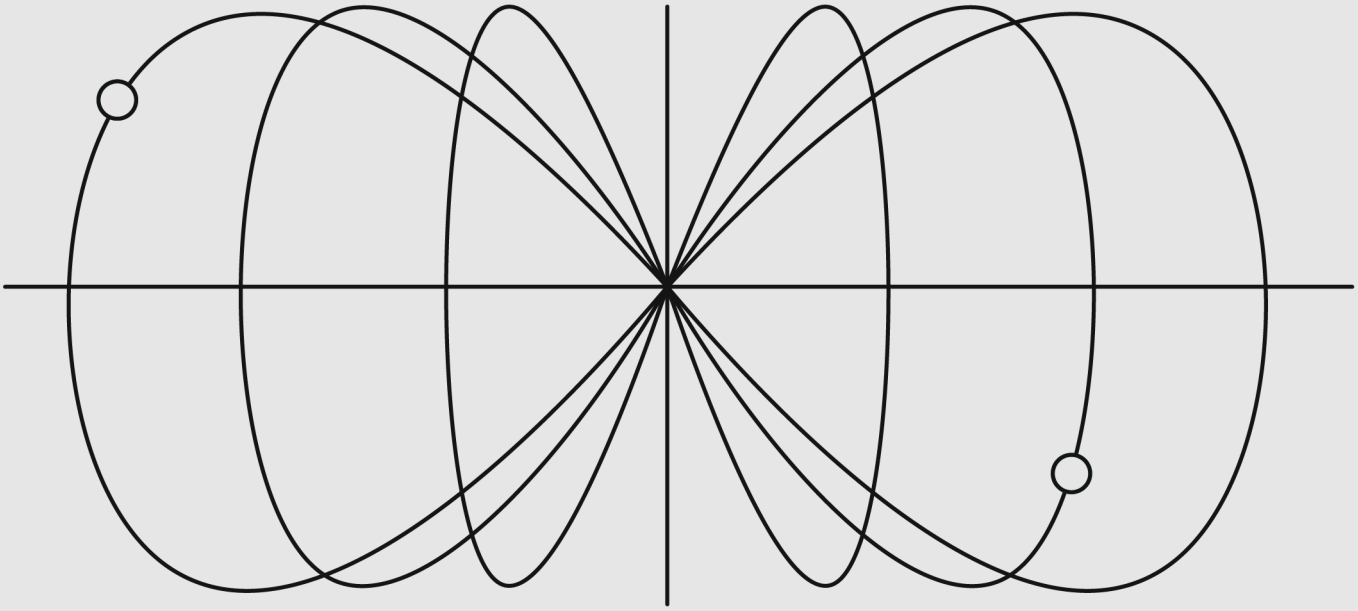
ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ



Статистика

ЛОНГРИД «ОПИСАТЕЛЬНЫЕ СТАТИСТИКИ»

Москва
2024

Цели и задачи предмета «Введение в статистику»

Статистика — это наука о данных. А что можно делать с данными? На этот вопрос отвечают 4 части курса.

Часть курса	Пример задачи	Что изучаем
1. **Описание и визуализация данных**	Подготовить слайды с таблицами и графиками. На слайдах показать, какая рекламная кампания принесла больше всего прибыли.	• Описательные статистики. • Графики. • Работа с данными в Python.
2. **Данные vs модель данных**	Узнать, как выбирать людей для пробных запусков рекламных кампаний, чтобы выводы можно было переносить на других людей.	• Выборка. • Генеральная совокупность. • Связь между ними.
3.**Вероятностные модели**	Спрогнозировать, какая доля сильных кандидатов будет проваливать отборочный тест на собеседовании на работу, как часто слабые кандидаты будут справляться с тестом успешно.	• Распределения (дискретное, Бернулли и биномиальное, нормальное). • Центральная предельная теорема.
4. **Калибровка и тестирование модели**	Выяснить, скольким людям нужно показать рекламу, чтобы выводы о её эффективности можно было распространять на других людей.	• Оценка среднего и пропорции. • Гипотеза. • Отвергающее правило (критерий). • Значимость. • Ошибки тестов, мощность.

Итого впереди 4 раздела и 15 тем, чтобы познакомиться со статистикой.

Тема 1. Обзор данных. Описательные статистики

Цели и задачи

Цель — научиться описывать данные небольшим количеством чисел.

Задачи:

- познакомиться с понятием «датасет» (от англ. data set, «набор данных»);
- потренироваться интерпретировать значения в датасете;
- научиться просматривать датасеты в Python;
- научиться по-разному оперировать данными в зависимости от типа данных;
- научиться обнаруживать некорректные данные и очищать от них датасет;
- познакомиться с описательными статистиками;
- научиться описывать датасет с их помощью;
- научиться интерпретировать описательные статистики.

Данные

При работе с данными их по возможности сводят к числовым, а затем оперируют числами или показывают графики, опирающиеся на числа.

[!] Инсайт Числа помогают людям прийти к общему мнению.

Если спросить у окружающих, что больше: «огромнейший ассортимент» или «гиганский ассортимент», возникнут споры: кто-то скажет, что первое, кто-то, что второе, а кто-то откажется сравнивать. Но если попросить сравнить 723 и 514, все согласятся, что 723 больше.

Встречаются и нечисловые данные. Например, пол человека принимает 2 нечисловых значения: «мужской» или «женский».

Определение 1

Наблюдение (observation) — это строка в таблице ***(датасете)***.

Определение 2

Датасет (dataset) — набор ***наблюдений***. Таблица, в которой:

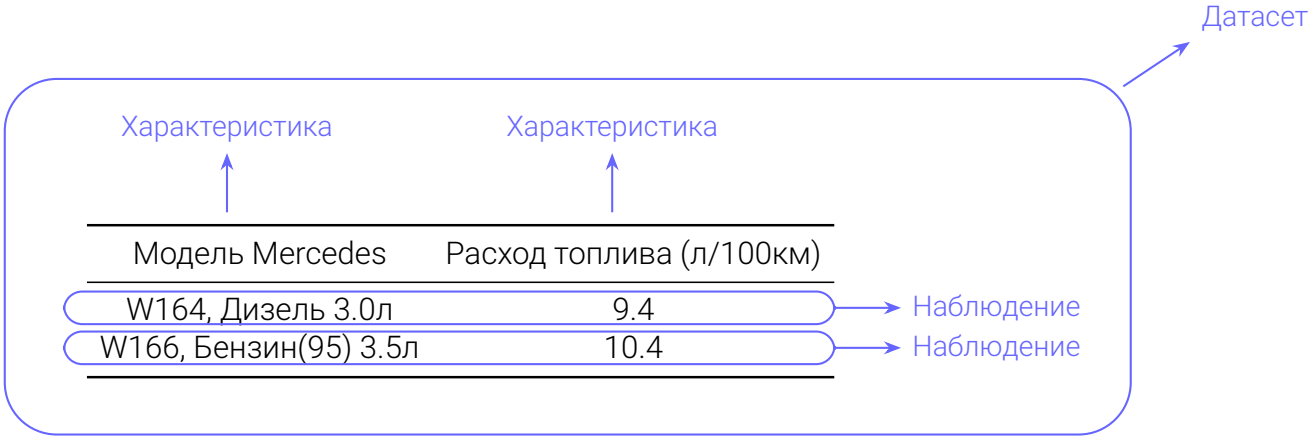
- строка соответствует одному наблюдению;
- столбец соответствует одной характеристике наблюдений.

Пример 1

Данные ниже взяты с сайта drom.ru

Модель Mercedes	Расход топлива (л/100км)
W164, Дизель 3.0л	9.4
W166, Бензин (95) 3.5л	10.4

- Датасет — вся таблица.
- Наблюдение — каждая строчка. Отдельное наблюдение соответствует конкретной модели автомобиля Mercedes.
- Столбцы — характеристики автомобилей:
  - первый столбец — это характеристика автомобиля «модель»;
  - второй столбец — это характеристика автомобиля «расход топлива».



1 Интерпретация

Если говорить с собеседником на его языке, удастся быстрее добиться понимания. Информационная апперцепция интенсифицируется при унификации эмитентом пространства коммуникации к аттрибутированным реципиенту возможностям.

Кстати, второе предложение имеет тот же смысл, что и первое.

Определение 3

Интерпретация (interpretation) — текст, переведённый на понятный язык без потери смысла, или с небольшой неточностью смысла. Требование «понятный язык» значит, что интерпретацию сможет понять неспециалист, не владеющий данными о текущей ситуации.

Определение 4

Интерпретация (interpretation) значения — текст, который:

- поясняет смысл данного значения в текущем контексте;
- сообщает единицы измерения, если они неочевидны (только для числовых значений);
- по возможности не использует специальные термины и обозначения из статистики или математики.

Пример 2

Дана таблица с данными (источник: drom.ru)

Модель Mercedes	Расход топлива (л/100км)
W164, Дизель 3.0л	9.4
W166, Бензин (95) 3.5л	10.4

Задача — проинтерпретировать выделенные числа.

Решение.

Модель W164 автомобиля Mercedes с дизельным двигателем объёмом 3 литра на 100 км пути расходует в среднем 9.4 литра дизельного топлива. Модель W166 автомобиля Mercedes с бензиновым двигателем объёмом 3.5 литра на 100 км пути расходует в среднем 10.4 литра бензина марки АИ-95. ◀

[!] Инсайт Цель интерпретации — объяснить смысл числового показателя человеку, который не знаком с этой сферой. Для тех, кто «в теме», и так всё понятно. Интерпретации нужны, чтобы поняли остальные.

«Расход 10.4 (95)» — плохая интерпретация. Она понятна автовладельцу, но может быть неясной человеку, который не стоял на заправке и не знает, что 95 — марка бензина, 10.4 — литры, а расход принято считать на 100 км пути.

2 Работа с данными в Python

Использовать датасет об автомобилях легко, ведь он маленький. Но чтобы обработать большие объёмы данных, нужны инструменты программирования.

Мы будем использовать язык программирования Python.

За работу с датасетами в Python отвечает модуль pandas, который часто сокращается до pd. Этот модуль умеет загружать датасеты с диска или прямо из интернета, сохранять датасеты на диск, отображать, модифицировать, выбирать подходящие под заданные условия данные, проводить расчёты над данными и показывать результаты.

Определение 5

Модуль ***pandas*** — модуль для работы с данными в языке Python. Подключается командой `import pandas as pd`.

В pandas чаще всего используются 2 типа объектов: DataFrame («интерфейс с данными» — англ.) и Series («ряд» — англ.).

Определение 6

DataFrame — датасет в виде двумерной таблицы. Часто обозначается как *df*.

Команда `df = pd.read_csv("filename.csv")` или `df = pd.read_excel("filename.xls")` создаёт датафрейм и загружает в него данные из файла.

Определение 7

Series — одномерный набор значений. Столбец и реже строка.

Пример 3

Датафрейм умеет выполнять операции с данными, ниже примеры команд:

```
# выбрать первые 2 наблюдения и положить в новый DataFrame
df2 = df.head(2)

# выбрать только столбцы с названиями Column1 и column 2 в новый DataFrame
df3 = df[['Column1', 'column 2']]
# название должно совпадать точно:
# важно, большая или маленькая буква, есть или нет пробела

# получился Series (столбец)
series1 = df3['Column1']

# получился DataFrame (с одним столбцом и множеством строк)
df4 = df3[['Column1']]
```

df =

Column1	column 2	Col3	...
1	2	Apple	...
3	4	Banana	...
5	6	Cactus	...
...	...	...	...
...	...	...	...

→ df2 =

Column1	column 2	Col3	...
1	2	Apple	...
3	4	Banana	...

df3 =

Column1	column 2
1	2
3	4
5	6
...	...
...	...

→ series1 = 1, 3, 5, ..., ...

→ df4 =

Column1
1
3
5
...
...

[!] Инсайт Визуально DataFrame отрисовывается лучше, чем Series: получается в конце DataFrame, даже если в нём только один столбец. То есть, нужно использовать `df3[['Column1']]` вместо `df3['Column1']`, когда планируется отобразить результат.

3 Работа в Google Colab

В курсе мы предлагаем работать в среде Google Colab. Это аналог файла, в котором код на языке Python разбит на «ячейки». Здесь можно запускать не весь код сразу, а по одной ячейке. На экране появляется результат выполнения последней команды в запущенной ячейке. Только последней команды! Результаты остальных команд будут вычислены, но невидимы.

Пример 4

Запуск ячейки в в Google Colab:

```
df=pd.read_csv('filename.csv')
df.head(5) # результат этой команды мы не увидим
df.head(2) # результат этой команды мы не увидим
df         # результат этой команды мы... увидим!!!
```

Увидим только результат последней команды: то есть весь датафрейм.

3.1 Ошибки при работе в Google Colab

Если при выполнении ячейки возникает ошибка , ячейка краснеет, а под ней появляется причина ошибки.

[!] Инсайт При возникновении ошибки нужно задуматься, какого типа объект/переменная используется.

Пример 5

Самая частая ошибка.

```
df[df['Age'] < 20]
```

Если при выполнении этой команды возникла ошибка, проблема в том, что датасет некорректно хранит возраст. Например, вместо числа 18, хранится строка ’18’, а это другой тип данных с точки зрения Python. Поэтому операция сравнения с числом невозможна. Исправить некорректное хранение чисел в виде строк можно следующей командой:

```
df['Age'] = df['Age'].apply(int)
```

После этого предыдущая команда выполнится без ошибок.

4 Виды датасетов

В процессе работы иногда хочется удалить отдельные строчки или отсортировать данные. Безопасность этого действия зависит от типа датасета.

Есть три основных типа:

- *кросс-секция (cross-section)*,
- *временной ряд (time series)*,
- *панельные данные (panel data)*.

Определение 8

Кросс-секция (cross-section) — данные, в которых разные наблюдения соответствуют разным объектам в один и тот же либо в неизвестный момент времени. На них, как правило, нет естественного порядка, кросс-секции можно сортировать, не задумываясь о нарушении порядка наблюдений.

Пример 6

Кросс-секция: строки — разные дома.

Адрес	Количество квартир
Кузнецкий проезд, 5	40
улица Ленивка, 1	160

Определение 9

Временной ряд (time series) — данные, в которых разные наблюдения соответствуют одному и тому же объекту в разные периоды (годы, дни, секунды). Существует естественный порядок наблюдений, хронологический: следующее наступило позже предыдущего. После сортировки это свойство, скорее всего, нарушится. Рядом окажутся наблюдения непоследовательных лет, может наступить хронологический хаос.

Пример 7

Временной ряд: строки — разные часовые интервалы, идущие по порядку.

Время	Число посетителей
9:00-9:59	5
10:00-10:59	3

Определение 10

Панельные данные (panel data) — данные, в которых разные наблюдения соответствуют разным объектам в нескольких периодах. Естественный порядок наблюдений — хронологический для каждого объекта. После сортировки он может нарушиться.

Пример 8

Панельные данные: строка — число заказов на определённый адрес в определённый месяц. Для каждого адреса — несколько месяцев. Для каждого месяца — несколько адресов.

Адрес	Период	Число заказов
улица Ленивка, 1	Январь 2020	4
улица Ленивка, 1	Февраль 2020	12
Кузнецкий проезд, 5	Январь 2020	150
Кузнецкий проезд, 5	Февраль 2020	210

5 Типы данных

В Python есть своё опредение типа данных. Но мы сейчас поговорим о типах с точки зрения статистики, а не того, как данные хранятся в байтах памяти компьютера.

Столбцы в датасетах могут иметь разный тип данных. Но во всех наблюдениях значения в одном и том же столбце имеют один и тот же тип данных. В зависимости от типа данных разные операции с такими данными будут осмысленными либо бессмысленными.

Определение 11

Тип данных ***числовой (numerical)***. Можно сравнивать на больше/меньше, а значит, и сортировать; кроме редких исключений можно выполнять арифметические операции

Определение 12

Тип данных ***упорядоченные категории (categorical ordinal)***. Можно сравнивать на больше/меньше, а значит, и сортировать; бессмысленно выполнять арифметические операции

Определение 13

Тип данных ***неупорядоченные категории (categorical nominal)***. Нельзя сравнивать на больше/меньше, а значит, почти всегда бессмысленно сортировать, бессмысленно

выполнять арифметические операции. К этому типу относят обычно и ***текстовые*** данные.

Операции со значениями	Числовой	Категориальный	
		Упорядоченный	Неупорядоченный
Сравнивать на больше/меньше	можно	можно	бессмысленно
Сортировать	можно	можно	бессмысленно
Выполнять арифметические действия	можно почти всегда	бессмысленно	бессмысленно

Пример 9

Дан датасет.

Occupation	AgeGroup	Salary (th.rub/month)	IELTS	Female
Elite bodyguard	<25	120	6.0	0
Python developer	25-30	100	3.0	0
Product manager	25-30	100	9.0	1

У характеристик Occupation, AgeGroup, Salary тип данных определяется «без подводных камней».

При определении типа значений в столбцах IELTS и Female вполне можно ошибиться. В дальнейшем это приведёт к некорректным операциям с данными в этих столбцах.

- Occupation (род занятий) — очевидно, неупорядоченная (иногда говорят «неупорядочиваемая») категорийная характеристика: нельзя сказать, что профессия телохранителя больше или меньше, чем профессия разрабочика или менеджера. Они разные. Точка.
- AgeGroup (возрастная группа) — очевидно, упорядоченная категорийная характеристика. Можно сказать, что второй и третий человек в одинаковой возрастной категории, но можно ещё сказать, что они оба старше, чем первый человек — сравнение на больше/меньше разрешено. Однако нельзя сказать, во сколько раз старше или на сколько лет старше: арифметические операции здесь некорректны.
- Salary (зарплата) — очевидно, имеет числовой тип. Можно сказать, что зарплата второго и третьего одинаковы, можно сравнить, что у первого человека зарплата выше, и можно даже сказать, что на 20% или что на 20 тысяч рублей. Всё это корректно делать с числовым типом данных.
- IELTS (балл на международно признаваемом экзамене на владение английским языком) — выглядит числовой характеристикой, но на самом деле она категорийная упорядоченная. Согласно результатам экзамена, первый человек знает язык лучше второго, но нельзя говорить, что в два раза лучше. Точно так же нельзя сказать, что экзамен показал втрое лучшие знания языка у третьего человека

по сравнению со вторым. Нельзя сказать, что сумма знаний первого и второго равна знаниям третьего. Некорректно проводить арифметические действия с категорийной переменной, даже если она выглядит, как числовая.

- Female (индикатор, показывающий пол женский или нет) — выглядит, будто тип характеристики числовой, но на самом деле он категорийный неупорядоченный. Некорректно сказать, что пол третьего человека больше, чем пол второго, или складывать сумму полов людей. У третьего человека женский пол, у первый двух мужской. Они разные. Точка. Больше ничего с категорийным неупорядоченным типом проделывать нельзя.

Характеристика Female из примера подпадает ещё под одну классификацию, потому что принимает только значения 0 или 1.

Определение 14

Индикатор (indicator) или ***дамми (dummy)*** характеристика — такая характеристика наблюдения, которая может принимать только значения 0 или 1.

Понятие индикатора понадобится дальше, когда будем разбираться с частотными описательными статистиками. Пока просто запомним, что «0 или 1» — особый случай.

6 Очистка данных

Если данные содержат ошибки, то и выводы из этих данных могут оказаться неверными.

[!] Инсайт Не нужно доверять выводам, основанным на ошибочных данных.

Также проблемой может стать примесь не интересующих нас наблюдений в датасете с важными для нас данными.

Определение 15

Очисткой данных (data cleaning) мы будем называть процедуру из двух действий:

1. Исправление ошибок в данных.
2. Удаление нерелевантных или некорректных данных.

Иногда починить наблюдение не получается и приходится его выбросить. К сожалению, ***выбрасывание (drop)*** данных также может ***подтолкнуть к неверным (misleading) выводам.***

Пример 10

В столбце «Рост спортсмена» аналитик оставил лишь наблюдения от 50 до 250, выбросив остальные, как испорченные данные. Оказалось, что в федерации баскетбола местные

решили вносить данные в метрах, но забыли об этом предупредить, когда передавали данные. Таким образом, рост баскетболистов, например, «2.02», попал под условие удаления. Аналитик удалил всех баскетболистов вместо того, чтобы умножить эти значения на 100 и не удалять.

[!] Инсайт Важно всегда обдумывать решения по очистке данных и смотреть на подозрительные данные собственными глазами. НЕ стоит пользоваться никакими «автоматическими» правилами для удаления данных, даже если они звучат очень умно и научно и опубликованы в интернете:

- «всегда удалять 1% самых маленьких значений»;
- «всегда удалять значения, которые меньше трёх сигм»;
- и подобными.

Баскетболисты не хотят, чтобы их удаляли, даже не взглянув...

Один из первых признаков наличия ошибочных данных — присутствие выбросов.

Определение 16

Выбросами (outliers) называются числовые значения, слишком большие или слишком маленькие — слишком непохожие на остальные данные, чтобы разумно было анализировать их вместе с остальными. Сами по себе выбросы необязательно ошибочны, но мы зачастую не хотим анализировать их вместе с остальными наблюдениями.

Чтобы найти выбросы, нужно:

1. Отсортировать датасет по возрастанию.
2. Посмотреть в начало и конец отсортированного датасета собственными глазами. Поискать выбросы с помощью здравого смысла.
   Если выбросы есть, они либо в начале, где самые маленькие значения, либо в конце, где самые большие — больше им быть негде.

Когда выбросы найдены и обследованы человеческим глазом, нужно:

1. Найти искажения.
2. Придумать правило для исправления искажений.
3. Исправить искажения.
4. Найти нерелевантные и некорректные значения.
5. Придумать правило для фильтрации: какие наблюдения выбрасывать, а какие оставлять.
6. Выбросить нерелевантные и некорректные наблюдения.

Пример 11

После сортировки взрослых спортсменов по росту отрезки данных с каждого края выглядят так:

head	-8945234	0	1.7	1.8	2.01	54	153	154	154	. . .	. . .

tail	. . .	. . .	216	217	312	3459870

Проводим очистку данных.

1. Выбросы: -8945234, 0, 1.7, 1.8, 2.01, 54, 312, 3459870. Эти значения не особенно похожи на рост взрослого человека в сантиметрах.
2. Значения 54 и 312 — предположительно, искажения из-за опечаток.
3. Заменяем 54→154 и 312→212.
4. Значения 1.7, 1.8, 2.01 — предположительно, искажения, и показывают рост в метрах, а не в сантиметрах, как в остальных наблюдениях.
5. Фильтруем наблюдения между 0 и 2.5, умножаем их на 100 и заменяем изначальные значения на исправленные (умноженные на 100).
6. Фильтруем то, что получилось, оставляя только числа от 50 до 250.

Рост	170	180	201	154	153	154	154	. . .	. . .	216	217	212

Можно ещё раз отсортировать для красоты, но очистка данных проведена.

[!] Инсайт Иногда при очистке данных приходится принимать спорные решения. В примере выше мы решили, что 54 — это опечатка от 154, 312 — опечатка от 212, а 3459870 и 0 — мусор. Возможно, другой аналитик предпочёл бы отбросить и 54 и 312 как мусор. А третий аналитик не разглядел бы в 1.7, 1.8, 2.01 метры.

На этапе очистки данных многое зависит от здравого смысла аналитика. Компьютер в таких случаях должен быть только инструментом. Доверять ему принимать решения по автоматической очистке данных не стоит.

В следующем примере проводится очистка неупорядоченных категориальных данных. Здесь сортировка, выбросы и прочее неактуальны. Нужно только смотреть.

Пример 12

Видим в характеристике «Пол человека» значения.

Gender	Male	MALE	male	female	Female	. . .	. . .	FEMALE

Такое часто бывает, когда данные хранятся в компьютере в типе данных «строка». Компьютер будет видеть много различных значений: для него люди Male и male имеют разный пол.

На деле различных значений должно быть только 2.

Очистку данных произведём за 1 шаг:

1. Заменим каждую строку на такую же, но из маленьких букв (lowercase).

Gender	male	male	male	female	female	...	...	female

Теперь и компьютер тоже видит данные как категориальный тип с 2 значениями.

Описательные статистики

Трудно без слёз долго смотреть на огромные таблицы с данными. Хочется простого и понятного описания. В идеале, одним показателем.

Определение 17

Описательная статистика или просто ***статистика (statistic)*** — это способ описать данные одним показателем. Математически ***статистика*** — это функция от набора наблюдений.

Описательных статистик можно придумать огромное количество.

Пример 13

Минимальное значение в числовом столбце датасета — полезная статистика. Если минимальный пассажиропоток на станции метро оказался равен отрицательному значению -234987 человек, значит, в данных выброс. Это ошибка, которую нужно удалить. Если же минимум равен 7200 человек в сутки, а максимум — 70500 человек в сутки, значит, мы узнали кое-что обо всех значениях в столбце датасета с помощью только пары чисел, не открывая датасет с сотнями дней наблюдений.

Пример 14

Разность между 125-м и 32-м наблюдениями в числовом столбце данных — бесполезная статистика. Если эта статистика (а она не имеет названия) равна 3455, мы не узнали о данных ничего важного. Эта статистика бесполезна, потому и не имеет названия.

Нужно знать, уметь находить и понимать смысл самых применяемых и распространенных статистик, у которых, конечно, есть названия в языке статистики. Это позволит подобрать подходящий показатель, чтобы коротко и ёмко описать данные.

1 Виды описательных статистик

Самих статистик очень много. Удобно их разделить на 3 вида:

1. *частотные (frequency-based)*,
2. *основанные на среднем (mean-based)*,
3. *ранговые/порядковые (rank-based/order-based)*.

Пример 15

Конверсия показов рекламы в продажи продукта равна 40% — пример частотной статистики. Интерпретация конверсии звучит так: 40% увидевших рекламу людей купили продукт.

Пример 16

Средний чек в кафе за день 350 рублей — пример усредняющей статистики. Значит, если в кафе было 1000 человек, суммарная выручка составила 350 тысяч рублей за день.

Пример 17

Третья квартиль времени, прошедшего от момента заказа до момента доставки продуктов домой клиентам, составляет 90 минут — пример ранговой статистики. Интерпретация: прилизительно три четверти от всех заказов уложились в 90 минут, а четверть — нет.

[!] Инсайт С интерпретациями примеры гораздо яснее. Их поймёт даже человек, который не знает что такое третья квартиль и конверсия!

Ниже будут даны определения и конверсии, и всех квартилей.

2 Частотные статистики

Частотные статистики описывают количество или долю. Количество или долю чего? Количество или долю события!

Определение 18

Событие (event) — свойство отдельного наблюдения, которое либо истинно, либо ложно для данного наблюдения в датасете. В первом случае говорят, что ***событие наступило (happened)*** или ***событие произошло (occured)***, во втором — ***не наступило*** или ***не произошло.***

Пример 18

События:

- Клиент купил товар? (если наблюдения — клиенты, а купил/не купил — один из столбцов).
- В кафе чек заказа больше 500 рублей? (если наблюдения — заказы в кафе, а сумма чека — один из столбцов).
- Время доставки заказа превысило 60 минут? (если наблюдения — заказы с доставкой, а время доставки — один из из столбцов).

Не события:

- Чек заказа (число, а не «случилось/не случилось»).
- Среднее время доставки всех заказов превысило 60 минут? (нельзя найти по отдельному наблюдению-заказу).

Пример 19

Конверсия рекламы в продажу продукта выше 30%?

- Событие, если наблюдения — продукты, а конверсия — один из столбцов.
- Не событие, если наблюдения — клиенты, а купил/не купил — один из столбцов (невозможно посчитать конверсию по отдельному наблюдению-клиенту).

Теперь можно вернуться к частотным статистикам. Их три.

Определение 19

Количество (count) — количество всех наблюдений в датасете. Обозначение: n

Определение 20

Абсолютная частота события (absolute frequency) — количество наблюдений, в которых событие наступило. Обозначения: $n_A = |A| = \#\{A\}$

Определение 21

Относительная частота события (relative frequency) — доля наблюдений, в которых событие наступило, от некоторого большего количества наблюдений. Обозначения: $f_A = \hat{p}_A = \frac{n_A}{n}$. Обычно подразумевается доля от всех наблюдений датасета. Но далеко не всегда.

Определение 22

Конверсия — относительная частота желаемого события, обычно продажи, подписки, перехода на более дорогой тарифный план. Конверсия — частный случай относительной частоты, а не новая описательная статистика.

Пример 20

В датасете 50 наблюдений — пользователей веб-сервиса. Среди них 30 женщин и 5 женщин купили платную подписку. Из 20 мужчин подписку купили тоже 5.

Опишем датасет:

- Количество наблюдений: 50.
- Абсолютные частоты по полу:

  - Женщины: 30 человек.
  - Мужчины: 20 человек.
- Относительные частоты по полу.
  - Женщины: 60%.
  - Мужчины: 40%.
- Количество платных подписок: 10.
- Конверсия в подписки: 20%.
- Конверсия в подписки для мужчин: 40%. (считается как 5/20, а не 5/50 — доля не от всех наблюдений)
- Конверсия в подписки для женщин: 16.7%

Пример 21

Сумма конверсий 40% и 16.7% не 100%, и не 20% — эти доли вычислялись в разных наборах наблюдений.

[!] Инсайт При виде относительной частоты (доли), первая мысль: «Среди кого эта доля?» Подвох доли, если он есть, может быть только в том, что доля найдена среди каких-то хитро выбранных наблюдений.

3 Сумма и среднее

Сумма баллов, сумма продаж, суммарная прибыль от всех клиентов — показатели с важным экономическим смыслом. Они говорят сами за себя.

Определение 23

Сумма (sum) сумма по всем наблюдениям числовой (!) характеристики обозначается: $\sum\limits_{i=1}^{n} X_i = X_1 + X_2 + \cdots + X_n$, где X_i — значение числовой характеристики i-го наблюдения.
Некорректно считать сумму категорийной характеристики. Исключение: индикатор.

[!] Инсайт *Сумма* индикаторов ***события*** равна ***абсолютной частоте события***

Пример 22

Дан датасет.

Female	Salary (th.rub)
1	40
1	30
0	30

Суммарный доход всех троих $40 + 30 + 30 = 100$ тысяч рублей в месяц. А сумма значений дамми Female равна... количеству женщин в датасете $1 + 1 + 0 = 2$ женщины.

В статистике очень любят среднее по двум причинам:

- Позволяет быстро находить сумму. А сумма имеет экономический смысл.
- Имеет хорошо предсказуемое поведение при увеличении количества собранных данных.

Определение 24

Среднее (mean, average) — это сумма, делённая на количество наблюдений. Обозначается $\overline{X} = mean = avg = \frac{\sum\limits_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$.

[!] Инсайт При виде среднего или суммы первая возникающая мысль: «Среди кого?»

Этим вопросом стоит задаться всерьёз. Например, в датасете средний доход мужчин — 30 000 рублей. Прежде чем размышлять о том, почему он такой низкий или высокий, нужно внимательно посмотреть, среди кого этот доход. Может выясниться, что в огромном датасете только один мужчина. И тогда размышления будут уже не о доходе, а о том, почему только один мужчина попал в датасет.

[!] Инсайт Одно повреждённое наблюдение портит среднее и сумму — нужно очищать данные, прежде чем показывать кому-то среднее и сумму.

Пример 23

В данные по заказам закрался выброс: среди тысячи заказов все по 500 рублей, но один из-за неправильного ввода данных записан как 500 000 рублей.

Настоящая выручка была бы: $\sum\limits_{i=1}^{1000} X_i = 500 + 500 + \cdots + 500 = 1000 \cdot 500 = 500\,000$ рублей.

Но из-за ошибки в данных компьютер вычислит сумму: $\sum\limits_{i=1}^{1000} X_i = 500 + 500 + \cdots + 500 + 500\,000 = 999 \cdot 500 + 500\,000 = 999\,500$ рублей — практически вдвое больше, чем на самом деле пришло денег.

Настоящий средней чек: $\overline{X} = \frac{500\,000}{1000} = 500$ рублей, и это логично ведь каждый чек — 500 рублей. Но из-за ошибки при вводе данных всего одного заказа, компьютер выдаст $\overline{X} = \frac{999\,500}{1000} = 999.5$ рублей, почти вдвое больше.

4 Статистики, основанные на среднем, и показывающие разброс в наблюдениях

Пример 24

Представим, что привлечение клиента стоит 300 рублей. А средний чек с клиента составляет 500 рублей.

Тут возможны разные ситуации, например:

1. все клиенты платят ровно по 500 рублей;
2. половина клиентов платит по 100 рублей, а половина — по 900 рублей.

Эти ситуации важно различать. В первом случае все клиенты приносят одинаковую прибыль 200 рублей, и задача — привлечь как можно больше клиентов. Во втором случае клиенты разные. Одни приносят прибыль 600 рублей, а другие — убыток 200 рублей. И было бы отлично таргетировать привлечение только на клиентов, которые приносят прибыль.

Итак, эти ситуации нужно различать, но средний чек в обеих одинаковый — 500 рублей.

Ситуации различаются разбросом значений чека.

1. Разброса значений никакого нет: 500, 500, 500, 500, ...
2. Разброс (большой?) между значениями: то 100, то 900.

Необходимо научиться измерять разброс между значениями в разных наблюдениях. Понадобится вспомогательное определение.

Определение 25

Централизация/отклонение от среднего (de-meaned value, mean-centered value) — это отклонение числовой характеристики конкретного наблюдения от среднего значения этой характеристики по всем наблюдениям:

$X_i - \overline{X}$.

Централизация не является описательной статистикой для измерения разброса. Централизация — это не одно число для датасета, а по одному числу для каждого наблюдения. Зато на основе централизаций такую статистику можно построить, и не одну.

Дадим определения сразу трёх, а затем покажем на примере их все, а заодно и централизацию.

Определение 26

Среднее абсолютное отклонение (mean absolute deviation = MAD) — это среднее от модулей централизаций: $MAD(X) = \frac{1}{n}\sum\limits_{i=1}^{n}|X_i - \overline{X}|$

Определение 27

Среднее квадратичное отклонение (mean squared deviation = MSD) — это среднее от квадратов централизаций: $MSD(X) = \frac{1}{n}\sum\limits_{i=1}^{n}(X_i - \overline{X})^2$

Определение 28

Неисправленное/смещённое стандартное отклонение (biased standard deviation) — это корень из среднего квадратичного отклонения: $sd(X) = \sqrt{MSD(X)} = \sqrt{\frac{1}{n}\sum\limits_{i=1}^{n}(X_i - \overline{X})^2}$

[!] Инсайт Все отклонения так или иначе измеряют разброс между значениями характеристики.

Пример 25

У официанта было 4 заказа с чеками на суммы: $500,500,1500,700$ рублей.

Найдём среднюю сумму чека у официанта: $\overline{X} = \frac{500+500+1500+700}{4} = 800$ рублей.

Найдём централизацию каждого чека: $(500-800,500-800,1500-800,700-800) = (-300,-300,700,100)$

Например, централизация чека со второго заказа равна минус 300 рублей. Значит, чек со второго заказа был на 300 рублей меньше, чем средний чек официанта.

Найдём $MAD(X) = \frac{-300-300+700+100}{4} = 350$ рублей — чеки отличаются от среднего чека на 350 рублей в среднем в ту или иную сторону.

Найдём $MSD(X) = \frac{(-300)^2+(-300)^2+700^2+100^2}{4} = 170000$ квадратных рублей. Это число, конечно, тоже как-то связано с разбросом, но мы не рекомендуем пытаться объяснить незнакомым людям, что такое рубли в квадрате

Найдём $sd(X) = \sqrt{170000} \approx 412.31$ рублей — чеки отличаются от среднего чека на 412.31 в среднеквадратичном смысле. Как видим, стандартное отклонение тоже не особенно интерпретируется

[!] Инсайт Стандартное отклонение позволяет многое сказать о количестве наблюдений в определённом интервале с использованием неравенства Чебышёва или Центральной Предельной Теоремы, оно имеет большой смысл при тестировании гипотез, работает ли модель на данных или нет, но пока мы не изучили ничего из перечисленного, и смысла в стандартном отклонении не видим

[!] Инсайт Стандартное отклонение (пока что) — это статистика, которая как-то показывает разброс, но напрямую не интерпретируется.

Не добавляйте стандартное отклонение в слайды, пока не можете объяснить, зачем оно в них нужно

[!] Инсайт Есть два разных стандартных отклонения: исправленное и неисправленное — при большом числе наблюдений они дают практически одинаковое значение, но при маленьком — нет, и нужно не перепутать.

Применение стандартного отклонения с использованим неравенства Чебышёва есть в необязательной задаче домашней работы чёрного уровня.

5 Ранговые статистики

Но что делать, если данные нечисловые?

> Пример 26
>
> «Неоконченное среднее», «Полное среднее», «Неоконченное высшее», «Высшее», и т. п. — уровни образования нельзя сложить и поделить на количество. Некорректно искать среднее между «Неоконченным высшим» и «Высшим», и тем более среднее абсолютное отклонение среди них.

Не только для числовых, но и для упорядоченных категорий можно находить ранговые статистики: значения, меньше которых определённая доля наблюдений, и больше которых — оставшаяся доля наблюдений.

Определение 29

Минимум (min) — наименьшее значение среди всех наблюдений. Обозначения: min.

Определение 30

Максимум (max) — наименьшее значение среди всех наблюдений. Обозначения: max.

Определение 31

Медиана (median) — такое значение, что приблизительно половина наблюдений меньше медианы, а половина больше медианы. Обозначения: $median$.

Определение 32

Квантиль (quantile) с уровнем p — такое значение, что приблизительно доля p наблюдений меньше p-квантили, а $1-p$ больше p-квантили. Обозначения: $quantile(X,p)$.

Определение 33

Персентиль (percentile) с уровнем $p\%$ — такое значение, что приблизительно $p\%$ наблюдений меньше p-персентили, а $(100-p)\%$ больше p-персентили. Обозначения: $percentile(X,p)$.

Определение 34

Нижняя квартиль/первая квартиль (lower quartile, first quartile) — такое значение, что приблизительно 25% наблюдений меньше нижней квартили, а 75% больше. Обозначения: $Q_1 = LQ$.

Определение 35

Верхняя квартиль/третья квартиль (upper quartile, third quartile) — такое значение, что приблизительно 75% наблюдений меньше нижней квартили, а 25% больше. Обозначения: $Q_3 = UQ$.

[!] Инсайт Названий много — суть одна:

Медиана $=$ вторая квартиль $=$ 50%-персентиль $=$ 0.5-квантиль.

Медиана разделяет 50% самых маленьких и 50% самых больших значений.

Первая квартиль $=$ 25%-персентиль $=0.25$-квантиль.

Первая квартиль разделяет четверть самых маленьких и три четверти самых больших значений.

[!] Инсайт Названий ещё больше:

третья дециль $=$ 30%-персентиль $=0.3$-квантиль.

Пример 27

Ниже приведена статистика по зарплатам выпускников.

Зарплата выпускника (тыс. руб в месяц)

	mean	min	Q1	median	Q3	max
Университет A	80	20	30	80	130	140
Университет B	80	50	60	80	100	110

- Средние зарплаты не отличаются, но вот ранговые статистики показывают интересные выводы.
- Самые сильные 25% выпускников Университета A зарабатывают от 130 до 140 тысяч рублей в месяц.
- Самые сильные 25% выпускников Университета B зарабатывают от 100 до 110 тысяч рублей в месяц.
- Если абитуриент планирует учиться и выбирает между A и B — есть аргумент выбирать A.
- Самые слабые 25% выпускников A зарабатывают от 20 до 30 тысяч рублей в месяц.
- Самые слабые 25% выпускников B зарабатывают от 50 до 60 тысяч рублей в месяц.
- Если абитуриент после поступления не планирует напрягаться, есть аргумент выбирать B.

Со способом расчёта ранговых статистик будут трудности.

Что такое первая квартиль для зарплаты трёх человек? Меньше неё должна быть зарплата у 25% человек. Но что такое 25% от трёх человек? Это сколько человек? Дробное количество — ерунда получается! А если нужна не квартиль, а 17%-персентиль?

[!] Инсайт Нет идеального способа вычислять ранговые статистики. Есть приемлемый. Придётся округлять.

Напоминание из школы: числа можно округлять вверх и вниз до целого.

Определение 36

floor(x) — округление x вниз, ***ceil(x)*** — округление x вверх.

Чтобы найти p-квантиль, нужно:

1. Упорядочить значения по возрастанию:
   - новые числа обозначаются с круглыми скобками в индексах $X_{(i)}$ и нумеруются ***с нуля***.
2. Вычислить ***локацию*** по формуле ниже:
   - $i = p \cdot (n - 1)$
3. Вычислить квантиль, наконец-то!

- $quantile(X,p) = X_{(floor(i))} + (i - floor(i)) \cdot (X_{(ceil(i))} - X_{(floor(i))})$

Пример 28

Найдём третью квартиль среди чисел: $40,\ 20,\ 70$. Цель примера — разобраться, как происходит расчёт. Интерпретации здесь нет, интерпретацию мы видели в примере выше с зарплатами выпускников. Итак, к расчётам.

Третья квартиль это 0.75-квантиль.

1. Упорядочиваем числа: $40, 20, 70 \longrightarrow \overbrace{20}^{X_{(0)}}, \overbrace{40}^{X_{(1)}}, \overbrace{70}^{X_{(2)}}$
2. $i = 0.75 \cdot (3 - 1) = 1.5$, округляем вниз: $floor(1.5) = 1$ и вверх $ceil(1.5) = 2$
3. $Q_3 = quantile(X,0.75) = X_{(1)} + (1.5 - 1) \cdot (X_{(2)} - X_{(1)}) = 40 + (1.5 - 1) \cdot (70 - 40) = 55$

Правда ли, что меньше 55 ровно 75% наблюдений? Нет, конечно, 75% от 3-х наблюдений будет 2.25 наблюдений, нецелое число. Меньше 55 имеется 66.7% наблюдений, а не 75%, но учитывая, что у нас всего 3 наблюдения — неплохое приближение.

[!] Инсайт Если упорядоченные категории закодированы числами, например 2 — неоконченное высшее, 3 — высшее образование, то любые дробные числа не имеют смысла: 2.5 не соответствует никакой категории. Поэтому если оказалось $Q_3 = 2.72$, то правильная интерпретация: около 25% людей имеют уровень образования 3-й и выше (высшее и более одного высшего, степень и т п), а около 75% людей имеют уровень образования не выше 2-го: (неоконченное высшее или ниже).

6 Когда можно применять статистики

Какие виды статистик можно применять зависит от типа данных:

- Неупорядоченные категории — только частотные статистики.
- Упорядоченные категории — частотные или ранговые.
- Числовые — все виды.

Пример 29

Большой датасет после очистки данных имеет вид:

AgeGroup	Salary (th.rub/month)	Female
18-21	30	0
...	...	...

Первый этап его описания, показать статистики, описывающие:

- Весь датасет — число наблюдений.
- Характеристику Female — только частотные статистики. Формально говоря, можно посчитать сумму и среднее, но они не дадут ничего нового. Сумма совпадёт с количеством женщин, а среднее — с долей женщин.

- Характеристику AgeGroup — частотные статистики и ранговые статистики.
- Характеристику Salary — полный набор всех статистик, что мы определяли выше.

Пример результата:

Female	Count	Frequency
0	2800	0.56
1	2200	0.44
Всего	5000	1.00

Statistics	Female
sum	2200
mean	0.44

Мужчин в датасете чуть больше, чем женщин.

AgeGroup	Count	Frequency
18-21	700	0.14
22-24	900	0.18
25-29	900	0.18
30-34	800	0.16
35-39	600	0.12
40-44	600	0.12
45-49	300	0.06
50-70	200	0.04
Всего	5000	1.00

Statistics	AgeGroup
min	18-21
Q1	22-24
med	25-29
Q3	35-39
max	50-70

По любой таблице видно, что возраст людей в датасете — от 18 до 70 лет.

По правой таблице — возраст приблизительно четверти людей до 24 лет, примерно четверти людей от 35 лет.

Более точно по левой — 32% до 24 лет и 36% от 35 лет.

По правой таблице — приблизительно половина до 29 и приблизительно половина от 25.

Более точно по левой — 50% до 29 лет и 68% от 25 лет.

Левая таблица точнее, но пользоваться ей тем труднее, чем больше категорий.

SalaryGroup	Count	Frequency
16-50	3500	0.70
50-100	1350	0.27
100-200	140	0.028
200-800	10	0.002
Всего	5000	1.00

Statistics	Salary
min	16
Q1	20
med	30
Q3	70
max	800
count	5000
mean	40
mad	35
sd	30

Чтобы сделать таблицу частот, приходится придумывать события. Например здесь это «зарплата от 16 до 50 тысяч рублей в месяц», или «от 50 до 100 тысяч рублей в месяц».

> Причём из таблицы неясно, куда включается ровно 50 тысяч рублей, если такое было. А ведь важно, чтобы человек не был посчитан дважды и попал только в одну из категорий. Кроме того, мы не особенно угадали с категориями: 70% попало в категорию «зарплата от 16 до 50 тысяч рублей в месяц», но у скольки из них 16, а у сколько 50 — неясно.
>
> Правая таблица сразу показывает, что у половины зарплата до 30 тысяч рублей в месяц. А у четверти людей — вообще от 16 до 20 тысяч рублей в месяц.
>
> Ещё мы видим, что у 25% людей зарплата от 70 до 800 тысяч рублей в месяц — звучит потрясающе! Но из левой таблицы можно уточнить, что от 100 до 800 тысяч рублей лишь у 3% людей. Так что у остальных 22% от 70 до 100 тысяч рублей, уже не так классно. Видим, что средняя зарплата — 40 тысяч рублей в месяц, и люди отклоняются от неё в среднем на 35 тысяч рублей. Это потому, что очень-очень мало людей отклоняются от неё на 760 тысяч рублей в плюс, а довольно много отклоняется в плюс или в минус на 5, на 7, на 20 тысяч рублей: вот в среднем отклонение и выходит 35 тысяч рублей.
>
> На втором этапе описания датасета можно было бы отдельно рассмотреть зарплаты мужчин и женщин и сравнить. Или зарплаты в каждой категории по возрасту. Или в сочетаниях категорий пола и возраста. Но это уже совсем другая история...

Итоги

Надеемся, этот материал помог понять:

- что такое датасеты, как их открыть и посмотреть в pandas;
- что находится в строках и столбцах датасета, что такое наблюдения и характеристики;
- что такое интерпретация, зачем это понятие нужно и какие есть правила интерпретации значения;
- что такое кросс-секция, временной ряд и панельные данные;
- почему важно отличать числовой тип значения и категориальные: ordinal и nominal;
- как искать выбросы, как они помогают найти некорректные, нерелевантные и искажённые значения;
- как и зачем проводить очистку данных;
- когда и как можно найти частотные статистики: абсолютную, относительную, и что они показывают;
- когда и как можно применять статистики, основанные на среднем;
- что такое разброс между значениями, зачем его измерять;
- как найти и проинтерпретировать среднее абсолютное отклонение;
- как найти среднеквадратичное отклонение, почему не нужно пытаться его интерпретировать;
- когда и как можно найти ранговые статистики, и что они показывают;
- почему хотя ранговых статистик много, на самом деле они все опираются на p-квантиль.

В этих последних строках надеемся, что хоть что-нибудь было понятно и хоть что-нибудь было новым. Громадный респект всем, кто дочитал!