Авива Петри, Кэролайн Сэбин

Наглядная МЕДИЦИНСКАЯ СТАТИСТИКА

Перевод с английского
под редакцией В.П. Леонова

3-е издание,
переработанное и дополненное

Москва
Издательская группа «ГЭОТАР-Медиа»
2015

1 Типы 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

Рис. 1.1. Схема, показывающая различные типы переменных

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».

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, FEV1) 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, т.е. 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 рассмотривать 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, т.е. 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.

2 Ввод данных

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

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 ребенка			Дата			
№ 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

1=Гемофилия А
2=Гемофилия В
3=Болезнь Виллебренда
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=1 раз в день
3=1 раз в неделю
4=1 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

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.

3 Проверка ошибок и выбросов

При любом исследовании всегда существует возможность обнаружить ошибки в наборе данных либо вначале при измерениях, либо при сборе, переписывании и вводе данных в компьютер. Довольно-таки 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 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**, т.е. 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. В случае, если отсутствует большая часть данных, и результаты анализа вряд ли будут надежны. 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. Мы можем столкнуться с определенными проблемами, когда высока вероятность того, что отсутствующие данные тесно связаны с переменной, представляющей наибольший интерес в нашем исследовании (см., например, результаты по регрессионному анализу, глава 27). В такой ситуации получаемые результаты анализа могут быть сильно искажены (глава 34). Например, предположим, что нас интересуют значения величины, которая отражает состояние здоровья пациентов, однако эта информация пропущена, не измерена для некоторых пациентов, потому что они не чувствовали себя достаточно хорошо, чтобы присутствовать в клинике на приеме у врача; в этом случае мы, скорее всего, получим излишне оптимистичное представление о здоровье пациентов, поскольку мы не учитываем в анализе отсутствующие данные. Уменьшить такое смещение результата можно с помощью подходящих статистических методов[1] или путем оценки некоторыми способами значений отсутствующих данных[2], однако наиболее предпочтительным вариантом является сведение к минимуму количества пропущенных (неизмеренных) данных в самом начале исследования.

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 перемен-

[1] *Laird N.M.* Missing data in longitudinal studies // Statistics in Medicine. — 1988. — Vol. 7. — N 305. — P. 315.

[2] *Engels J.M. and Diehr P.* Imputation of missing longitudinal data: a comparison of methods // Journal of Clinical Epidemiology. — 2003. — Vol. 56. — P. 968–976.