Авива Петри • Кэролайн Сэбин

# НАГЛЯДНАЯ МЕДИЦИНСКАЯ СТАТИСТИКА

4-е издание,
переработанное и дополненное

Перевод с английского
под редакцией В.П. Леонова

Москва
ИЗДАТЕЛЬСКАЯ ГРУППА
«ГЭОТАР-Медиа»
2021

# Содержание

# Обработка данных Часть 1

## Главы

# Типы данных 1

**Цели изучения**

К концу этой главы вы должны овладеть следующими знаниями.

- Различие выборки и популяции (генеральной совокупности).
- Различие категоризованных (качественных) и числовых данных.
- Описание различных типов категоризованных и числовых данных.
- Объяснение значения терминов «переменная», «процент», «отношение», «доля», «оценка», «метка».
- Объяснение, что означает цензурирование данных.

## Данные и статистика

Цель большинства исследований состоит в сборе данных, которые впоследствии помогают получить информацию о какой-либо области исследования. Наши данные основываются на наблюдениях одной или нескольких переменных; термин **переменная** означает количественный показатель, способный изменяться. Например, мы можем собрать основную клиническую и демографическую информацию на больных со специфической болезнью. Переменные, вызывающие интерес, могут включать пол, возраст и рост больного.

Обычно мы получаем свои данные из **выборки** индивидуумов, которые представляют **популяцию** — группу индивидуумов, которая представляет для нас интерес. Наша цель состоит в том, чтобы сгруппировать эти данные и извлечь из них полезную информацию. **Статистика** охватывает различные методы, например сбор данных, их обобщение, анализ данных и подведение итогов, основанных на полученных данных: мы используем статистические методы, чтобы достичь своей цели.

Данные могут принимать различные формы. Первое, что мы должны знать, прежде чем примем решение о том, какой статистический метод окажется наиболее подходящим для его использования, — это то, какой тип данных принимает каждая переменная. Каждая переменная и результирующие показатели будут принимать один из двух типов: **категориальная (качественная)** или **числовая (количественная) переменная** (рис. 1.1).

## Категориальные (качественные) данные

Встречаются в том случае, когда индивидуум может принадлежать только к одной из множества категорий переменной.

- **Номинальные данные —** категории не упорядочиваются, а просто имеют названия. Например, группа крови (A, B, AB и 0) 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. Примеры включают стадии болезни (заболевание в запущенной стадии, средняя, легкая форма болезни или ее отсутствие) и степень боли (сильная, умеренная, слабая, отсутствие боли).

Категориальная (качественная) переменная — это **бинарная** или **дихотомическая переменная**, когда имеются только две возможные категории. Примеры включают «да/нет», «умер/жив» или «больной имеет заболевание/больной не имеет никаких заболеваний».


**Рис. 1.1.** Схема, показывающая различные типы переменных

## Числовые (количественные) данные

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

- **Дискретные данные** имеют место, когда переменная может принимать только определенные числовые значения. Часто ведется подсчет количества событий, таких как количество посещений врача в год или количество заболеваний человека за последние пять лет.
- **Непрерывные данные** имеют место, когда нет ограничений в отношении данных, которые переменная может принимать, например, вес или рост.

## Различие между типами данных

Мы часто используем различные статистические методы в зависимости от того, являются ли данные категориальными или числовыми. Хотя различие между категориальными и числовыми данными и так понятны, в некоторых случаях оно не совсем ясно. Например, когда у нас есть переменная с множеством установленных категорий (например, боль может иметь семь категорий), могут возникнуть трудности, как отличить ее 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 является примером переменной, которую часто трактуют как дискретную, хотя на самом деле она является непрерывной. Обычно мы ссылаемся на «возраст в последний день рождения», нежели на «возраст по состоянию на сегодняшний день», и поэтому женщина, которая сообщает, что ей 30, может иметь в виду, что ей только что исполнилось 30 или ей может быть почти 31.

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.

## Производные (вторичные) данные

Мы можем столкнуться с множеством других типов данных в области медицины. Они включают следующие.

- **Проценты** — они могут возникать при рассмотрении вопроса относительно улучшения состояния больного во время лечения, например состояние больного (объем форсированного выдоха в 1 с, $FEV_1$) может улучшиться на 24% после лечения новым препаратом. В этом случае имеет место степень улучшения, а не абсолютные данные, которые представляют интерес.
- **Пропорции** или **отношения** — иногда встречаются два варианта пропорций или отношений. Например, индекс массы тела (индекс Кетле) высчитывается следующим образом: вес индивидуума (кг) делят на квадрат его/ее роста ($м^2$), таким образом, делается оценка, превышает ли ее/его вес норму или, наоборот, имеется недостаток веса.
- **Интенсивность** — относительная частота заболеваний, где количество заболеваний делят на общее число лет, в течение которых были прослежены все пациенты в этом исследовании (глава 31), является общепринятой при эпидемиологическом исследовании (глава 12).
- **Метки, оценки** — иногда мы пользуемся произвольными значениями, то есть метками, в том случае, когда мы не можем измерить количество. Например, ряд вопросов на ответы относительно качества жизни можно суммировать, для того чтобы дать полную оценку качества жизни каждого индивидуума.

Все эти переменные можно рассматривать как непрерывные переменные в большинстве исследований. 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 рассматривать цензурированные данные в ситуациях, иллюстрированных следующими примерами.

- 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, то есть она будет подвержена цензуре. Например, при измерении уровней вируса, у которых действительный уровень ниже измерительного предела «X», их уровень определяется как «необнаруживаемый» или «невычисляемый», даже при том, что в образце может находиться какой-нибудь вирус. В этой ситуации, если такой уровень меньше нижнего инструментального уровня, результат измерения может быть представлен как «<X». Аналогично некоторые инструменты могут давать надежные результаты измерения лишь при условии измерения величин менее определенного максимального значения «Y». Поэтому любые значения выше этой величины также будут цензурированы, и результат измерения будет представлен выражением «>Y».
- Мы можем столкнуться с цензурированными данными, например, когда во время прохождения испытания некоторые больные выбывают или отстраняются от него до того, как это испытание будет окончено. Этот тип данных подробно обсуждается в главе 44.

# Ввод данных 2

**Цели изучения**

К концу этой главы вы должны овладеть следующими знаниями.

- Описание различных форматов, для того чтобы ввести данные в компьютер.
- Описание в общих чертах дизайна анкетного опроса.
- Отличие между отдельной и многомерной переменной.
- Описание, как кодировать пропущенные (неизмеренные) значения переменных.

При проведении какого-либо исследования вам почти всегда необходимо будет вводить данные в компьютерный пакет прикладных программ. Компьютеры — неоценимая вещь, при помощи которой вы можете проверить правильность данных, ускорить сбор данных и анализа, а также с ней гораздо проще проверять ошибки, производить графические подсчеты данных и новых переменных. 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 и ограничения: 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** (**A**merican **S**tandard **C**ode for **I**nformation **I**nterchange — стандартный код информационного обмена США) или в **текстовом** файле. Данные в ASCII формате могут читаться большинством пакетов. ASCII формат состоит из текста, который вы можете читать с компьютера. Обычно каждая переменная в файле отделяется от следующей каким-нибудь **разделителем**, часто пространством или запятой. Такой формат известен как **свободный формат**.

Самый простой способ ввода данных в ASCII формате — это печатать данные непосредственно в нем, используя текстовый редактор либо иной редакторский пакет. В качестве альтернативы данные, находящиеся в пакете электронных таблиц (Excel), могут быть сохранены в текстовом формате. Используя любой подход при исследовании, общепринято, чтобы каждой строке данных соответствовал отдельный индивидуум, а каждая колонка соответствовала переменной, хотя может возникнуть необходимость в продолжении последовательных рядов, если на каждого индивидуума собрано большое количество переменных.

## Планирование ввода данных

При сборе данных вам необходимо будет использовать форму или анкету для занесения данных. Если они хорошо разработаны, они помогут 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.

## Категориальные данные

Если вы имеете дело 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», «легкая боль», «средняя боль» и «сильная боль» соответственно. Эти коды могут быть добавлены к формам при сборе данных. Для бинарных данных, например ответы 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: «Каковы симптомы у больного?» В этом случае пациент может испытывать разные симптомы. Существует два способа обработки этих данных, в зависимости от того, какую из двух следующих ситуаций использовать.
  - **Существует несколько возможных симптомов, и многие из них человек может испытывать.** Можно создать ряд различных бинарных переменных, все зависит от того, ответит ли больной «да» или «нет» на присутствие возможных симптомов. Например: «Был ли кашель у больного?», «Болело ли у больного горло?»
  - **Существует огромное количество возможных симптомов, но больной может иметь только некоторые из них.** Можно создать ряд различных номинальных переменных; каждая из следующих друг за другом переменных позволит вам определить наличие того или иного симптома у больного. Например: «Какой симптом был первым у больного?», «Каким был второй симптом?». Вы заранее должны определить максимальное количество симптомов, которые, как вы полагаете, больной может иметь.

## Числовые данные

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

## Множественные формы на одного больного

Иногда информация собирается 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 объединить все данные, собранные на одного человека при исследовании.

## Проблемы с датами и периодами

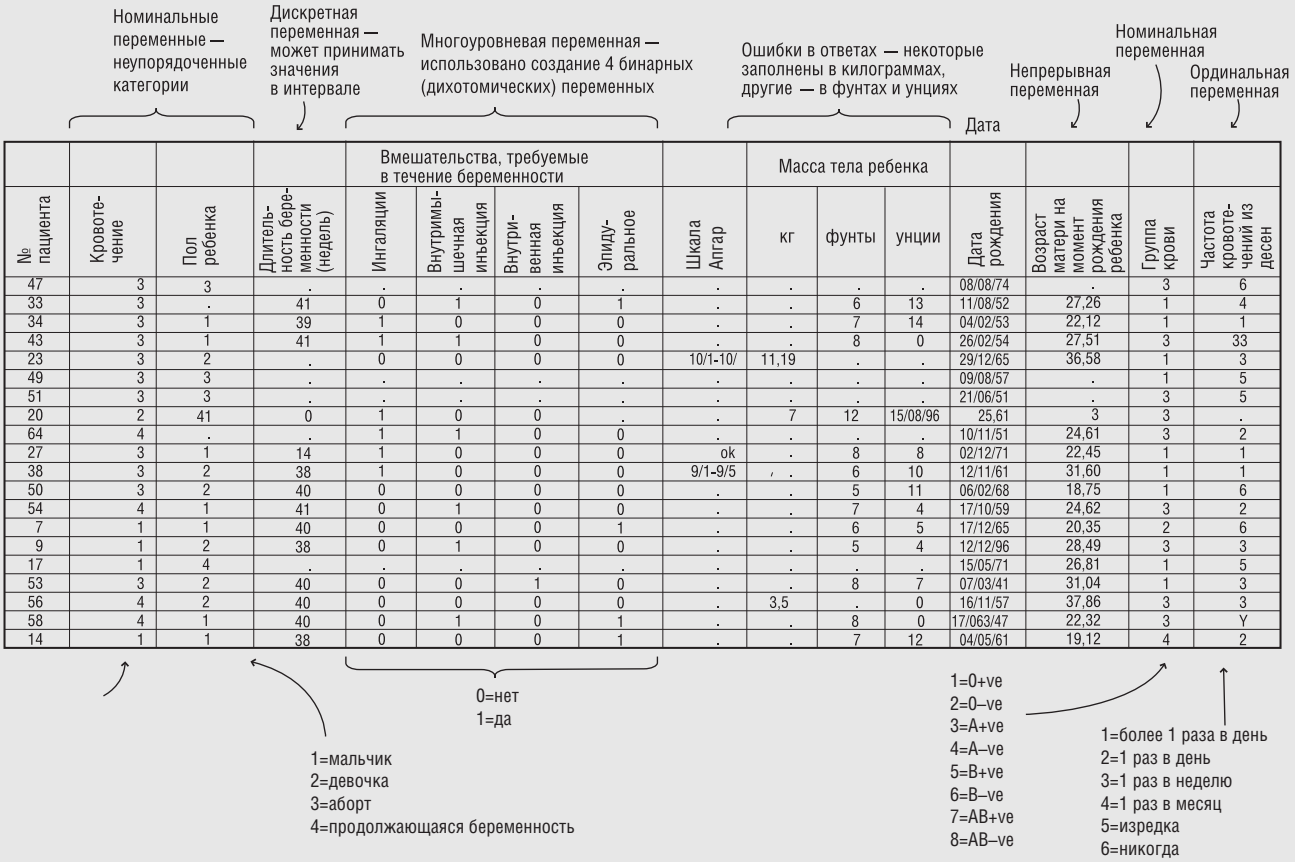
Даты и периоды должны вводиться последовательно, например: либо день/месяц/год, либо месяц/день/год, но они не должны быть взаимозаменяемыми. Важно установить, какой формат может читаться в данном статистическом пакете.

## Кодирование отсутствующих (пропущенных) данных

Вам следует определиться, что вы будете делать с отсутствующими данными, прежде чем вводить данные. В большинстве случаев вы будете вынуждены использовать какой-нибудь символ для недостающих данных. Статистические пакеты предлагают различные способы обозначения недостающих данных. Некоторые пакеты

используют специальные символы (например, точка или звездочка) для обозначения пропущенных данных, принимая во внимание это во время анализа, тогда как другие требуют от вас, чтобы вы ввели свой код для обозначения отсутствующих данных (обычно используемые значения 9, 999 или –9999). Выбранное значение должно быть одно для всех переменных, и его нельзя использовать для другой переменной. Например, при вводе категориальной переменной с четырьмя категориями (имеющиеся коды 1, 2, 3 и 4) вы можете выбрать цифру 9 для недостающих данных. Однако, если этой переменной является «возраст ребенка», необходимо выбрать другой код, например «–9». Более подробно отсутствующие данные рассматриваются в главе 3.

## Пример




| № пациента | Кровотечение | Пол ребенка | Длительность беременности (недель) | Ингаляции | Внутримышечная инъекция | Внутривенная инъекция | Эпидуральное | Шкала Апгар | кг | фунты | унции | Дата рождения | Возраст матери на момент рождения ребенка | Группа крови | Частота кровотечений из десен |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 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 |

**Рис. 2.1.** Часть электронной таблицы, в которой показаны собранные данные на примерах 64 женщин с наследственными беспорядочными кровотечениями

Данная 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 о возникновении 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 рис. 2.1 приведена кодовая схема для категориальных переменных. Каждый ряд отведен для отдельного пациента; каждая колонка — для отдельной переменной. В случае, если женщина все еще беременна, возраст женщины во время рождения высчитывался на день рождения младенца. Данные, касающиеся живорожденных 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.

# Проверка ошибок и выбросов 3

**Цели изучения**

К концу этой главы вы должны овладеть следующими знаниями.

- Описание, как проверять наличие ошибок в данных.
- Различение случайно полностью пропущенных данных, пропущенных случайно и пропущенных не случайно.
- Описание в общих чертах действий с пропущенными данными, различение одиночных и множественных расчетных данных.
- Определение выбросов (аномальных значений).
- Объяснение, как проверять наличие выбросов (аномальных значений).

При любом исследовании всегда существует возможность обнаружить ошибки в наборе данных либо вначале при измерениях, либо при сборе, переписывании и вводе данных в компьютер. Довольнотаки трудно устранить все эти ошибки. Однако можно сократить количество опечаток и описок путем тщательной проверки данных, как только они будут введены. Даже просто просмотрев глазами, можно сразу же обнаружить очевидные ошибки. В этой главе мы предлагаем ряд подходов, которые вы можете использовать при проверке данных.

## Опечатки

Опечатки — это самые распространенные ошибки при вводе данных. Если количество данных невелико, вы можете сравнить уже напечатанные данные с оригинальными, просто просмотрев их, и таким образом проверить, есть ли ошибки. Однако на это потребуется много времени, если количество данных большое. Также можно ввести данные дважды и сравнить эти данные при помощи компьютерной программы. Любые различия между двумя наборами данных будут обнаружены. Хотя этот подход не исключает возможность, что та же самая ошибка была введена неправильно в обоих случаях или то, что данные в форме/анкете неправильные, но, по крайней мере, хотя бы 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

- **Категориальные данные.** Относительно легко проверить категориальные данные, так как отклики на каждую переменную (переменная отклика) могут принимать только одно значение из ряда ограниченных данных. Поэтому данные, которые не допустимы, должны считаться ошибочными.
- **Числовые (количественные) данные.** Числовые данные часто трудно проверить, но и здесь встречаются ошибки. Например, достаточно просто поменять местами цифры или не туда поставить десятичную запятую при вводе числовых данных. Числовые данные могут быть проверены по **размаху**, то есть верхние и нижние ограничения могут быть заданы 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 определенный 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 и не может быть больше 12 месяцев. Также можно применять и некоторые логические проверки. Например, дата рождения больного должна соответствовать ее/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 очевидно, что была допущена ошибка. Не следует менять данные только потому, что они выглядят необычными.

## Обработка пропущенных данных

Всегда может случиться так, что некоторые данные будут отсутствовать. Если доля пропущенных данных слишком высока, получение надежных результатов маловероятно. Необходимо выяснить причины, почему данные отсутствуют: если произошло так, что данные отсутствуют на какой-то одной переменной и/или в отдельной подгруппе индивидуумов, это может указывать на то, что данная переменная не используется или никогда не была измерена для данной группы индивидуумов. В этом случае необходимо исключить из исследования данную переменную или данную группу индивидуумов. Имеются различные типы пропущенных данных[1].

**Полностью пропущенные случайно (ППС).** Пропущенные значения действительно случайно распределены в наборе данных, а их пропуск не имеет отношения к изучаемой переменной. Маловероятно, что результирующий параметр расценивается как смещенный (глава 34). Таким примером является случай невыполнения пациентом предписаний врача, когда первый попал в автокатастрофу.

**Пропущенные случайно (ПС).** Пропущенные значения переменной не зависят от этой переменной, но могут быть полностью объяснены имеющимися в наличии значениями одной или других переменных. Например, предположим, что индивидуумов попросили вести дневник питания, если их ИМТ превышает 30 кг/м$^2$: данные дневника питания являются ПС, так как пропуск обусловлен исключительно ИМТ (дневник не ведется, когда ИМТ ниже установленного верхнего предела).

**Пропущенные не случайно (ПНС).** Случай, когда данные касательно определенной переменной пропущены, имеет прямое отношение к этой переменной. В такой ситуации наши результаты будут в большой степени смещены. Например, предположим, что нам интересны измерения, отражающие состояние здоровья пациентов, и такого рода информация пропущена у некоторых из них по причине недостаточно четкого выполнения врачебных предписаний: вероятнее всего, мы получим слишком оптимистическую общую картину состояния здоровья данных пациентов, если не примем во внимание пропущенные данные в процессе их анализа.

Если данные не принадлежат к типу ПНС, мы, вероятно, можем оценить (рассчитать) пропущенные данные. Простой поход к данному вопросу — замена пропущенных наблюдений их средним значением для этой переменной или, если данные продольные, данными последнего наблюдения. Это примеры единичного условного расчета. В случае множественного условного расчета мы формируем некоторое число (обычно около пяти) наборов условных расчетных данных, взятых из исходного набора, с пропущенными значениями, замененными условно расчетными данными. Последние выводятся из подходящей модели, позволяющей случайное изменение данных.

Далее мы применяем стандартные статистические процедуры по отношению к каждому полному условному набору данных и в конечном счете группируем результаты для последующего их ана-

[1] *Bland M.* An Introduction to Medical Statistics. 4th ed. Oxford University Press, 2015.

лиза. Также доступен альтернативный статистический подход к работе с пропущенными данными[1], но лучший вариант — минимизировать количество недостающих данных в самом начале.

## Выбросы (аномальные значения)

### Что такое выбросы?

Выбросы — это наблюдения, которые отличаются от главной группы данных и не совместимы с остальными данными. Эти данные могут быть подлинными наблюдениями с экстремальными уровнями переменной. Однако они могут появиться также в результате опечаток, и в этом случае любые данные, вызывающие подозрение, должны быть проверены. Важно проверить, имеются ли выбросы в наборе данных, так как они в значительной степени могут повлиять на результаты некоторых исследований (глава 29).

К примеру, женщина, у которой рост 2,1 м, вероятнее всего, воспринималась бы как выброс в большинстве наборов данных. Однако, хотя и очевидно, что эта величина является довольнотаки высокой, по сравнению с обычным ростом женщин, эти данные могут быть подлинными, так как эта женщина может быть очень высокой. В этом случае вам следует исследовать это наблюдение и далее, может быть, проверить другие ее показатели, такие, как возраст и вес, прежде чем принимать решение истинности этой величины. И только в том случае, если стало очевидно, что эти данные неверны, следует изменить данное значение.

[1] *Horton N.J., Kleinman K.P.* Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models // American Statistician. 2007. Vol. 61, N 1. P. 71–90.

### Проверка выбросов

Самый простой метод состоит в том, чтобы во время набора данных проверять их на глаз[2]. Это приемлемо, если количество наблюдений не слишком большое и если значения потенциального выброса намного ниже или выше, чем остальная часть данных. Проверка по интервалу изменения также должна идентифицировать возможные выбросы. В качестве альтернативы данные могут быть представлены для обнаружения иным способом (глава 4) — выбросы будут идентифицированы на гистограммах и диаграммах рассеяния (см. также главу 29 с дискуссией о выбросах в регрессионном анализе).

### Обработка выбросов

Важно не удалять индивидуума из анализа только потому, что его/ее данные выше или ниже, чем можно было бы ожидать. Однако включение выбросов в анализ может повлиять на результаты, когда используются какие-нибудь статистические методы. Самый простой метод состоит в том, чтобы повторить анализ как с включенными, так и с исключенными данными — это вид анализа чувствительности (см. главу 35). Если результаты подобны, 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 результат анализа данных. Они включают применение преобразований (глава 9) и непараметрических критериев (глава 17).

[2] Наглядные примеры таких проверок приведены в [252].

### Пример


| № пациента | Кровотечение | Пол ребенка | Длительность беременности (недель) | Вмешательства, требуемые в течение беременности: ингаляции | внутримышечная инъекция | внутривенная инъекция | эпидуральное | Шкала Апгар | Масса тела ребенка: кг | фунты | унции | Дата рождения | Возраст матери на момент рождения ребенка | Группа крови | Частота кровотечений из десен |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 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 | 3 |
| 51 | 3 | 3 | . | . | . | . | . | . | . | . | . | 21/06/51 | . | 3 | 3 |
| 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 |

**Рис. 3.1.** Проверка ошибок в наборе данных

После введения данных, как описано в главе 2, набор данных следует проверить на наличие ошибок. Обведенные цифры — это ошибочно введенные данные. Например, код «41» в колонке «пол ребенка» неверен, и в результате этого у больной 20 обнаружилось отсутствие данных; оставшаяся часть данных была введена в неверные колонки. Другие (например, необычные данные в колонках «срок беременности» и «вес») 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-я неделя, а вес 11,19 записан неверно. Так как оказалось невозможным найти вес ребенка, эти данные были введены (закодированы) как пропущенные.

# Графическое представление данных 4

## Цели изучения

К концу этой главы вы должны овладеть следующими знаниями.

- Объяснение, что подразумевается под распределением частот.
- Описание формы распределения частот.
- Описание следующих схематичных графиков: столбиковая диаграмма или гистограмма, круговая диаграмма, точечный график, график «стебель и листья», график Box-plot («ящик с усами»), двумерная диаграмма рассеяния (скаттерплот).
- Объяснение, как идентифицировать выбросы из схемы в различных ситуациях.
- Описание ситуации, когда уместно использовать соединительные линии в диаграмме.

Первое, что вы захотите сделать после ввода данных в компьютер, — это обобщить их таким образом, чтобы можно было «ощутить» эти данные. Это можно сделать, создавая диаграммы, таблицы или статистическую сводку (главы 5 и 6). Диаграммы — это мощный инструмент передачи информации о данных, для представления простых итоговых изображений, обнаружения выбросов и тенденций, до того как будет проведен какой-либо запланированный анализ.

## Одна переменная

### Частотное распределение

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

### Представление частотных распределений

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

- **Столбчатая или колончатая диаграмма** — для каждой категории чертится отдельный горизонтальный или вертикальный столбик, длина которого пропорциональна частоте для данной категории. Столбики отделяются друг от друга небольшим пробелом, для того чтобы показать, что эти данные являются категориальными или дискретными (рис. 4.1, а).
- **Круговая диаграмма** — диаграмма, которая делится на секции, причем каждая из них отводится для определенной категории таким образом, чтобы площадь каждого сектора была пропорциональна частоте этой категории (рис. 4.1, б).

Часто бывает трудно отобразить *непрерывные числовые* данные, так как, прежде чем их начертить, чтобы обнажить суть, их

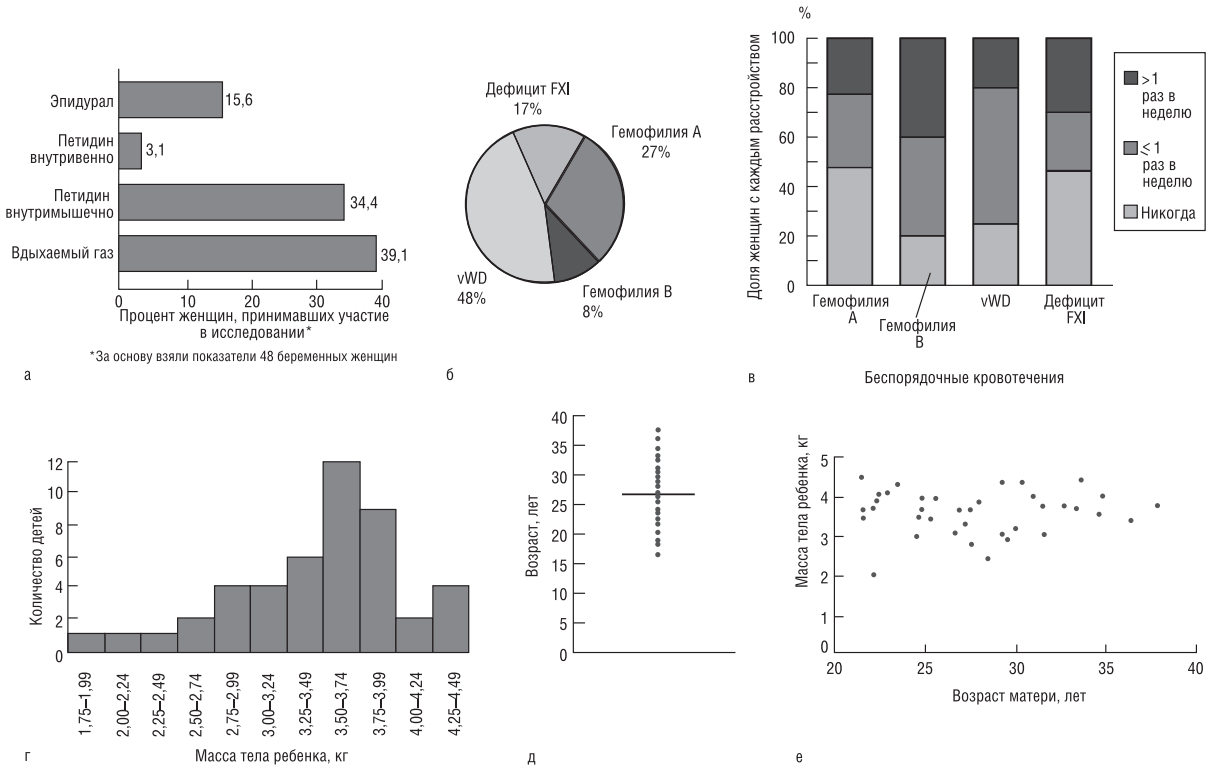



**Рис. 4.1.** Различные графические результаты, которые можно создать путем обобщения акушерских данных, полученных при исследовании женщин с беспорядочными кровотечениями (глава 2): а — столбчатая диаграмма показывает процентное отношение исследуемых женщин, которым необходимо облегчить боль во время родов от перечисленных вмешательств; б — круговой график показывает процентное отношение исследуемых женщин с беспорядочными кровотечениями; в — сегментированная столбчатая диаграмма показывает частоту, с которой женщины с беспорядочными различными кровотечениями испытывают кровотечения десен; г — гистограмма показывает вес ребенка при рождении; д — точечный график показывает возраст матери во время рождения ребенка в сравнении со средним возрастом, который отмечен горизонтальной линией; е — двумерный график показывает соотношение между возрастом матери во время родов (на горизонтальной, или *x*-оси) и весом ребенка (на вертикальной, или *y*-оси).

нужно 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 для данной переменной. Например, вес ребенка (см. рис. 4.1, г) может быть от 1,75–1,99 кг, от 2,00–2,24 кг, от 4,25–4,49 кг. Площадь столбца пропорциональна частоте в данном интервале. Поэтому, если одна из групп охватывает более широкий интервал, чем другие, основание столбца будет шире, а высота, соответственно, меньше. Обычно выбирают между 5 и 20 группами; интервал должен быть достаточно узким, чтобы отобразить структуру данных, но не должен быть настолько узким, чтобы они стали исходными данными (то есть в каждом интервале по одному наблюдению). Гистограмма должна быть четко обозначена, чтобы было понятно, где находятся границы.
- **Точечный график** — каждое наблюдение отображено одной точкой на горизонтальной (или вертикальной) линии (рис. 4.1, д). Этот тип графика очень просто чертить, но при большом объеме данных это затруднительно. Часто на диаграмме отображается обобщающая характеристика данных, такая, как среднее или медиана (глава 5). Этот график можно использовать и для дискретных данных.
- **График «стебель и листья»** — смесь диаграммы и таблицы; он похож на гистограмму и эффективен для отображения данных по увеличению порядка величины. Обычно чертится вертикальный **стебель**, который состоит из нескольких первых цифр данных, приведенных по порядку. Выходящие наружу от этого стебля **листья** — это конечная цифра всех данных по порядку, которые написаны горизонтально (рис. 4.2) в порядке увеличения порядка расположения числа.
- **График Box-plot** (часто называют **график «ящик с усами»**) — это вертикальный или горизонтальный прямоугольник, где две параллельные стороны прямоугольника отвечают верхнему и нижнему квартилям данных (глава 6). Линия, проведенная поперек прямоугольника, отвечает значению среднего (глава 5). Усы, начинающиеся в конце прямоугольника, обычно показывают минимальные и максимальные значения, но иногда указывают и особые процентили, например 5-й и 95-й процентиль (глава 6, рис. 6.1). Здесь же могут быть обозначены и выбросы (аномально большие или малые значения).

| | | |
|---|---|---|
| 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 объем форсированного выдоха (ОФВ, в литрах) у детей, вдохнувших дипропионат беклометазона или плацебо (глава 21)

### Форма частотного распределения

Выбор наиболее подходящего статистического метода часто зависит от формы распределения. Распределение данных чаще всего **унимодальное**, то есть имеет одну вершину. Иногда распределение **бимодальное** (две вершины) или **равномерное** (каждая величина одинаково вероятна и нет вершин). Когда распределение унимодальное, главная цель состоит в том, чтобы увидеть, где находится большая часть данных относительно максимальных и минимальных значений. В частности, важно определить, является ли распределение:

- **симметричным** — сосредоточенным вокруг средней точки, одна сторона которой является симметричным отражением другой (рис. 5.1);
- **скошенным вправо (положительная асимметрия)** — длинный правый хвост с одним или несколькими большими значениями. Такие данные весьма часты в медицинском исследовании (рис. 5.2);
- **скошенным 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 **двумерной диаграммы рассеяния (скаттерплот)** (рис. 4.1, е). Это двумерный 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 удаленное, аномальное значение. Однако иногда выбросы могут стать очевидными только при рассмотрении соотношения между двумя переменными. Например, для женщины, рост которой 1,6 м, вес 55 кг не выглядит необычным, однако для женщины, рост которой 1,9 м, такой вес будет необычно маленьким.

## Использование соединительных линий на диаграммах

Использование соединительных линий на диаграммах может быть обманчивым. Соединение линиями говорит о том, что значения по оси *x* должны быть каким-то образом упорядочены, например это может быть, если ось *x* отражает некоторую величину времени или дозу. Там, где этого нет, точки (наблюдения) не должны соединяться с помощью линий. И наоборот, если имеется зависимость между различными точками (к примеру, потому что они являются результатами для одного и того же пациента в два разных момента времени, например до и после лечения), полезно для наглядности отображения такой взаимосвязи соединять соответствующие точки прямой линией (рис. 20.1). При отсутствии таких соединительных линий осознание важности такой информации может отсутствовать.

# Описание данных: «меры положения» 5

**Цели изучения**

К концу этой главы вы должны овладеть следующими знаниями.

- Объяснение, что является средним значением.
- Описание соответствующего использования каждого из следующих типов среднего значений: среднее арифметическое, мода, медиана, геометрическое среднее, взвешенное среднее значение.
- Объяснение, как вычислить каждый вид среднего значения.
- Перечисление преимуществ и недостатков каждого вида средних величин.

## Обобщение данных

Довольно трудно прочувствовать числовые измерения до тех пор, пока данные не будут обобщены содержательным образом. Диаграмма (глава 4) часто полезна в качестве отправной точки. Также мы можем сжать информацию, определив величины, которые представляют важные характеристики данных. В частности, если бы мы знали, из чего состоит представленная величина или насколько широко рассеяны наблюдения, тогда бы мы смогли сформировать образ этих данных. **Мера положения** — это общее понятие для числового выражения **локализации** (на числовой оси), которое описывает типичный результат измерения. Мы посвящаем эту главу мерам положения, самыми распространенными из них являются 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 вычисляется следующим образом: складывают все значения и делят их 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 $x_1, x_2, x_3, \ldots\ldots x_n$. Например, $x$ мог бы обозначать рост индивидуума (см), так чтобы $x_1$ обозначал рост первого индивидуума, а $x_i$ — рост $i$-го индивидуума и т.д. Мы можем написать формулу для среднего арифметического наблюдений (пишется $\overline{x}$, а произносится «$x$ с чертой»[1]):

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

Используя математическую систему обозначения, мы можем сократить это выражение:

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

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

Это выражение часто сокращают еще больше:

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

## Медиана

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

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

Медиана подобна среднему значению, если данные симметричные (рис. 5.1), меньше среднего значения, если данные скошены вправо (рис. 5.2), и больше среднего значения, если данные скошены влево.

## Мода

**Мода** — это значение, которое встречается наиболее часто в наборе данных; если данные непрерывные, мы обычно группируем данные 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 в том случае, когда два или больше значений 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.

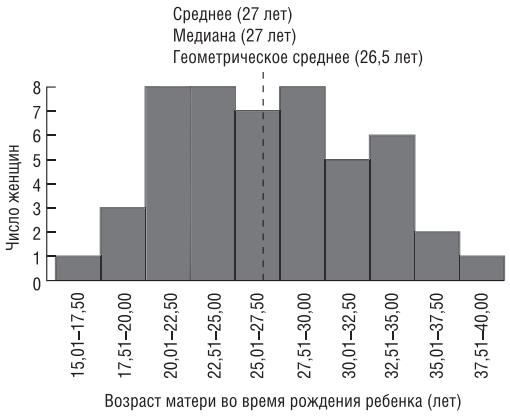



**Рис. 5.1.** Средняя, медиана и геометрическое среднее возраста женщин в исследовании, описанном в главе 2, во время рождения ребенка. Распределение возраста довольно симметрично, поскольку все три меры положения дают близкие значения, обозначенные на графике пунктирной линией

[1] В последние 10–15 лет для обозначения выборочного арифметического среднего вместо символа $\overline{x}$ чаще используют латинскую букву M. Тогда как генеральное (популяционное) среднее обозначают греческой буквой μ (читается: мю). Обращаем также внимание на то, что в статистике принято генеральные параметры распределений обозначать греческими буквами, а соответствующие выборочные параметры — соответствующими аналогами латинского алфавита. Например, генеральное стандартное отклонение обозначается греческой буквой σ (читается: сигма), а выборочное стандартное отклонение латинской буквой S. — *Прим. ред.*

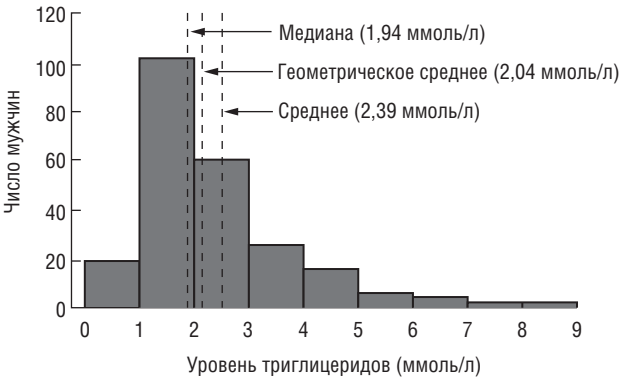



**Рис. 5.2.** Средняя, медиана и геометрическое среднее уровня триглицеридов в выборке из 232 мужчин с развившимся сердечным заболеванием (глава 19). Распределение уровня триглицеридов скошено вправо, среднее (арифметическое) дает более высокое значение меры положения, чем медиана или геометрическое среднее

## Среднее геометрическое

В случае, если наши данные имеют несимметричное распределение, среднее арифметическое не будет являться обобщающим показателем такого распределения. Если данные скошены вправо, мы можем создать распределение, которое будет более симметричным, если мы возьмем логарифм (по основанию 10 или по основанию **e**) каждого значения переменной в наборе данных (глава 9). Среднее арифметическое значений этих логарифмов — это характеристика распределения для преобразованных данных. Чтобы получить меру, которая будет иметь те же самые единицы измерения, как и первоначальные наблюдения, мы должны осуществить обратное преобразование — потенцирование (то есть взять антилогарифм) средней логарифмированных данных; мы называем такую величину **среднее геометрическое**. При условии, что распределение данных логарифма приблизительно симметричное, среднее геометрическое подобно медиане и меньше, чем среднее необработанных данных (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 вес $w_i$ к каждому значению $x_i$ в нашей выборке, для того чтобы учесть эту важность. Если значения $x_1$, $x_2$, $x_3$,..., $x_n$ имеют соответствующий вес $w_1$, $w_2$, $w_3$,..., $w_n$, взвешенное арифметическое среднее выглядит следующим образом:

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

Предположим, что мы заинтересованы в определении средней продолжительности пребывания госпитализированных больных в каком-либо районе и мы 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 средней | Преимущества | Недостатки |
|---|---|---|
| Среднее | Используются все значения набора данных.<br>Определяется математически выполнимым алгебраическим выражением.<br>Известно выборочное распределение (см. главу 9) | Искажено выбросами.<br>Искажается асимметричными данными |
| Медиана | Не искажается выбросами.<br>Не искажается асимметричными данными | Игнорирует большую часть информации.<br>Не определяется алгебраически.<br>Усложняется в выборочном распределении |
| Мода | Легко определяется для категориальных данных | Игнорирует большую часть информации.<br>Не определяется алгебраически.<br>Неизвестно выборочное распределение |
| Среднее геометрическое | 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.<br>Приписывает соответствующий вес каждому наблюдению.<br>Алгебраически определяется | Веса должны быть известны или оценены |

# Описание данных: «меры рассеяния» 6

**Цели изучения**

К концу этой главы вы должны овладеть следующими знаниями.

- Определение следующих терминов: «процентиль», «дециль», «квартиль», «медиана» и объяснение их взаимоотношения.
- Объяснение, что обозначается описываемым интервалом/размахом, а также нормальным диапазоном.
- Определение следующих параметров рассеяния: размах (диапазон), междецильный размах, дисперсия, SD, коэффициент вариации.
- Перечисление преимуществ и недостатков различных параметров рассеяния (вариабельности).
- Распознавание внутри- и междуобъектной вариабельности.

## Обобщение данных

Если мы сможем кратко изложить две меры непрерывной переменной, одна из которых показывает средние данные, а другая описывает рассеяние наблюдений, то мы значительным образом сконцентрируем наши данные. Мы объяснили, как выбрать соответствующую меру положения в главе 5. В этой главе мы обсудим самые обычные способы описания меры **рассеяния** (**разброса** или **вариабельности**), сравнение которых см. в табл. 6.1.

## Размах (интервал изменения)

Размах — это разность между максимальным и минимальным значениями переменной в наборе данных; вы найдете эти две величины, на которые ссылаются вместо их разности. Обратите внимание, что данный размах вводит в заблуждение, если одно из этих значений есть выброс (глава 3).

## Размах, полученный из процентилей

### Что такое процентили?

Предположим, что мы расположим наши данные упорядоченно, начиная с самой маленькой величины переменной $X$ и заканчивая самой большой величиной. Величина $X$, до которой расположен 1% наблюдений, находящихся ниже $X$ (и 99% наблюдений, находящихся выше этого значения $X$), называется первый **процентиль**. Величина $X$, до которой находится 2% наблюдений, называется второй процентиль и т.д. Величины $X$, которые делят упорядоченный набор значений на 10 равных групп, то есть 10-й, 20-й, 30-й, ... 90-й процентиль, называются **децили**. Величины $X$, которые делят упорядоченный набор значений на 4 равные группы, то есть 25-й, 50-й и 75-й процентиль, называется **квартили**. 50-й процентиль — это **медиана** (глава 5).

### Применение процентилей

Мы можем добиться такой формы описания рассеяния, на которую не повлияет **выброс (аномальное значение)**, при этом исключая экстремальные величины в наборе данных и определяя размах остающихся наблюдений. **Межквартильный размах** — это разница между первым и третьим квартилем, то есть между 25-м и 75-м процентилем (рис. 6.1). В него входят центральные 50% наблюдений в упорядоченном наборе, где 25% наблюдений находятся ниже центральной точки и 25% — выше. **Интердецильный размах** содержит в себе центральные 80% наблюдений, то есть те наблюдения, которые располагаются между 10-м и 90-м процентилями. Мы часто используем размах, который содержит 95% наблюдений, то есть он исключает 2,5% наблюдений снизу и 2,5% сверху (рис. 6.1). Мы можем применить этот интервал при диагностике болезни. В этом случае он называется **референтный интервал, референтный размах** или **нормальный размах** (см. главу 38).

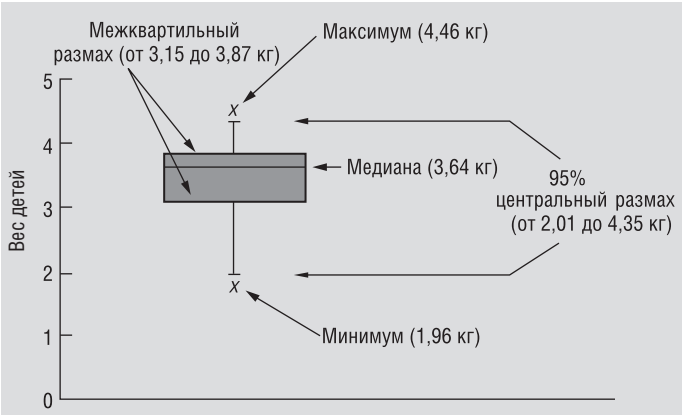



**Рис. 6.1.** График «ящик с усами» — вес ребенка при рождении (глава 2). На этом рисунке отображена медиана, межквартильный размах, размах, который содержит центральные 95% наблюдений, и максимальная и минимальная величины


**Рис. 6.2.** Диаграмма показывает рассеяние выбранных значений возраста матери во время рождения ребенка (глава 2) вокруг средней величины. Вариация вычисляется путем суммирования разностей от каждой точки до среднего значения, возведенных в квадрат, и делением этой суммы на ($n$–1)

## Дисперсия

Один из способов измерения рассеяния данных заключается в том, чтобы определить степень отклонения каждого наблюдения от средней арифметической. Очевидно, чем больше отклонение, тем больше изменчивость, вариабельность наблюдений. Однако мы не можем использовать среднее этих отклонений как меру рассеяния, потому что положительные отклонения компенсируют отрицательные отклонения (их сумма тождественно равна нулю). Для того чтобы решить эту проблему, мы возводим в квадрат каждое отклонение и находим среднее возведенных в квадрат отклонений (рис. 6.2); эта величина называется **вариацией**, или **дисперсией**. Возьмем, например, $n$ наблюдений, $x_1$, $x_2$, $x_3$,...$x_n$, средняя которых равняется $\bar{x} = (\sum x_i)/n$. Мы вычисляем дисперсию, обычно обозначаемую как $s^2$, этих наблюдений следующим образом:

$$s^2 = \frac{\sum (x_i - \overline{x})^2}{n-1}.$$

Мы видим, что это не одно и то же, что и среднее арифметическое возведенных в квадрат отклонений, потому что мы делим на ($n$–1) вместо $n$. Причина этого состоит в том, что мы почти всегда полагаемся на *выборочные* данные в своих исследованиях (глава 10). Теоретически можно показать, что мы получим более точную дисперсию, если разделим не на $n$, а на ($n$ – 1).

Единицы измерения (размерность) вариации — это квадрат единиц измерения первоначальных наблюдений, например: если переменная вес измеряется в кг, то единицей измерения вариации будет кг$^2$.

## Стандартное отклонение

**Стандартное (среднеквадратичное) отклонение** — это положительный квадратный корень из дисперсии. На примере $n$ наблюдений это выглядит так:

$$s = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n-1}}.$$

Мы можем размышлять о стандартном отклонении, как о своего рода среднем отклонении наблюдений от среднего. Оно вычисляется в тех же единицах (размерностях), что и исходные данные.

Если разделить стандартное отклонение на среднее арифметическое и выразить этот показатель в процентах, мы получим **коэффициент вариации**. Он является мерой рассеяния, который не зависит от единиц измерения (безразмерный), но имеет некоторые теоретические неудобства и поэтому не очень одобряется статистиками[1].

## Вариация в пределах и между субъектами

Если мы проведем повторные измерения непрерывной переменной у одного и того же пациента, то мы увидим, какие изменения происходят в ответах этого индивидуума (**внутрисубъектные** изменения). Это может происходить по той причине, что данный индивидуум не всегда дает точные и те же самые ответы и/или из-за ошибки измерения (см. глава 39). Однако если проводить измерения в пределах одного пациента, то вариация обычно меньше, чем вариация, если проводить единичное измерение на каждом индивидууме в группе (**межсубъектные** изменения). Например, вместимость легкого 17-летнего мальчика между 3,60 и 3,87 литрами при повторных измерениях не менее 10 раз; если же проводить однократные измерения на 10 мальчиках того же возраста, то объем будет варьировать между 2,98 и 4,33 литрами. Эти концепции важны при планировании исследования (глава 13).

[1] Основным преимуществом коэффициента вариации $C_v$ является его безразмерность. Это позволяет, вычислив величину $C_v$ для признаков с различными размерностями (см, кг, мм рт.ст., ммоль/л и т.д.), сравнить масштабы разброса значений этих признаков. Например, признак с размерностью [см] имеет значение $C_v$ = 20%, а признак с размерностью [мм рт.ст.] имеет значение $C_v$ = 90%. — *Прим. ред.*

**Таблица 6.1.** Преимущества и недостатки мер рассеяния

| Мера рассеяния | Преимущества | Недостатки |
|---|---|---|
| Размах | Легко определить | Использует даже два наблюдения.<br>Искажается выбросами.<br>Имеет тенденцию к увеличению при росте объема выборки |
| Размахи, основанные на процентилях | Не подвержены влиянию выбросов.<br>Не зависят от размера выборки.<br>Пригодны для асимметричных распределений данных | Грубый расчет.<br>Невозможно рассчитать для маленьких выборок.<br>Может использоваться даже при двух наблюдениях.<br>Не имеет алгебраического выражения для вычисления |
| Дисперсия | Использует каждое наблюдение.<br>Имеет алгебраическое выражение для вычисления | Размерность величины — квадрат размерности исходных данных.<br>Чувствительна к выбросам.<br>Не подходит для асимметричных распределений данных |
| Стандартное отклонение | Те же самые преимущества, что и у дисперсии.<br>Единицы измерения те же, что и у исходных данных.<br>Легко интерпретируемо | Чувствительно к выбросам.<br>Не подходит для асимметричных распределений данных |

# Теоретические распределения: нормальное распределение 7

**Цели изучения**

К концу этой главы вы должны овладеть следующими знаниями.

- Определение терминов: «вероятность», «условная вероятность».
- Распознавание различия между субъективным, частотным и априорным подходами к вычислению вероятности.
- Определение сложения и правила умножения вероятности.
- Определение терминов «случайная переменная», «распределение вероятности», «параметр», «статистический параметр», «показатель», «функция плотности вероятности».
- Различие дискретного и непрерывного распределения вероятности и перечисление свойства каждого из этих распределений вероятности.
- Перечисление свойств нормального и стандартного нормального распределения вероятности.
- Определение стандартизированного нормального отклонения (SND).

В главе 4 мы показали, как создать **эмпирическое распределение частоты** исследуемых данных. Оно контрастирует с теоретическим **распределением вероятности**, которое можно описать при помощи математической модели. Когда наше эмпирическое распределение аппроксимирует некоторое распределение вероятности, мы можем применить теоретические знания об этом распределении, для того чтобы ответить на вопросы, касающиеся данных. Часто это требуется для оценки вероятностей.

## Понимание вероятности

**Вероятность** измеряет неопределенность. Она находится в самом центре статистической теории. Вероятность измеряет возможность появления данного события. Это положительное число, которое находится в интервале между 0 и 1. Если она равна нулю, никакого события и *быть не может*. Если она равна единице, тогда *событие должно обязательно произойти*. Вероятность **дополнительного** события (события *непроисходящего*) равна единице минус вероятность появления события. Мы рассмотрим **условную вероятность** как вероятность события с учетом того, что другое событие произошло, в главе 45.

Мы можем вычислить вероятность, используя различные подходы.

- **Субъективная** — индивидуальная степень уверенности, что данное событие произойдет (например, что случится конец света в 2050 году).
- **Частотная** — соотношение количества событий, которые могли бы произойти, если бы мы повторяли эксперимент огромное количество раз (например, если бы мы бросали монету 1000 раз, сколько бы раз выпал «орел»).
- **Априорная** — требует знания теоретической *модели*, называемой **распределением вероятности**, которая отображает вероятности всех возможных результатов эксперимента. Например, генетическая теория позволяет нам отобразить вероятность распределения цвета глаз у ребенка при рождении, если у женщины голубые глаза, а у мужчины карие, первоначально определяя весь возможный генотип цвета глаз у ребенка и их вероятности.

## Правила вероятности

Мы можем применять правила вероятности, для того чтобы складывать и умножать вероятности.

- **Правило сложения** — если два события, A и B, *взаимоисключающие, несовместимые* (то есть каждое событие исключает другое), вероятность того, что произойдет то или иное событие, равна сумме их вероятностей:

$$\text{Prob (A или B)} = \text{Prob (A)} + \text{Prob (B)}.$$

Например, у взрослого больного в обычной зубоврачебной практике нет отсутствующих зубов, некоторые зубы отсутствуют или он беззубый (то есть нет зубов), вероятности равны 0,67, 0,24 и 0,09 соответственно, тогда вероятность того, что у больного есть несколько зубов, равна 0,67+0,24=0,91.

- **Правило умножения** — если два события, A и B, *независимы* (то есть возникновение одного события не влияет на возможность появления другого), вероятность того, что оба события произойдут, равна произведению вероятности каждого:

$$\text{Prob (A и B)} = \text{Prob (A)} \times \text{Prob (B)}.$$

Например, если два, не имеющих отношения друг к другу, больных ожидают в кабинете хирургической стоматологии, вероятность того, что у обоих больных нет отсутствующих зубов, равна $0{,}67 \times 0{,}67 = 0{,}45$.

## Распределения вероятности: теория

**Случайная переменная** — это величина, которая может принимать любое из набора взаимоисключающих значений с определенной вероятностью. **Распределение вероятности** показывает вероятности всех возможных значений случайной переменной. Это теоретическое распределение, которое выражено математически и имеет среднее и дисперсию, которые являются аналогами среднего и дисперсии в эмпирическом распределении. Каждое распределение вероятности определяется определенными параметрами, которые являются обобщающими величинами (например, среднее, дисперсия), характеризующими данное распределение (то есть знание их позволит подробно описать распределение). При помощи соответствующей **статистики** можно оценить эти параметры в выборке. В зависимости от того, является ли случайная переменная дискретной или непрерывной, распределение вероятности может быть либо дискретным, либо непрерывным.

- **Дискретное** (например, биноминальное, распределение Пуассона) — мы можем получить вероятности, соответствующие **каждому возможному значению** случайной переменной. **Сумма всех таких вероятностей равняется единице**.
- **Непрерывное** (например, нормальное, $\chi^2$, $t$ и $F$) — мы можем получить вероятность случайной переменной $x$, только **принимающей значения в определенных интервалах** (потому что существует бесконечное множество значений $x$). Если горизонтальная ось изображает значения $x$, мы можем начертить кривую из уравнения распределения (**функция плотности распределения вероятности**); она имеет сходство с эмпирическим относительным частотным распределением (глава 4). **Общая площадь под кривой равняется единице**; эта площадь отражает вероятность всех возможных событий. Вероятность того, что $x$ находится между двумя граничными значениями, равна площади под кривой между этими значениями (рис. 7.1). Для удобства была создана таблица (приложение A), для того чтобы предоставить нам возможность оценить рассматриваемую вероятность для обычно используемых непрерывных распределений вероятности. Особенно они важны применительно к построению доверительных интервалов (глава 11) и проверке гипотез (глава 17).

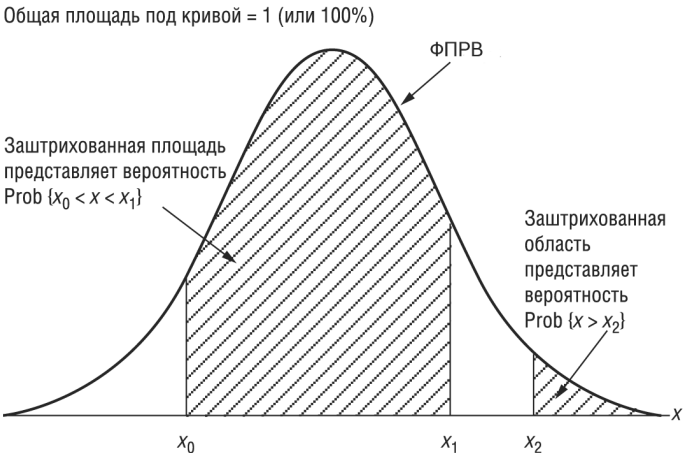



**Рис. 7.1.** Функция плотности распределения вероятности[1], ФПРВ, $x$

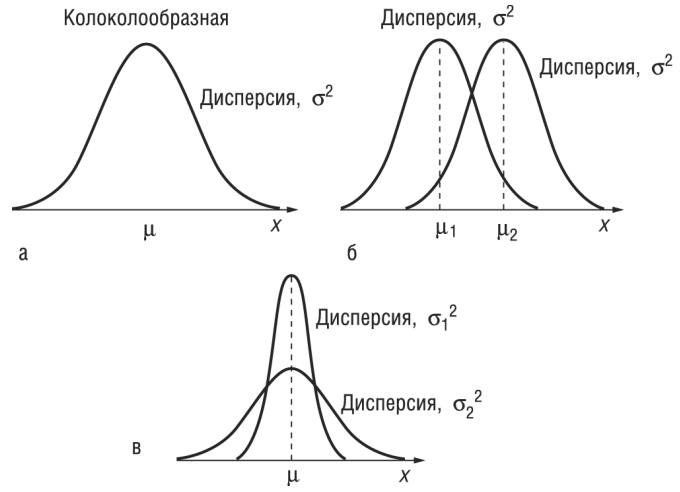



**Рис. 7.2.** Функция плотности распределения вероятности нормального распределения переменной, $x$: а — симметричное относительно среднего μ, дисперсия $\sigma^2$; б — результат изменения среднего ($\mu_2 > \mu_1$); в — результат изменения дисперсии ($\sigma_1^2 < \sigma_2^2$)

## Стандартное нормальное распределение

Одним из самых важных распределений в статистике является **нормальное распределение**. Его функция плотности распределения вероятности (рис. 7.2):

- полностью определяется двумя параметрами: среднее (μ) и дисперсия ($\sigma^2$);
- колоколообразная (унимодальная);
- симметричная относительно среднего;
- сдвигается вправо, если среднее увеличивается, и влево, если среднее уменьшается (при постоянной дисперсии);
- сплющивается, если дисперсия увеличивается, но становится более остроконечной, если дисперсия уменьшается (для постоянного среднего).

[1] На всех рисунках этой главы отсутствует вертикальная ось $Y$ — функция плотности распределения вероятности (ФПРВ), обозначаемая в русскоязычной литературе как $f(x)$.

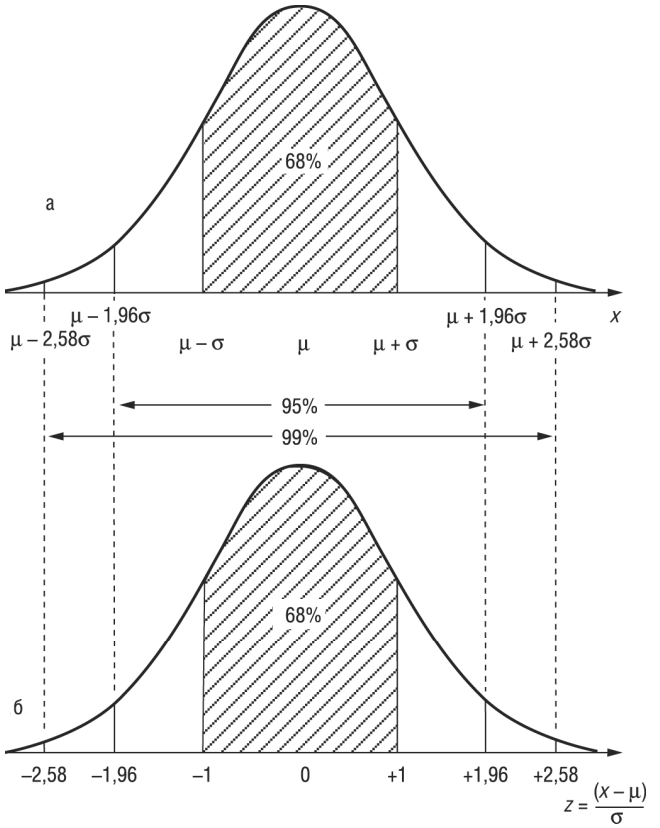



**Рис. 7.3.** Площади (проценты общей вероятности) под кривой для (а) нормального распределения $x$, со средним μ и дисперсией $\sigma^2$ и (б) стандартного нормального распределения $z$

Дополнительные свойства:

- среднее и медиана нормального распределения равны;
- вероятность (рис. 7.3, а) того, что нормально распределенная случайная переменная $x$ со средним μ и стандартным отклонением σ, находится между:
  - (μ − σ) и (μ + σ) равна 0,68;
  - (μ − 1,96σ) и (μ + 1,96σ) равна 0,95;
  - (μ − 2,58σ) и (μ + 2,58σ) равна 0,99.

Эти интервалы можно использовать для определения **референтных интервалов** (главы 6 и 38).

В главе 35 мы покажем, как оценить нормальность.

## Стандартное нормальное распределение

Существует бесконечно много (семейство) нормальных распределений в зависимости от значений μ и σ. Стандартное нормальное распределение — это особое нормальное распределение, вероятности для которого приведены в таблице (приложения A1, A4).

Стандартное нормальное распределение имеет среднее, равное 0, и дисперсию, равную 1.

Если случайная переменная $x$ имеет нормальное распределение со средним μ и дисперсией $\sigma^2$, тогда стандартизованное нормальное отклонение (СНО) $z = \frac{x-\mu}{\sigma}$ будет являться случайной переменной, которая имеет стандартное нормальное распределение.

# Теоретические распределения: другие распределения 8

**Цели изучения**

К концу этой главы вы должны овладеть следующими знаниями.

- Перечисление важнейших свойств *t*-распределения Стьюдента, $\chi^2$-распределения Пирсона, *F*-распределения Фишера–Снедекора и логнормального распределения вероятностей.
- Объяснение, когда каждое из этих распределений вероятности наиболее полезно.
- Перечисление важнейших свойств биномиального распределения и распределения Пуассона.
- Объяснение, когда наиболее полезны биномиальное распределение и распределение Пуассона.

## Несколько слов для поддержки духа

Не стоит волноваться, если теория, лежащая в основе распределения вероятности, вам покажется сложной. Наш опыт показывает, что единственное, что вы хотели бы знать, — это когда и как применять эти распределения. Поэтому мы изложили основы и опустили уравнения, которые характеризуют распределения вероятности. И вы поймете, что единственное, что вам нужно, — это освоить основные понятия, терминологию и, возможно, знать, как работать с таблицами.

## Непрерывное распределение вероятностей

Эти распределения основаны на непрерывных случайных переменных. Часто это не непосредственно измеряемая переменная, которая отвечает такому распределению, а параметр, **статистика**, полученная из этой переменной. Общая площадь под кривой функции плотности распределения вероятности есть сумма вероятностей всех возможных значений, и она равна 1 (глава 7). Мы рассмотрели нормальное распределение в главе 7; в этой главе мы опишем другие наиболее известные распределения.

### *t*-распределение (приложение A2, рис. 8.1)

- Получено Вильямом Госсетом, который публиковался под псевдонимом Student (Студент)[1], поэтому его часто называют ***t*-распределением Стьюдента**.

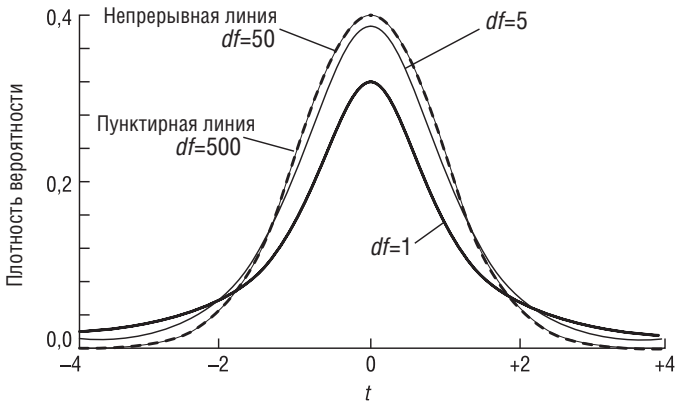



**Рис. 8.1.** *t*-распределение со степенями свободы (*df*)=1; 5; 50 и 500

[1] Статья была опубликованна в 1908 г. в журнале Biometrika (см. http://www.biometrica.tomsk.ru/student_1908_1.pdf и http://www.biometrica.tomsk.ru/student_1908_2.pdf). — *Прим. ред.*

- Параметрами, которые характеризуют *t*-распределение, являются **степени свободы** (*df*), так, мы сможем начертить функцию плотности распределения вероятности только в том случае, если мы будем знать уравнение *t*-распределения и степени свободы. В главе 11 мы рассмотрим степени свободы; обратите внимание, что они часто выражаются через объем выборки.
- Форма подобна такой же для стандартизованного нормального распределения, но более приплюснута и с более длинными хвостами. Форма приближается к нормальной кривой, по мере того как увеличиваются степени свободы.
- В частности, его применяют для вычисления доверительных интервалов и исследования гипотез с одной или двумя средними (главы 19–21).

### Хи-квадрат ($\chi^2$) распределения Пирсона (приложение A3, рис. 8.2)

- Является скошенным вправо распределением, принимающим только положительные значения.
- Характеризуется **степенями свободы** (глава 11).
- Форма зависит от числа степеней свободы; становится более симметричным и приближается к нормальному с их ростом.
- Особенно часто используется для анализа категориальных данных (главы 23–25).

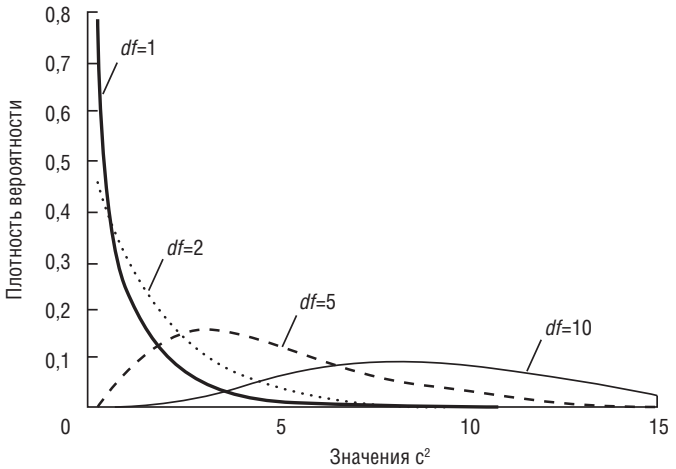



**Рис. 8.2.** $\chi^2$ распределения Пирсона со степенями свободы (*df*)=1, 2, 5 и 10

### F-распределение (приложение A5)

- Является скошенным вправо.
- Определяется как отношение. Распределение отношения двух оценок дисперсий, вычисленных для нормально распределенных данных, аппроксимируются *F*-распределением.
- Два параметра, которые характеризуют его, — степени свободы (глава 11) числителя и знаменателя отношения.
- *F*-распределение особенно полезно для сравнения двух дисперсий (глава 35) и более чем двух средних при использовании дисперсионного анализа (ANOVA) (глава 22).

### Логнормальное распределение

- Это распределение вероятности случайной переменной, логарифм которого (по основанию 10 или *e* — основание натурального логарифма) имеет нормальное распределение.
- Сильно скошено вправо (рис. 8.3, а).
- Если мы возьмем логарифмы наших исходных данных, которые скошены вправо, мы создадим эмпирическое распределение,

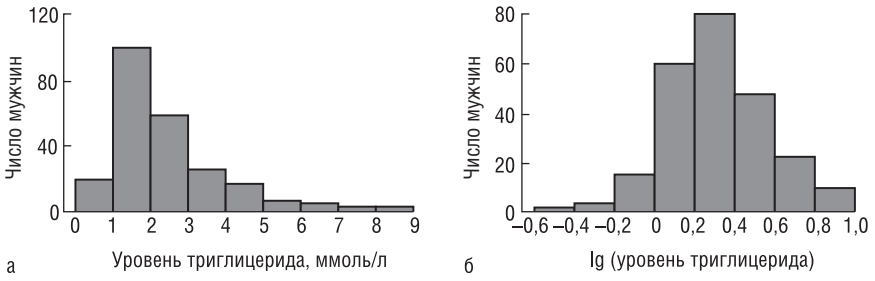



**Рис. 8.3.** Логнормальное распределение уровня триглицерида у 232 мужчин с заболеваниями сердца (глава 19) (а). Почти нормальное распределение $\log_{10}$ (уровень триглицерида) (б)

которое почти нормальное (рис. 8.3, б), и тогда наши данные соответствуют приближенно логнормальному распределению.

- Многие переменные в медицине имеют логнормальное распределение. Мы можем использовать свойства нормального распределения (глава 7), для того чтобы сделать выводы этих переменных после логарифмического преобразования данных.
- Если набор данных имеет логнормальное распределение, мы используем среднее геометрическое (глава 5) как обобщающий показатель положения.

## Дискретные распределения вероятностей

Случайная переменная, которая имеет такое распределение вероятности, является дискретной. Сумма вероятностей всех возможных взаимоисключающих событий равна 1.

### Биноминальное распределение

- Предположим, в данной ситуации существует только два результата — «успех» и «неудача». Например, нас интересует, забеременеет или нет женщина при экстракорпоральном оплодотворении (IVF). Если мы посмотрим на $n$=100, не имеющих отношение друг к другу женщин, подвергающихся IVF (каждая с той же вероятностью беременности), то биноминальная случайная переменная — это наблюдаемое количество зачатий. Часто это понятие объясняется в терминах n независимых повторных испытаний (например, 100 раз подбросить монету), при которых результатом будет являться либо успех (например, орел), либо неудача.
- Два параметра, которые описывают биноминальное распределение, — это $n$ — количество индивидуумов в выборке (или повторения испытания) и $p$ — точная вероятность успеха для каждого индивидуума (или при каждом испытании).
- Среднее (значение для случайной переменной, которое мы ожидаем, если мы осматриваем $n$ индивидуумов или повторим испытание $n$ раз) — это $np$. Дисперсия — $np(1-p)$.
- Когда $n$ невелико, распределение будет скошено вправо, если $p<0,5$, и влево, если $p>0,5$. Распределение становится более симметричным, по мере того как объем выборки будет увеличиваться (рис. 8.4), и приблизится к нормальному распределению в том случае, если $np$ и $n(1-p)$—>5.
- Мы можем использовать свойства биномиального распределения, для того чтобы сделать выводы относительно **пропорций**. Особенно часто мы используем аппроксимацию биномиального распределения с помощью нормального распределения при анализе пропорций (долей).

### Распределение Пуассона

- Пуассоновская случайная переменная — это **число** событий, которые происходят независимо и случайно во времени или пространстве со средней интенсивностью **μ**. Например, количество госпитализаций в день типично отвечает распределению Пуассона. Мы используем распределения Пуассона, для того чтобы вычислить вероятность конкретного количества госпитализаций в любой отдельный день.
- Параметр, которым описывают распределение Пуассона, — это **среднее**, то есть средняя интенсивность, **μ**.
- **Среднее** равняется **дисперсии** в распределении Пуассона.
- Если среднее маленькое, то распределение будет скошено вправо и будет становиться более симметричным; по мере того как среднее будет увеличиваться, оно приближается к нормальному распределению.

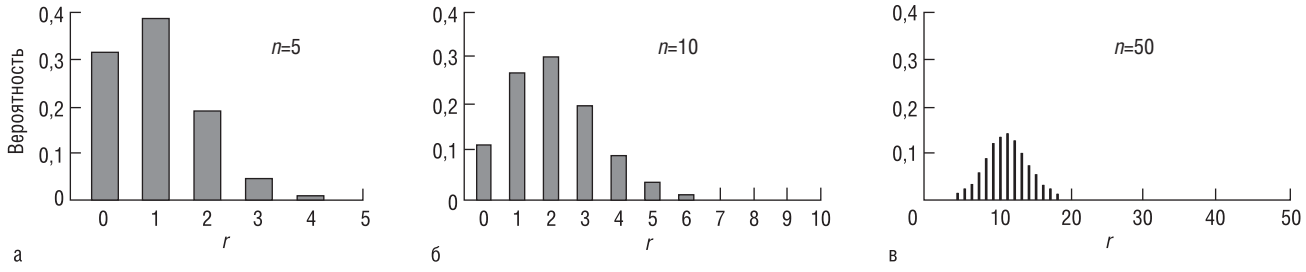



**Рис. 8.4.** Биноминальное распределение показывает количество успехов $r$, когда вероятность успеха $p$=0,20 для объема выборки $n$=5 (а), $n$=10 (б) и $n$=50 (в) (NB! В главе 23 наблюдаемая серораспространенность HHV-8 была $p$=0,185≈0,2 и объем выборки был 271: для пропорции было использовано приближение нормальным распределением)

# Преобразования 9

**Цели изучения**

К концу этой главы вы должны овладеть следующими знаниями.

- Описание ситуаций, в которых преобразование данных может быть полезным.
- Объяснение, как преобразовать набор данных.
- Объяснение, когда применять и что достигается логарифмированием, извлечением квадратного корня, получением обратной величины, возведением в квадрат и логит-преобразованием.
- Описание, как интерпретировать итоговые результаты, полученные с помощью логарифмирования данных, после их обращения к оригинальной шкале.

## Зачем преобразовывать?

В нашем исследовании наблюдения могут не подчиняться требованиям предполагаемого статистического анализа (глава 35).

- Переменная может иметь ненормальное распределение — требование, обязательное для множества различных параметрических методов анализа (например, *t*-критерий Стьюдента, ANOVA и т.д.).
- Рассеяния (размах) признака в наблюдениях каждой из групп ряда могут быть разными, но равенство дисперсий — это необходимое условие корректного использования **параметра** при сравнении средних с помощью *t*-критерия Стьюдента и классического дисперсионного анализа (ANOVA) (главы 21–22).
- Две переменные не могут быть линейно зависимы (**линейность** — это требование во многих видах регрессионного анализа, см. главы 27–33 и 42).

Очень часто полезно **преобразовывать** данные, для того чтобы удовлетворить требованиям, которые лежат в основе предлагаемых статистических методов[1].

## Как мы преобразовываем?

Мы превращаем исходные данные в преобразованные, используя одно и то же математическое преобразование для каждого наблюдения. Предположим, что у нас есть $n$ наблюдения ($y_1, y_2, ..., y_n$) с переменной $y$ и мы принимаем решение, что нам подходит логарифмическое преобразование. Мы берем логарифм каждого наблюдения, для того чтобы образовать ($\log y_1, \log y_2, ..., \log y_n$). Если мы назовем преобразованную переменную $z$, тогда $z = \log y_i$ для каждой $i$ ($i = 1, 2, ..., n$) и наше преобразование данных можно записать ($z_1, z_2, ... z_n$).

Мы проверяем, достигло ли это преобразование своей цели при создании набора данных, который удовлетворяет предположениям запланированного статистического анализа, и переходим к анализу преобразования данных ($z_1, z_2, ..., z_n$). Мы часто делаем обратные преобразования обобщающих мер (таких, как среднее) до первоначальной размерности; те выводы, которые мы сделали из гипотез по нашим тестам (глава 17) на преобразованных данных, применимы для исходных данных[2].

[1] При этом важно помнить, что результат применения статистического критерия к преобразованной переменной нельзя априорно переносить на исходную переменную. Далее преобразование может изменить и размерность числовой переменной, что затрудняет смысловую интерпретацию новой переменной. Например, систолическое давление имеет размерность (мм рт.ст.). После возведения в квадрат этой переменной размерность станет (мм рт.ст.)$^2$. Каков физический смысл это новой переменной? — *Прим. ред.*

[2] Такое согласие наблюдается далеко не всегда, и потому оно требует дополнительной проверки. — *Прим. ред.*

## Типичные преобразования

### Логарифмическое преобразование, $z = \log(y)$

При логарифмическом преобразовании данных мы можем выбирать, взять логарифмы по основанию 10 ($\log_{10}(y)$ — десятичный логарифм) или по основанию $e$ ($\log_e(y) = 1n(y)$ — натуральный или неперовский логарифм), но они должны быть одинаковы для отдельной переменной в наборе данных. Обратите внимание, что мы не можем брать логарифм отрицательного числа или нуля. Обратное преобразование логарифма (потенцирование) называется антилогарифмом; антилогарифм неперовского логарифма есть экспонента — $e$.

- Если распределение $y$ скошено вправо, преобразование $z = \log(y)$ часто дает в результате приближенно **нормальное распределение** (рис. 9.1, а). Тогда $y$ имеет логнормальное распределение (глава 8).
- Если существует экспоненциальное соотношение между $y$ и другой переменной $x$, такое, что конец кривой загибается вверх, в то время как $y$ (по вертикальной оси) наносится соответственно $x$ (по горизонтальной оси), то соотношение между $z = \log(y)$ и $x$ приблизительно **линейное** (рис. 9.1, б).
- Предположим, что у нас есть различные группы наблюдений, причем включая измерения непрерывной переменной $y$. Мы можем обнаружить, что группы, которые имеют более высокие значения $y$, имеют и большие вариации. В частности, если коэффициент вариации (стандартное отклонение, деленное на среднее), постоянен для всех групп, преобразование логарифма $z = \log(y)$ создает группы с **равными дисперсиями** (рис. 9.1, в).

В медицине часто применяется логарифмическое преобразование из-за логической интерпретации и потому, что многие переменные имеют скошенное вправо распределение. Например, если исходные данные подвергнуть логарифмическому преобразованию, тогда разница между двумя средними величинами на логарифмической шкале равна отношению двух соответствующих средних в исходной шкале. Антилогарифмы границ 95% доверительного интервала (глава 11) для среднего значения логарифмически преобразованных данных дают границы 95% доверительного интервала для геометрического среднего. Если мы используем для независимых переменных (предикторов) в регрессионном анализе (глава 29) логарифмирование по основанию 10, увеличение переменной в логарифмической шкале на единицу представляет собой 10-кратное увеличение переменной в исходной, оригинальной шкале. Обратите внимание, что логарифмическое преобразование в регрессии зависимой, результирующей переменной дает возможность обратного преобразования регрессионных коэффициентов и мультипликативного эффекта, а не суммирования эффекта в оригинальной, исходной шкале (см. главы 30 и 31).

### Преобразование квадратного корня $z = \sqrt{y}$

Это преобразование имеет свойства такие же, как и логпреобразование, хотя результаты, после того как они были преобразованы обратно, объяснить сложнее. В дополнение к способностям **нормализации** и **линеаризации**, эффективно использовать преобразование, **стабилизирующее дисперсию**, если дисперсия возрастает при увеличении значений $y$, то есть тогда дисперсия, деленная на среднее, постоянна. Мы применяем преобразование квадратного корня, если $y$ — это количество редких явлений, встречающихся во времени и пространстве, то есть это пуассоновская переменная (глава 8). Помните, что мы не можем извлечь квадратный корень из отрицательного числа.

### Обратное преобразование $z = 1/y$

Мы часто применяем обратное преобразование к периодам жизни (выживаемости), если не используем специальные методы для анализа выживаемости (глава 44). Обратное преобразование имеет

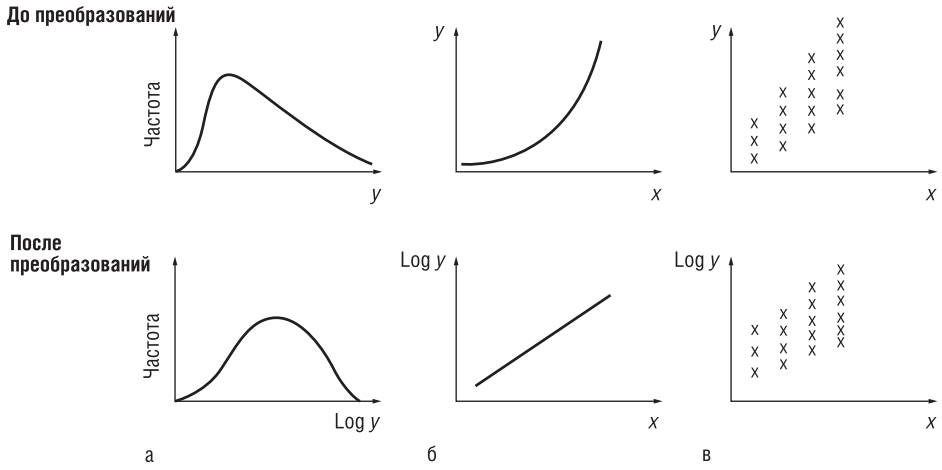



**Рис. 9.1.** Свойства логарифмического преобразования: нормализация (а), линеаризация (б), выравнивание дисперсий (в)

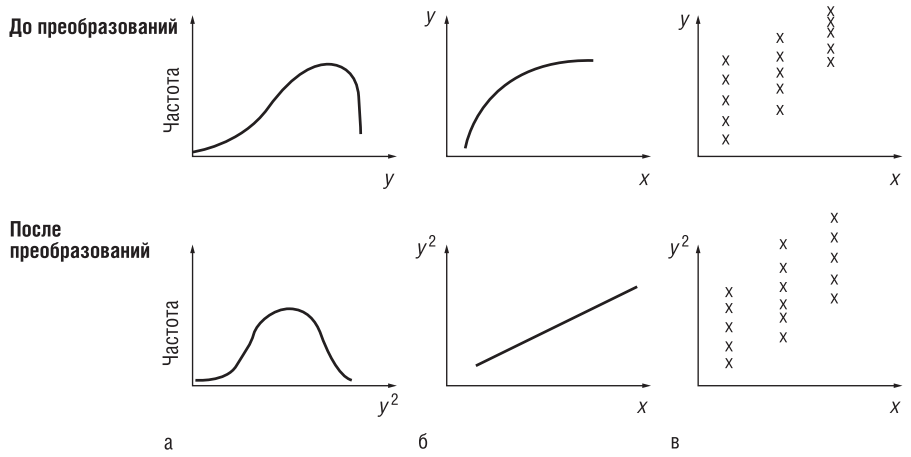



**Рис. 9.2.** Эффект квадратичного преобразования: нормализация (а), линеаризация (б), выравнивание дисперсий (в)

такие же свойства, как и лог-преобразование. В дополнение к способностям **нормализации** и **линеаризации**, оно гораздо эффективнее для **стабилизации дисперсии**, чем логарифмическое преобразование, если дисперсия очень заметно увеличивается при увеличении значений $y$, то есть для стабилизации частного от деления дисперсии на среднее. Заметьте, что мы не можем делить на нуль.

## Квадратичное преобразование $z = y^2$

Квадратичное преобразование достигает результат, обратный лог-преобразованию.

- Если $y$ скошено влево, то распределение $z = y^2$ часто является приближенно **нормальным** (рис. 9.2, а).
- Если соотношение между двумя переменными $x$ и $y$ такое, что, когда мы наносим $y$ против $x$, кривая загибается вниз, тогда соотношение между $z = y^2$ и $x$ почти **линейное** (рис. 9.2, б).
- Если дисперсия непрерывной переменной $y$ уменьшается, по мере того как значение $y$ увеличивается, тогда квадратное преобразование $z = y^2$ **стабилизирует дисперсию** (рис. 9.2, в).

## Логит (логистическое) преобразование $z = \ln \frac{p}{1-p}$

Это преобразование, которое мы наиболее часто применяем к каждой пропорции, доле $p$, в наборе пропорций. Мы не можем применять логистическое преобразование в том случае, если $p$=0 либо $p$=1, потому что соответствующие значения логита — это $-\infty$ и $+\infty$. Единственное решение — это взять $p$ как $1/(2n)$ вместо 0 и как $\{1-1/(2n)\}$ вместо 1. Оно линеаризует сигмовидную кривую (рис. 9.3), см. главу 30 по использованию логит-преобразования в логистической регрессии[1].

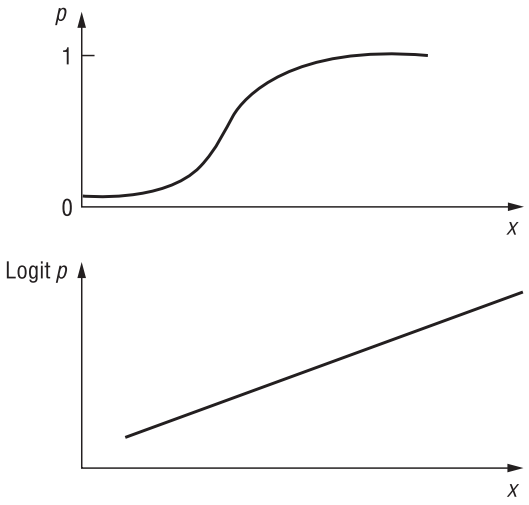



**Рис. 9.3.** Эффект логистического преобразования сигмовидной кривой

[1] Достаточно подробное описание специфики использования этого мощного метода с большим количеством реальных примеров приведено в статье «Логистическая регрессия в медицине и биологии» (см. http://www.biometrica.tomsk.ru/logit_0.htm).