ГЛАВА 1

Разведочный анализ данных

Основной постулат этой книги гласит: данные можно использовать для ответа на вопросы, разрешения споров и принятия более эффективных решений.

Эта глава познакомит вас с процедурой, которой мы будем для этого придерживаться: загрузкой, проверкой и разведкой данных, а также выбором статистических показателей, измеряющих то, что нас интересует. В качестве примера воспользуемся данными Национального исследования роста семьи (National Survey of Family Growth, NSFG) США для ответа на вопрос, с которым я столкнулся, когда мы с женой ожидали нашего первенца: правда ли, что первые дети появляются на свет позже?

Данные

Возможно, вы слышали, что первые дети чаще рождаются позже срока. Поиск в интернете выдаст множество дискуссий на эту тему. Одни утверждают, что это правда, другие — что это миф, а некоторые заявляют ровно обратное: первые дети появляются на свет раньше срока.

Часто в подобных спорах оппоненты в подтверждение собственных слов ссылаются на данные. Вот несколько примеров такой аргументации:

> «У ОБЕИХ моих подруг, недавно родивших первенцев, роды проходили почти на две недели позже срока — только тогда начались схватки или было проведено стимулирование родов».
>
> «Мой первый появился с опозданием на две недели, и поэтому я думаю, что второй появится на две недели раньше!»
>
> «Я не верю этому, потому что моя сестра, которая была первым ребенком у моей матери, родилась рано — как и многие мои двоюродные братья и сестры».

Подобные сообщения называют **бытующим, или неподтвержденным, мнением**, поскольку они основаны на неопубликованных и, как правило, личных данных. Нет ничего предосудительного, когда подобные факты из жизни приводятся в бытовых разговорах, и поэтому у меня нет намерения придираться к людям, которых я только что процитировал.

Однако нам понадобятся более убедительные доказательства и ответ, заслуживающий большего доверия. Чтобы их получить, «личного опыта» обычно бывает недостаточно по следующим причинам:

Малое количество наблюдений

Если беременность у родящих впервые длится дольше, то разница, вероятно, невелика по сравнению с естественным разбросом. В этом случае нам, возможно, придется сравнить большое количество беременностей, чтобы понять, имеет ли место различие.

Предвзятость в отборе данных

Вступающие в подобные дискуссии люди могут являться заинтересованными сторонами, если их первые дети родились поздно. В этом случае процесс отбора данных может повлиять на результаты.

Предвзятость подтверждения

Люди, верящие в некоторое утверждение, чаще будут приводить примеры, подтверждающие его. И наоборот, сомневающиеся в нем с большей вероятностью приведут контрпримеры.

Неточность

Бытующее мнение часто основано на личном опыте, который может быть неверно припомнен, искажен, неточно воспроизведен и т. д.

Чтобы устранить эти недостатки, мы будем использовать статистические инструменты, среди которых можно упомянуть следующие:

Сбор данных

Мы задействуем данные крупного национального опроса, который был специально разработан с целью получения статистически обоснованных выводов о населении США.

Описательная статистика

Мы вычислим статистические показатели, кратко описывающие данные, и сравним различные способы визуализации данных.

Разведочный анализ данных

Мы будем искать закономерности, различия и другие особенности, которые отвечают на интересующие нас вопросы. Помимо этого, мы будем проверять наличие несоответствий и выявлять ограничения.

Оценка

Мы будем использовать данные некоторой выборки для оценки характеристик генеральной совокупности.

Проверка гипотез

В тех случаях, когда наблюдается некоторый эффект, например разница между двумя группами, мы будем оценивать вероятность того, что этот эффект появился случайно.

Придерживаясь этой последовательности шагов и стараясь избегать ошибок, мы сможем прийти к выводам, которые будут более обоснованными и с большей вероятностью окажутся верными.

Национальное исследование роста семьи США

С 1973 года Центры по контролю и профилактике заболеваний (Centers for Disease Control and Prevention, CDC) США проводят Национальное исследование роста семьи (NSFG), целью которого является сбор «информации о семейной жизни, браке и разводах, беременности, бесплодии, использовании средств контрацепции, а также о здоровье мужчин и женщин. Результаты обследования используются... для планирования медицинских услуг и программ санитарного просвещения, а также для проведения статистических исследований семей, рождаемости и здоровья».

Мы воспользуемся собранными в этом исследовании данными, чтобы понять, рождаются ли первенцы позже, и ответить на другие вопросы. Чтобы эффективно использовать эти данные, нужно понимать дизайн статистического исследования.

Как правило, цель статистического исследования — сделать выводы о **генеральной совокупности** (популяции). В NSFG изучаемой популяцией являются жители Соединенных Штатов в возрасте от 15 до 44 лет.

В идеальном случае исследование должно содержать данные о каждом представителе популяции, но такое редко бывает возможно. Вместо этого собирают данные о некотором подмножестве генеральной совокупности, называемом **выборкой**. Люди, участвующие в опросе, зовутся **респондентами**.

NSFG представляет собой **поперечное, или одномоментное, исследование** (cross-sectional study); это означает, что оно фиксирует моментальный снимок популяции в определенный момент времени. Опрос в рамках NSFG проходил уже несколько раз; каждое его проведение называется **циклом**. Мы будем использовать данные 6-го цикла, который проходил с января 2002 по март 2003 года.

Обычно считается, что поперечные исследования являются **репрезентативными**; это означает, что выборка похожа на целевую популяцию во всех отношениях, которые важны для целей исследования. Такого идеала трудно достичь на практике, но специалисты, проводящие опросы, стараются приблизиться к нему настолько, насколько это возможно.

Исследование NSFG является не репрезентативным, а **стратифицированным**; это означает, что в нем для некоторых групп населения намеренно используется **избыточная выборка**. Разработчики исследования для трех групп населения — испаноязычных, афроамериканцев и подростков — увеличивали долю респондентов по сравнению с тем, как они представлены в населении США. Число респондентов в каждой группе, таким образом, становилось достаточным, чтобы можно было делать обоснованные выводы. Недостатком избыточной выборки является то,

что на основе ее статистических показателей делать выводы о генеральной совокупности становится труднее. Позже мы вернемся к этому вопросу.

При работе с такого рода данными важно ознакомиться с **кодбуком** (codebook), в котором документированы структура исследования, вопросы анкеты, а также коды ответов.

Считывание данных

Перед загрузкой данных NSFG требуется согласие с условиями использования:

Любая преднамеренная идентификация или раскрытие информации о физическом или юридическом лице нарушает гарантии конфиденциальности, предоставленные поставщикам информации. Таким образом, пользователи обязуются:

- Использовать данные из этого датасета исключительно для статистической отчетности и анализа.
- Не пытаться установить личность физических или идентичность юридических лиц, представленных в этих данных.
- Не связывать этот датасет с данными, позволяющими проводить идентификацию, как из других датасетов Национального центра статистики здравоохранения (National Center for Health Statistics, NCHS) США, так и ему не принадлежащих.
- Не предпринимать никаких попыток оценить методы раскрытия информации, применяемые для защиты физических и юридических лиц, а также не проводить никаких исследований по методам повторной идентификации физических и юридических лиц.

Если вы согласны соблюдать эти условия, то инструкции по загрузке данных можно найти в файле Jupyter-блокнота для этой главы.

Данные хранятся в двух файлах — «словаре» с описанием формата данных и файле данных:

```
dct_file = "2002FemPreg.dct"
dat_file = "2002FemPreg.dat.gz"
```

В Jupyter-блокноте для главы 1 определена функция, считывающая эти файлы. Она называется `read_stata`, поскольку формат этих данных совместим со статистическим программным пакетом Stata.

Вот пример ее использования:

```
preg = read_stata(dct_file, dat_file)
```

Функция возвращает объект типа `DataFrame`, который является структурой данных библиотеки Pandas и представляет табличные данные в виде строк и столбцов. В этой таблице каждая строка соответствует беременности, о которой сообщила респондентка, а столбец — одной из **переменных**. Каждая переменная может содержать ответы на вопросы анкеты или значения, рассчитанные на основе ответов на один или несколько вопросов.

Помимо данных, `DataFrame` содержит имена и типы переменных, а также предоставляет методы для доступа к данным и их модификации. Объект `DataFrame` имеет атрибут `shape`, содержащий количество строк и столбцов:

```
preg.shape
```

```
(13593, 243)
```

Этот датасет содержит 243 переменные с информацией о 13 593 беременностях. Объект `DataFrame` предоставляет метод `head`, который выводит на экран первые несколько строк:

```
preg.head()
```

	caseid	pregordr	howpreg_n	howpreg_p	moscurrp	nowprgdk	pregend1	pregend2	nbrnaliv	…
0	1	1	NaN	NaN	NaN	NaN	6.0	NaN	1.0	…
1	1	2	NaN	NaN	NaN	NaN	6.0	NaN	1.0	…
2	2	1	NaN	NaN	NaN	NaN	5.0	NaN	3.0	…
3	2	2	NaN	NaN	NaN	NaN	6.0	NaN	1.0	…
4	2	3	NaN	NaN	NaN	NaN	6.0	NaN	1.0	…

В левом столбце содержится индекс объекта `DataFrame`, который состоит из меток для каждой строки. В данном случае в качестве меток выступают целые числа, начинающиеся с 0, но они также могут быть строками или другими типами.

Также `DataFrame` имеет атрибут `columns`, содержащий имена переменных (столбцов):

```
preg.columns
```

```
Index(['caseid', 'pregordr', 'howpreg_n', 'howpreg_p', 'moscurrp', 'nowprgdk',
       'pregend1', 'pregend2', 'nbrnaliv', 'multbrth',
       …
       'poverty_i', 'laborfor_i', 'religion_i', 'metro_i', 'basewgt',
       'adj_mod_basewgt', 'finalwgt', 'secu_p', 'sest', 'cmintvw'],
      dtype='object', length=243)
```

Имена столбцов содержатся в объекте типа `Index` — еще одной структуре данных Pandas. Чтобы получить доступ к столбцу объекта `DataFrame`, можно воспользоваться именем столбца в качестве ключа:

```
pregordr = preg["pregordr"]
type(pregordr)
```

```
pandas.core.series.Series
```

Результатом является объект типа `Series` библиотеки Pandas, представляющий собой последовательность значений. У `Series` тоже имеется метод `head`, отображающий первые несколько значений и их метки:

```
pregordr.head()
```

```
0    1
1    2
2    1
3    2
4    3
Name: pregordr, dtype: int64
```

Последняя строка содержит название объекта `Series` и `dtype`, являющийся типом значений. В данном примере `int64` указывает, что значения являются 64-разрядными целыми числами.

Датасет NSFG содержит в общей сложности 243 переменные. Вот некоторые из тех, что мы будем использовать для разведочного анализа в этой книге:

`caseid`
: Целочисленный идентификатор (ID) респондентки.

`pregordr`
: Порядковый номер беременности: для первой беременности респондентки код равняется 1, для второй беременности — 2 и т. д.

`prglngth`
: Продолжительность беременности в неделях (целое число).

`outcome`
: Целочисленный код, обозначающий исход беременности. Код 1 указывает на рождение живого ребенка.

`birthord`
: Порядковый номер рождения живого ребенка: для первого ребенка респондентки код равен 1 и т. д. Для других исходов это поле остается пустым.

`birthwgt_lb` *и* `birthwgt_oz`
: Содержат вес ребенка при рождении в фунтах и унциях.

`agepreg`
: Возраст матери на момент окончания беременности.

`finalwgt`
: Статистический вес, связанный с данной респонденткой. Это число с плавающей точкой, указывающее на количество женщин в популяции США, которых респондентка представляет.

Если вы внимательно ознакомитесь с кодбуком, то обнаружите, что многие переменные являются **перекодированными**. Это означает, что они не принадлежат к **исходным данным**, собранным в ходе опроса, а являются результатом расчета с использованием исходных данных.

Например, значение `prglngth` в случае живорождения равно исходной переменной `wksgest` (недели беременности) при ее наличии; в противном случае оно оценивается по формуле `mosgest * 4.33` (месяцы беременности, умноженные на среднее количество недель в месяце).

Перекодирование переменных часто подразумевает проверку согласованности и точности данных. Поэтому обычно рекомендуется пользоваться перекодированными переменными, когда они доступны, — если только нет веских причин для самостоятельной обработки исходных данных.

Проверка данных

При экспорте данных из одной программной среды и последующем импорте в другую могут возникать ошибки. Когда вы только знакомитесь с новым датасетом, то можете неправильно декодировать данные или неверно интерпретировать их значение. Если вы уделите время проверке данных, то сможете впоследствии его сэкономить и предотвратить появление ошибок.

Один из способов проверки корректности данных — вычисление основных статистических показателей и их сравнение с опубликованными результатами. Например, кодбук NSFG содержит сводные таблицы для всех переменных. Вот таблица для переменной `outcome`, содержащей коды исходов всех беременностей:

Value (Значение)	Label (Метка)	Total (Всего)
1	LIVE BIRTH (ЖИВОРОЖДЕНИЕ)	9148
2	INDUCED ABORTION (ИСКУССТВЕННЫЙ АБОРТ)	1862
3	STILLBIRTH (МЕРТВОРОЖДЕНИЕ)	120

Value (Значение)	Label (Метка)	Total (Всего)
4	MISCARRIAGE (ВЫКИДЫШ)	1921
5	ECTOPIC PREGNANCY (ВНЕМАТОЧНАЯ БЕРЕМЕННОСТЬ)	190
6	CURRENT PREGNANCY (ТЕКУЩАЯ БЕРЕМЕННОСТЬ)	352
Total (Итого)		13593

В столбце «Всего» указано количество беременностей для каждого исхода. Чтобы проверить эти значения, воспользуемся методом `value_counts`, который подсчитывает количество появлений каждого значения, и методом `sort_index`, сортирующим результаты в соответствии со значениями колонки `Index` (слева):

```
preg["outcome"].value_counts().sort_index()
```

```
outcome
1    9148
2    1862
3     120
4    1921
5     190
6     352
Name: count, dtype: int64
```

Сравнивая наши результаты с опубликованной таблицей, мы убеждаемся, что значения в столбце `outcome` верны. Вот аналогичная официальная таблица для `birthwgt_lb`:

Value (Значение)	Label (Метка)	Total (Всего)
.	inapplicable (неприменимо)	4449
0–5	UNDER 6 POUNDS (МЕНЕЕ 6 ФУНТОВ)	1125
6	6 POUNDS (6 ФУНТОВ)	2223
7	7 POUNDS (7 ФУНТОВ)	3049
8	8 POUNDS (8 ФУНТОВ)	1889
9–95	9 POUNDS OR MORE (9 ФУНТОВ И БОЛЕЕ)	799
97	Not ascertained (Не установлено)	1
98	REFUSED (ОТКАЗ)	1
99	DON'T KNOW (НЕИЗВЕСТНО)	57
Total (Итого)		13593

Вес при рождении указан только для беременностей, окончившихся рождением живого ребенка. В таблице указано, что в 4449 случаях эта переменная неприменима. Кроме того, есть один случай, когда вопрос не был задан, один случай, когда респондентка не ответила, и 57 случаев, когда вес был неизвестен опрашиваемым.

Здесь, чтобы сравнить суммарные значения из датасета с опубликованными в кодбуке, тоже можно использовать метод `value_counts`:

```
counts = preg["birthwgt_lb"].value_counts(dropna=False).sort_index()
counts
```

```
birthwgt_lb
0.0        8
1.0       40
2.0       53
3.0       98
4.0      229
5.0      697
6.0     2223
7.0     3049
8.0     1889
9.0      623
10.0     132
11.0      26
12.0      10
13.0       3
14.0       3
15.0       1
51.0       1
97.0       1
98.0       1
99.0      57
NaN     4449
Name:  count, dtype: int64
```

Аргумент `dropna=False` указывает `value_counts` не игнорировать значения «неприменимо» («Not applicable» или «NA»). Они помечены в результатах как `NaN`, что означает «не число» (Not a Number), а количество таких случаев соответствует числу значений «неприменимо» в кодбуке.

Число записей для 6, 7 и 8 фунтов соответствует кодбуку. Чтобы проверить подсчет для диапазона веса от 0 до 5 фунтов, можно воспользоваться атрибутом `loc` (сокращение от «location» — местоположение) и индексом среза для выбора подмножества значений:

```
counts.loc[0:5]
```

```
birthwgt_lb
0.0      8
1.0     40
2.0     53
3.0     98
```

```
4.0    229
5.0    697
Name: count, dtype: int64
```

И можно задействовать метод `sum`, чтобы их суммировать:

```
counts.loc[0:5].sum()
```

```
1125
```

Итоговая сумма соответствует данным кодбука.

Значения 97, 98 и 99 соответствуют случаям, когда вес при рождении неизвестен. Существуют различные способы работы с отсутствующими данными. Простой вариант — заменить эти значения на `NaN`. Так же мы поступим и с очевидно неверным значением в 51 фунт[1].

Мы можем использовать метод `replace` («заменить») следующим образом:

```
preg["birthwgt_lb"] = preg["birthwgt_lb"].replace([51, 97, 98, 99], np.nan)
```

Первый аргумент метода — это список значений, подлежащих замене. Второй аргумент, `np.nan`, представляет собой значение `NaN` из библиотеки NumPy.

Когда вы таким образом считываете данные, то часто их приходится проверять на наличие ошибок и работать со специальными значениями. Подобные операции называются **очисткой данных**.

Преобразование данных

В рамках очистки данных иногда требуется изменить формат данных или провести с ними другие расчеты.

Например, переменная `agepreg` содержит возраст матери на момент окончания беременности. Он, согласно кодбуку, представляет собой целое число сотых долей года (centiyears), в чем можно убедиться, если применить метод `mean` для вычисления среднего значения:

```
preg["agepreg"].mean()
```

```
2468.8151197039497
```

Чтобы получить значение в годах, можно поделить весь столбец на 100:

```
preg["agepreg"] /= 100.0
preg["agepreg"].mean()
```

```
24.6881511970395
```

[1] Примерно соответствует 23 кг. — *Примеч. пер.*

Теперь среднее значение выглядит более правдоподобным.

Еще один пример: вес при рождении представлен в виде двух столбцов, `birthwgt_lb` и `birthwgt_oz`, содержащих фунты и унции соответственно. Будет удобнее объединить их в один столбец, содержащий вес в фунтах и долях фунта.

Сначала проведем очистку `birthwgt_oz`, как мы это делали в случае с `birthwgt_lb`:

```
preg["birthwgt_oz"] = preg["birthwgt_oz"].replace([97, 98, 99], np.nan)
```

Теперь, используя очищенные значения, создадим новый столбец, объединяющий фунты и унции в единую величину:

```
preg["totalwgt_lb"] = preg["birthwgt_lb"] + preg["birthwgt_oz"] / 16.0
preg["totalwgt_lb"].mean()
```

```
7.265628457623368
```

Среднее выглядит правдоподобным[1].

Сводные статистические показатели

Статистический показатель — это число, полученное из датасета, обычно предназначенное для количественной оценки некоторого параметра данных. В качестве примера можно назвать количество значений (count), среднее (mean), дисперсию (variance) и стандартное отклонение (standard deviation).

Объект `Series` имеет метод `count`, который возвращает количество значений, отличных от `nan`:

```
weights = preg["totalwgt_lb"]
n = weights.count()
n
```

```
9038
```

Он также предоставляет метод `sum`, возвращающий сумму всех значений; его можно использовать для вычисления среднего следующим образом:

```
mean = weights.sum() / n
mean
```

```
7.265628457623368
```

Но, как вы уже знаете, существует также метод `mean`, дающий тот же результат:

```
weights.mean()
```

```
7.265628457623368
```

[1] Примерно 3,3 кг. — *Примеч. пер.*

В нашем датасете средний вес при рождении составляет около 7.3 фунта.

Дисперсия — это статистический показатель, оценивающий разброс множества значений. Это среднее квадратов отклонений (squared deviations), равных расстоянию каждой точки от среднего значения:

```
squared_deviations = (weights - mean) ** 2
```

Среднее квадратов отклонений можно получить следующим образом:

```
var = squared_deviations.sum() / n
var
```

```
1.983070989750022
```

Объект `Series` ожидаемо предоставляет метод `var`, который делает *почти* то же самое:

```
weights.var()
```

```
1.9832904288326545
```

Результат слегка отличается от нашего, потому что в методе `var` при вычислении среднего квадратов отклонений происходит деление на `n-1`, а не на `n`. Это связано с тем, что существует два способа вычисления дисперсии выборки, в зависимости от того, что вы хотите сделать. Я объясню эту разницу в разделе «Оценка дисперсии» главы 8 на с. ??, но в реальной жизни это обычно не так важно. Если вы предпочитаете вариант с `n` в знаменателе, то можете получить его, передав методу `var` в качестве именованного аргумента `ddof=0`:

```
weights.var(ddof=0)
```

```
1.983070989750022
```

У наших данных дисперсия веса при рождении составляет около 1.98, но это значение трудно интерпретировать — прежде всего потому, что оно выражено в квадратных фунтах. Дисперсия полезна для некоторых вычислений, но не является хорошим способом описания данных. Более удачный показатель — **стандартное отклонение** (standard deviation), которое представляет собой квадратный корень из дисперсии. Его можно вычислить следующим образом:

```
std = np.sqrt(var)
std
```

```
1.40821553384062
```

Или же воспользоваться методом `std`:

```
weights.std(ddof=0)
```

```
1.40821553384062
```

В нашем датасете стандартное отклонение веса при рождении составляет около 1.4 фунта. Говоря просто, значения, лежащие в пределах одного или двух стандартных отклонений от среднего, являются распространенными, а более удаленные — редкими.

Интерпретация данных

Чтобы эффективно работать с данными, нужно обращать внимание не только на статистические показатели, но и на контекст. В качестве примера выберем в таблице с данными о беременности те строки, где переменная `caseid` равна 10229. Метод `query` принимает на вход строку, которая, помимо прочего, может содержать имена столбцов, операторы сравнения и числа:

```
subset = preg.query("caseid == 10229")
subset.shape
```

```
(7, 244)
```

Результатом является объект `DataFrame`, содержащий только те строки, для которых запрос принимает значение `True` (истина). Эта респондентка сообщила о семи беременностях. Вот их исходы, записанные в хронологическом порядке:

```
subset["outcome"].values
```

```
array([4, 4, 4, 4, 4, 4, 1])
```

Код исхода `1` обозначает рождение живого ребенка. Код `4` указывает на выкидыш, то есть потеря ребенка, как правило, по неизвестной физиологической причине.

Статистически случай этой респондентки не является чем-то необычным. Потеря ребенка — нередкое явление, к тому же есть другие респондентки, которые сообщили о таком же количестве случаев. Но эти данные рассказывают историю женщины, которая была беременна шесть раз, и каждый из этих случаев заканчивался выкидышем. Ее седьмая и последняя беременность завершилась рождением живого ребенка. Если мы отнесемся к этой информации с сочувствием, то, конечно, будем тронуты историей, которую она рассказывает.

За каждой строкой в датасете NSFG стоит человек, честно ответивший на множество сложных вопросов личного характера. Мы можем использовать эти данные для ответа на статистические вопросы о семейной жизни, репродукции и здоровье. При этом мы обязаны считаться с людьми, которых эти данные описывают, и проявлять к ним уважение и благодарность.

Глоссарий

В конце каждой главы книги приводится глоссарий терминов, которые в ней были определены.

Бытующее мнение (anecdotal evidence)

Данные, собранные неофициально на основе небольшого числа отдельных случаев и часто без применения систематической выборки.

Поперечное исследование (cross-sectional study)

Исследование, в ходе которого собираются данные о репрезентативной выборке из популяции за единичный момент или интервал времени.

Цикл (cycle)

Один интервал сбора данных в исследовании, где данные собираются за несколько интервалов времени.

Генеральная совокупность, или популяция (population)

Вся группа лиц или предметов, которые являются объектом исследования.

Выборка (sample)

Подмножество генеральной совокупности, часто отбираемое случайным образом.

Респонденты (respondents)

Люди, которые участвуют в опросе и отвечают на вопросы.

Репрезентативный (representative)

Выборка является репрезентативной, если она сходна с генеральной совокупностью по тем параметрам, которые важны для целей исследования.

Стратифицированный (stratified)

Выборка является стратифицированной, если в нее намеренно избыточно включены представители некоторых групп. Это делается, как правило, для того, чтобы набрать количество членов, достаточное для получения достоверных выводов.

Избыточная выборка (oversampled)

Группа является избыточной выборкой, если ее члены имеют повышенную вероятность попадания в выборку.

Переменная (variable)

В данных опроса переменная представляет собой совокупность ответов на вопросы или вычисленных на их основе значений.

Кодбук (codebook)

Документ, описывающий переменные датасета и содержащий иную информацию о данных.

Перекодированная переменная (recode)

Переменная, вычисленная на основе других переменных в датасете.

Исходные, или сырые, данные (raw data)

Данные, не прошедшие обработку после сбора.

Очистка данных (data cleaning)

Процесс выявления и исправления ошибок в датасете, работы с пропущенными значениями и вычисления перекодированных переменных.

Статистический показатель (statistic)

Значение, которое описывает или обобщает некоторое свойство выборки.

Стандартное отклонение (standard deviation)

Статистический показатель, количественно оценивающий разброс данных вокруг среднего значения.

Упражнения

Упражнения этой главы требуют использования файла NSFG о беременностях.

Упражнение 1.1

В таблице `preg` выберите столбец `birthord`, выведите на экран количество значений и сравните с результатами, опубликованными в кодбуке для 6-го цикла опроса NSFG по беременностям (https://oreil.ly/M2hDe).

Упражнение 1.2

Создайте новый столбец с именем `totalwgt_kg`, который содержит вес при рождении в килограммах (килограмм — это примерно 2.2 фунта). Для нового столбца вычислите среднее значение и стандартное отклонение.

Упражнение 1.3

Каковы продолжительности беременностей у респондентки с `caseid` 2298?

Каков был вес при рождении первого ребенка респондентки с `caseid` 5013? Подсказка: в запросе можно использовать оператор `and` для проверки нескольких условий.