающаяся беременность

0=Нет
1=Да

1=0+ve
2=0–ve
3=A+ve
4=A–ve
5=B+ve
6=B–ve
7=AB+ve
8=AB–ve

1=Более 1 раза в день
2=Один раз в день
3=Один раз в неделю
4=Один раз в месяц
5=Изредка
6=Никогда

Рис. 2-1. Часть электронной таблицы, в которой показаны собранные данные на примерах 64 женщин с наследственными беспорядочными кровотечениями.

Эта часть исследования показывает, как влияют наследственные беспорядочные кровотечения на беременность и роды. Данные были собраны при исследовании 64 женщин, зарегистрированных в одном и том же Центре гемофилии в Лондоне. Женщин опрашивали относительно их кровотечений и их первой беременности (или их текущей беременности, если они были беременны в первый раз на день интервьюирования). На рис. 2-1 представлены данные, которые были собраны при исследовании небольшого количества женщин после того, как данные были введены в электронную таблицу, но до того, как их проверили на наличие ошибок. Внизу на рис. 2-1 приведена кодовая схема для категориальных переменных. Каждый ряд отведён для отдельного пациента; каждая колонка для отдельной переменной. В случае если женщина всё ещё беременна, возраст женщины высчитывался на день рождения младенца. Данные, касающиеся живорождённых детей, описаны в главе 37.

Данные любезно предоставлены Dr R.A. Kadir, University Department of Obstetrics and Gynaecology, and Professor C.A. Lee, Haemophilia Centre and Haemostasis Unit, Royal Free Hospital, London.

3 Проверка ошибок и выбросов

При любом исследовании всегда есть опасность допустить ошибки при наборе данных либо вначале, при измерениях, либо при сборе, переписывании и вводе данных в компьютер. Довольно трудно избежать этих ошибок. Однако можно сократить количество опечаток и описок путём тщательной проверки данных, как только они будут введены. Даже бегло просмотрев таблицу, можно обнаружить очевидные ошибки. В этой главе мы предлагаем ряд подходов, которые можно использовать при проверке данных.

ОПЕЧАТКИ

Опечатки — самые распространённые ошибки при вводе данных. Если количество данных невелико, вы можете сравнить уже напечатанные с оригинальными, просто просмотрев их, и проверить, нет ли ошибок. Однако при большом объёме данных на это потребуется слишком много времени. Можно ввести данные дважды и сравнить их при помощи компьютерной программы. Любые различия между двумя вариантами будут обнаружены. Хотя не исключено, что одна и та же ошибка может быть допущена в обоих случаях или данные в форме/анкете неправильные, но по крайней мере это сводит к минимуму количество ошибок. Недостаток этого метода заключается в том, что приходится дважды вводить данные, а это может повлечь большие затраты денег и времени.

ПРОВЕРКА ОШИБОК[1]

- Категориальные данные. Относительно легко проверить категориальные данные, так как отклики на каждую переменную (переменная отклика) могут принимать только одно из ограниченного ряда значений, поэтому данные, которые недопустимы, должны считаться ошибочными.
- Числовые (количественные) данные. При вводе числовых данных достаточно просто поменять местами цифры или не туда поставить десятичную запятую — и данные искажены, поэтому их довольно трудно проверить, но и здесь надо попытаться устранить ошибки. Числовые данные можно проверить по размаху, т.е. задать верхние и нижние ограничения для каждой переменной. Если величина находится за пределами этого интервала, то она не используется при дальнейшем исследовании.
- Даты. Часто трудно проверить точность дат, хотя иногда вам следует знать, что в определённый период времени данные могут выпадать (исчезать). Даты необходимо проверять хотя бы ради того, чтобы удостовериться, что они действительны. Например, 30 февраля не существует, как и не может быть в месяце больше 31 дня, а в году — больше 12 месяцев. Можно применять и некоторые логические проверки. Например, дата рождения больного должна соответствовать его возрасту, больной должен родиться до начала исследования (по крайней мере, в большинстве исследований). Кроме того, больной, который умер, не может осуществлять последующие визиты!

Во всех проверках величина должна быть исправлена только в том случае, если ошибка очевидна. Не следует менять данные только потому, что они выглядят необычными.

ОБРАБОТКА ПРОПУЩЕННЫХ ДАННЫХ

Всегда может случиться так, что некоторых данных не будет. Если не хватает большого объёма информации, результаты анализа вряд ли будут надёжны. Необходимо выяснить, почему сведения отсутствуют. Если нет данных для какой-то одной переменной и/или в отдельной подгруппе индивидуумов, это может указывать на то, что переменная не используется или никогда не была измерена для этой подгруппы. Тогда эту переменную надо исключить из исследования по данной группе индивидуумов. Может оказаться, что эти данные просто ещё в работе, на бумаге и пока не введены!

ВЫБРОСЫ (АНОМАЛЬНЫЕ ЗНАЧЕНИЯ)

Что такое выбросы?

Выбросы — наблюдения, которые отличаются от главной группы данных и несовместимы с остальными. Эти данные могут быть подлинными наблюдениями с очень экстремальными величинами переменной. Однако они могут появиться также в результате опечаток и в этом случае любые данные, вызывающие подозрение, должны быть проверены. Важно выяснить, имеются ли выбросы в наборе данных, так как они могут в значительной степени повлиять на результаты некоторых исследований.

К примеру, женщина, у которой рост 2,1 м, вероятнее всего воспринималась бы как выброс в большинстве наборов данных. Однако, хотя очевидно, что эта величина довольно высокая по сравнению с обычным ростом женщин, эти данные могут быть подлинными, так как эта женщина может быть просто очень высокой. В этом случае надо исследовать это наблюдение и дальше, может быть, проверить другие её показатели, такие, как возраст и масса, прежде чем принимать решение относительно истинности этой величины.

[1] Достаточно эффективно при небольших и средних объёмах введённых данных использовать последовательные сортировки наблюдений (желательно ещё при вводе в таблицы EXCEL), по отдельным признакам (столбцам). В этом случае ошибочные значения будут размещаться в первой или в последней (в зависимости от типа сортировки — по возрастанию или уменьшению значений) строке таблицы. Кроме того, отчётливо будет видно различие между ошибочным значением и ближайшим к нему. Прим. переводчика.

И только в том случае, если стало очевидно, что эти данные неверны, следует изменить значение.

Проверка выбросов

Самый простой метод состоит в том, чтобы во время набора данных проверять их глазами. Это приемлемо, если количество наблюдений не слишком большое и потенциальный выброс намного ниже или выше, чем остальная часть данных. Проверка по интервалу изменения также должна идентифицировать возможные выбросы. В качестве альтернативы данные могут быть представлены иным способом — выбросы будут идентифицированы на гистограммах и диаграммах рассеяния (см. также главу 29 с дискуссией о выбросах в регрессионном анализе)[1].

Обращение с выбросами

Нельзя убрать индивидуума из анализа только потому, что его/её данные выше или ниже, чем должны быть. Однако включение выбросов может повлиять на результаты, когда используются какие-нибудь статистические методы. Самый простой метод состоит в том, чтобы повторить анализ как с включёнными, так и с исключёнными данными. Если результаты окажутся одинаковыми, то в этом случае выбросы не окажут большого влияния на результаты. Однако, если результаты сильно отличаются, следует применить соответствующие методы, при которых выбросы не повлияют на исследование данных. Они включают в себя применение преобразований и непараметрических критериев.

ПРИМЕР

Значения были введены ошибочно из колонки с пропуском?

Пропуски закодированы как «.»

Действительно ли это так? Вряд ли это корректно

				Вмешательства, требуемые в течение беременности					Масса тела ребёнка						
№ пациента	Кровотечение	Пол ребёнка	Длительность беременности (недель)	Ингаляции	Внутримышечная инъекция	Внутривенная инъекция	Эпидуральное	Apgar Метка	кг	фунты	унции	Дата рождения	Возраст матери на момент рождения ребёнка	Группа крови	Частота кровотечений из дёсен
47	3	3	.	.	.	.	.	.	.	.	.	08/08/74	.	3	6
33	3	.	41	0	1	0	1	.	.	6	13	11/08/52	27,26	1	4
34	3	1	39	1	0	0	0	.	.	7	14	04/02/53	22,12	1	1
43	3	1	41	1	1	0	0	.	.	8	0	26/02/54	27,51	3	33
23	3	2	.	0	0	0	0	10/1-10/	11,19	.	.	29/12/65	36,58	1	3
49	3	3	.	.	.	.	.	.	.	.	.	09/08/57	.	1	5
51	3	3	.	.	.	.	.	.	.	.	.	21/06/51	.	3	5
20	2	41 →	0	1	0	0	.	.	7	12	15/08/96	25,61	3	3	.
64	4	.	.	1	1	0	0	.	.	.	.	10/11/51	24,61	3	2
27	3	1	14	1	0	0	0	ok	.	8	8	02/12/71	22,45	1	1
38	3	2	38	1	0	0	0	9/1-9/5	, .	6	10	12/11/61	31,60	1	1
50	3	2	40	0	0	0	0	.	.	5	11	06/02/68	18,75	1	6
54	4	1	41	0	1	0	0	.	.	7	4	17/10/59	24,62	3	2
7	1	1	40	0	0	0	1	.	.	6	5	17/12/65	20,35	2	6
9	1	2	38	0	1	0	0	.	.	5	4	12/12/96	28,49	3	3
17	1	4	.	.	.	.	.	.	.	.	.	15/05/71	26,81	1	5
53	3	2	40	0	0	1	0	.	.	8	7	07/03/41	31,04	1	3
56	4	2	40	0	0	0	0	.	3,5	.	0	16/11/57	37,86	3	3
58	4	1	40	0	1	0	1	.	.	8	0	17/063/47	22,32	3	Y
14	1	1	38	0	0	0	1	.	.	7	12	04/05/61	19,12	4	2

Переставлены цифры? Должно быть 41?

Это правильно? Слишком молодая, чтобы иметь ребенка!

Опечатка? Должно быть 17/06/47?

?

Рис. 3-1. Проверка ошибок в наборе данных.

После введения данных, как описано в главе 2, набор данных следует проверить на наличие ошибок. Жирно обведённые цифры — ошибочно введённые данные. Например, код «41» в колонке «пол ребёнка» неверен, и в результате этого у больной 20 отсутствуют данные; оставшаяся часть данных была введена в неверные колонки. Другие (например, необычные данные в колонках срок беременности и масса) тоже похожи на ошибки, но эти пометки должны быть проверены, прежде чем будет принято какое-либо решение, так как они могут отразить подлинность выбросов. В этом случае срок беременности пациентки за номером 27 был 41 нед, а масса 11,19 записана неверно. Так как оказалось невозможным найти массу ребёнка, эти данные были введены (закодированы) как пропущенные.

[1] Простым и эффективным средством поиска ошибок ввода и выбросов является построение двухмерных диаграмм рассеяния для всех возможных пар признаков. Несмотря на то что уже при 10 признаках число таких графиков будет равно (10×9)/2=45, даже беглый просмотр этих графиков позволяет сразу же увидеть аномальные значения в виде точек, расположенных на значительном удалении от основного массива точек. Кроме того, полезно провести сортировку наблюдений по значениям одного столбца (признака) по убыванию и по возрастанию. В результате этого аномальные значения и ошибки ввода также будут сразу видны. Прим. переводчика.

4 Графическое представление данных

Первое, что вы захотите сделать после ввода данных в компьютер, — обобщить их таким образом, чтобы можно было «ощутить» их. Это можно сделать, создавая диаграммы, таблицы или статистическую сводку. Диаграммы — мощный инструмент передачи информации о данных для представления простых итоговых изображений, для обнаружения выбросов и тенденций до того, как будет проведён какой-либо запланированный анализ.

ОДНА ПЕРЕМЕННАЯ

Частотное распределение

Эмпирическое частотное распределение переменной связывает каждое возможное наблюдение, группу наблюдений (т.е. интервал значений) или категории с их наблюдаемой частотой появления. Если мы заменим каждую частоту относительной частотой (процент от общей частоты), то мы сможем сравнить распределения в двух и более группах индивидуумов.

Представление частотных распределений

Как только получены частоты (или относительные частоты) для категориальных или дискретных числовых данных, их можно наглядно представить.

Столбчатая или колончатая диаграмма

Для каждой категории чертят отдельный горизонтальный или вертикальный столбик, длина которого пропорциональна частоте данной категории. Столбики отделяются друг от друга небольшим пробелом для того, чтобы показать, какие это данные, категориальные или дискретные (рис. 4-1, а).

Круговая диаграмма

Круговая диаграмма делится на секции, каждой из которых отводится определённая категория, таким образом, чтобы площадь каждого сектора была пропорциональна частоте этой категории (рис. 4-1, б).

Бывает трудно отобразить непрерывные числовые данные: чтобы изобразить их графически, их нужно

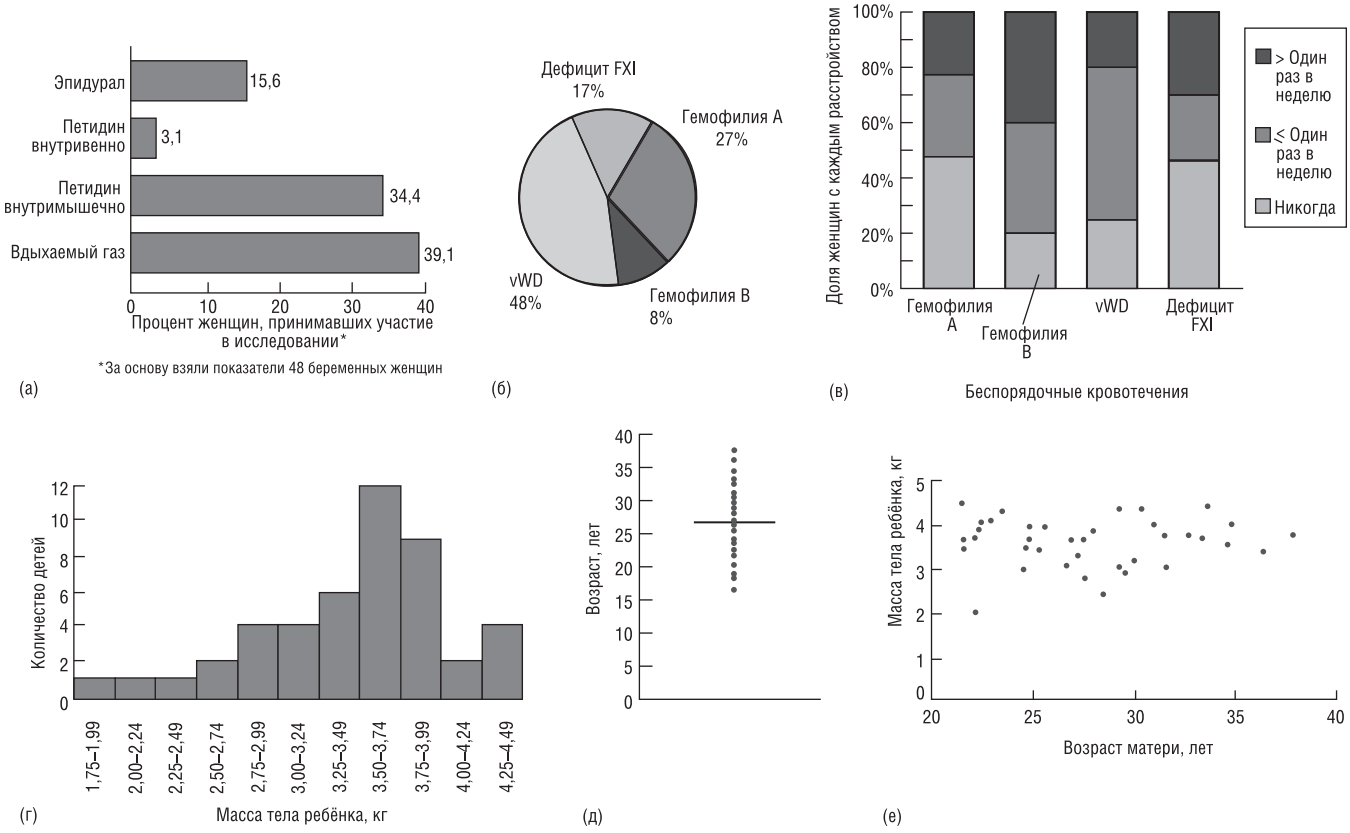

Рис. 4-1. Различные графические результаты, которые можно создать путём обобщения акушерских данных, полученных при исследовании женщин с беспорядочными кровотечениями (глава 2). Столбчатая диаграмма (а) показывает процентное отношение исследуемых женщин, которым необходимо облегчить боль во время родов от перечисленных вмешательств. Круговой график (б) показывает процентное отношение исследуемых женщин с беспорядочными кровотечениями. Сегментированная столбчатая диаграмма (в) показывает частоту, с которой женщины с беспорядочными различными кровотечениями испытывают кровотечения дёсен. Гистограмма (г) показывает массу тела ребёнка при рождении. Точечный график (д) показывает возраст матери во время рождения ребёнка со средним возрастом, который отмечен горизонтальной линией. Двухмерный график (е) показывает соотношение между возрастом матери во время родов (на горизонтальной, или *x*-оси) и массой тела ребёнка (на вертикальной, или *y*-оси).

сначала обобщить. Обычно используемые диаграммы включают следующее.

Гистограмма

Подобна круговой диаграмме, но здесь не должно быть пробелов между столбцами, так как данные непрерывны (рис. 4-1, г). Ширина каждого столбца гистограммы должна соответствовать интервалу значений данной переменной. Например, масса ребёнка (рис. 4-1, г) может быть от 1,75 до 1,99 кг, от 2 до 2,24 кг, от 4,25 до 4,49 кг. Площадь столбца пропорциональна частоте в данном интервале, поэтому если одна из групп охватывает более широкий интервал, чем другие, то основание столбца будет шире, а высота, соответственно, меньше. Обычно выбирают между пятью и 20 группами; интервал должен быть достаточно узким, чтобы отобразить структуру данных, но не должен быть настолько узким, чтобы они стали исходными данными (т.е. в каждом интервале по одному наблюдению). Гистограмма должна быть четко обозначена, чтобы было понятно, где находятся границы.

Точечный график

Каждое наблюдение изображено одной точкой на горизонтальной (или вертикальной) линии (рис. 4-1, д). Этот тип графика очень просто чертить, но только при небольшом объёме данных. Часто на диаграмме отображается обобщающая характеристика данных, такая, как среднее или медиана. Этот график можно использовать и для дискретных данных.

3	1,0	04
665	1,1	39
53	1,2	99
9751	1,3	1135677999
955410	1,4	0148
987655	1,5	00338899
9531100	1,6	0001355
731	1,7	00114569
99843110	1,8	6
654400	1,9	01
6	2,0	
7	2,1	19
10	2,2	

Беклометазон Плацебо

Рис. 4-2. График «стебель и листья» показывает объём форсированного выдоха в литрах у детей, вдохнувших беклометазон или плацебо (глава 21).

График «стебель и листья»[1]

Это смесь диаграммы и таблицы; он похож на гистограмму и эффективен для отображения данных по увеличению порядка величины. Обычно чертят вертикальный стебель, который состоит из нескольких первых цифр данных, приведённых по порядку. Выходящие наружу от этого стебля листья — конечная цифра всех данных по порядку, которые написаны горизонтально (рис. 4-2) в порядке увеличения порядка расположения числа.

График *Box-plot*

Этот тип графиков часто называют «ящиком с усами». Это вертикальный или горизонтальный прямоугольник, где две параллельных стороны отвечают верхнему и нижнему квартилям данных. Линия, проведённая поперёк прямоугольника, отвечает значению среднего. «Усы», начинающиеся в конце прямоугольника, обычно показывают минимальные и максимальные значения, но иногда указывают и особые процентили, например, 5-й и 95-й процентили (глава 6, рис. 6-1). Здесь же могут быть обозначены и выбросы (аномально большие или малые значения).

Форма частотного распределения

Выбор наиболее подходящего статистического метода часто зависит от формы распределения. Распределение данных чаще всего унимодальное, т.е. имеющее одну «вершину». Иногда распределение бимодальное (две «вершины») или равномерное (каждая величина одинаково вероятна и нет «вершин»). При унимодальном распределении главная цель состоит в том, чтобы увидеть, где находится бóльшая часть данных, относительно максимальных и минимальных значений. В частности, важно определить, каково распределение:

- симметричное — сосредоточенное вокруг средней точки, когда одна сторона — симметричное отражение другой (рис. 5-1);
- скошенное вправо (положительная асимметрия) — длинный правый «хвост» с одним или несколькими большими значениями. Такие данные являются весьма частыми в медицинском исследовании (рис. 5-2);
- скошенное влево (отрицательная асимметрия) — длинный левый «хвост» с одним или несколькими малыми значениями (рис. 4-1, г).

ДВЕ ПЕРЕМЕННЫЕ

Если одна переменная категориальная, тогда отдельные диаграммы, показывающие распределение второй переменной, должны быть начерчены для каждой категории. Другие графики, подходящие для таких данных, включают групповые или сегментные линии или графики с колонками (рис. 4-1, в).

Если обе переменные непрерывные или ординальные, то связь между ними можно изобразить при помощи двухмерной диаграммы рассеяния (скаттерплот) (рис. 4-1, е). Это двухмерный график, где оси переменных перпендикулярны друг другу. Одна переменная обычно называется *x* переменная и отображается на горизонтальной оси. Вторая переменная, известная как *y* переменная, наносится на вертикальную ось.

ИДЕНТИФИКАЦИЯ ВЫБРОСОВ ПРИ ИСПОЛЬЗОВАНИИ ГРАФИЧЕСКИХ МЕТОДОВ

Мы часто используем только одну переменную, отображающую данные, чтобы обнаружить выбросы. Например, длинный хвост на одной стороне гистограммы может указывать на удалённое, аномальное значение. Однако иногда выбросы могут стать очевидными только при рассмотрении соотношения между двумя переменными. Например, для женщины, рост которой 1,6 м, масса 55 кг не выглядит необычной, однако для женщины ростом 1,9 м такая масса будет необычно мала.

[1] В отечественных публикациях данный график незаслуженно не пользуется популярностью. Прим. переводчика.

5 Описание данных: «меры положения»

ОБОБЩЕНИЕ ДАННЫХ

Довольно трудно «прочувствовать» числовые измерения, до тех пор пока данные не будут обобщены содержательным образом. Диаграмма бывает полезна в качестве отправной точки. Также можно сжать информацию, представив величины, наиболее важные для характеристики данных. В частности, если знать, из чего состоит представленная величина или насколько широко рассеяны наблюдения, можно сформировать образ этих данных. Мера положения — общее понятие для числового выражения локализации (на числовой оси); которое описывает типичный результат измерения. Мы посвящаем эту главу мерам положения, самые распространённые из которых — среднее и медиана (табл. 5-1). Характеристики, которые отображают разброс или рассеяние наблюдений, мы включим в главу 6.

СРЕДНЕЕ АРИФМЕТИЧЕСКОЕ

Среднее арифметическое, которое очень часто называют просто среднее, для набора значений вычисляют следующим образом: складывают все значения и делят эту сумму на количество значений в этом наборе. Можно суммировать это буквальное выражение при помощи алгебраической формулы. Используя математическую систему обозначения, мы можем изобразить набор n наблюдений переменной x, как $x_1, x_2, x_3, .., x_n$. Например, x мог бы обозначать рост индивидуума (сантиметры), так чтобы x_1 обозначал рост первого индивидуума, а x_i — рост i индивидуума и т.д. Мы можем написать формулу для среднего арифметического наблюдений, пишется $\overline{x}$, а произносится «x с чертой»:

$$\overline{x} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n}.$$

Используя математическую систему обозначения, мы можем сократить это выражение:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n},$$

где Σ (греческая буква «сигма») означает «суммирование», а индексы внизу и вверху над этой буквой означают, что суммирование проводится от i=1 до i=n.

Это выражение часто сокращают ещё больше:

$$\overline{x} = \frac{\sum x_i}{n}, \text{ или } \overline{x} = \frac{\sum x}{n}.$$

МЕДИАНА

Если мы упорядочим данные по величине, начиная с самой маленькой величины и заканчивая самой большой, то медиана также будет характеристикой усреднения в упорядоченном наборе данных. Медиана делит ряд упорядоченных значений пополам, с равным числом этих значений как выше, так и ниже её (левее и правее медианы на числовой оси).

Вычислить медиану легко, если количество наблюдений n нечётное. Это будет наблюдение с номером $(n+1)/2$ в нашем упорядоченном наборе данных. Например, если n=11, то медиана — (11+1)/2=12/2=6, 6-е наблюдение в упорядоченном наборе данных. Если n чётное, тогда, строго говоря, медианы нет. Однако обычно мы вычисляем её как среднее арифметическое двух соседних средних наблюдений в упорядоченном наборе данных [т.е. наблюдений с номерами $(n/2)$ и $(n/2+1)$]. Так, например, если n=20, то медиана — среднее арифметическое из наблюдений с номерами 20/2=10 и (20/2+1)=11 в упорядоченном наборе данных.

Медиана подобна среднему значению, если данные симметричны (рис. 5-1), меньше, чем среднее значение, если данные скошены вправо (рис. 5-2), и больше, чем среднее значение, если данные скошены влево.

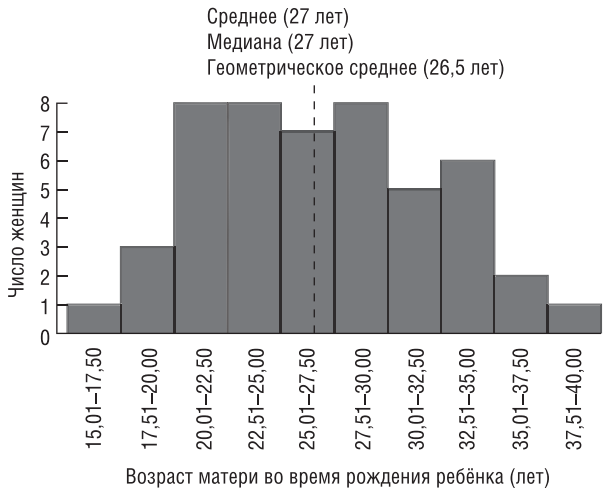

Рис. 5-1. Средняя, медиана и геометрическое среднее возраста женщин в исследовании, описанном в главе 2, во время рождения ребёнка. Распределение возраста довольно симметрично, поскольку все три «меры положения» дают близкие значения, обозначенные на графике пунктирной линией.

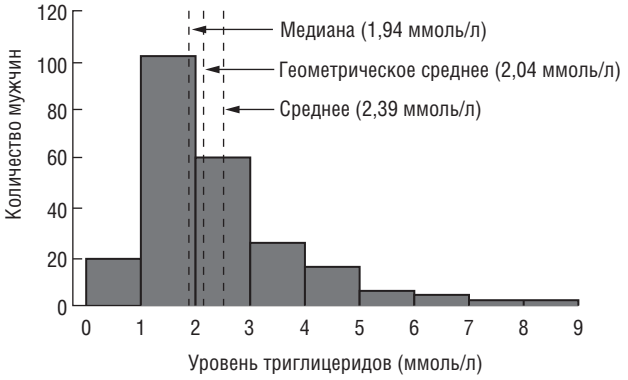

Рис. 5-2. Средняя, медиана и геометрическое среднее коцентрации триглицеридов в выборке из 232 мужчин с развившимся сердечным заболеванием (глава 19). Распределение концентрации триглицеридов скошено вправо, среднее арифметическое даёт более высокое значение «меры положения», чем медиана или геометрическое среднее.

МОДА

Мода — значение, которое встречается наиболее часто в наборе данных; если данные непрерывные, то мы обычно группируем их и вычисляем модальную группу. Некоторые наборы данных не имеют моды, потому что каждое значение встречается только один раз. Иногда можно встретить более одной моды; это происходит тогда, когда два значения или более встречаются одинаковое количество раз и частота встречаемости каждого из этих значений больше, чем таковые для любого другого значения. Мы редко используем моду как обобщающую характеристику.

СРЕДНЕЕ ГЕОМЕТРИЧЕСКОЕ

В случае если данные имеют несимметричное распределение, среднее арифметическое не будет обобщающим показателем такого распределения. Если данные скошены вправо, то можно создать распределение, которое будет более симметричным, если взять логарифм (по основанию 10 или по основанию *e*) каждого значения переменной в наборе данных. Среднее арифметическое значений этих логарифмов — характеристика распределения для преобразованных данных. Чтобы получить меру, которая будет иметь те же самые единицы измерения, как первоначальные наблюдения, мы должны осуществить обратное преобразование — потенцирование (т.е. взять антилогарифм) средней логарифмированных данных; мы называем такую величину среднее геометрическое. При условии, что распределение данных логарифма приблизительно симметричное, то среднее геометрическое подобно медиане и меньше, чем среднее необработанных данных (см. рис. 5-2).

ВЗВЕШЕННОЕ СРЕДНЕЕ

Мы используем взвешенное среднее в том случае, когда некоторые значения интересующей нас переменной x более важны, чем другие. Мы присоединяем массу w_i к каждому из значений x_i в выборке для того, чтобы учесть эту важность. Если значения x_1, x_2, x_3, .., x_n имеют соответствующие значения массы w_1, w_2, w_3, ... w_n, то взвешенное арифметическое среднее выглядит следующим образом:

$$\frac{w_1x_1 + w_2x_2 + ... + w_nx_n}{w_1 + w_2 + ... + w_n} = \frac{\sum w_ix_i}{\sum w_i} \quad .$$

Например, предположим, что мы заинтересованы в определении средней продолжительности пребывания госпитализированных больных в каком-либо районе и знаем средний реабилитационный период для больных в каждой больнице. Необходимо учитывать количество информации, в первом приближении принимая за вес каждого наблюдения такой показатель, как количество больных в больнице.

Взвешенное среднее и среднее арифметическое идентичны, если каждый вес равен единице.

Таблица 5-1. Преимущества и недостатки мер положения

Тип среднего	Преимущества	Недостатки
Среднее	Используются все значения набора данных. Определяется математически выполнимым алгебраическим выражением. Известно выборочное распределение (см. главу 6)	Искажается выбросами. Искажается асимметричными данными
Медиана	Не искажается выбросами. Не искажается асимметричными данными	Игнорирует большую часть информации. Не определяется алгебраически. Усложняется в выборочном распределении
Мода	Легко определяется для категориальных данных	Игнорирует большую часть информации. Не определяется алгебраически. Неизвестно выборочное распределение
Среднее геометриическое	До обратного преобразования имеет те же самые преимущества, что и среднее	Подходит, если логарифмическое преобразование образует симметричное распределение
Взвешенное среднее	Те же самые преимущества, что и у среднего. Приписывает соответствующий вес каждому наблюдению. Алгебраически определяется	Вес должен быть известен или оценён