ПРИКЛАДНАЯ СТАТИСТИКА

Описательные статистики

«Чиновники едят мясо, я — капусту. В среднем мы едим голубцы.»

Анекдот категории Б

Упражнение 1 (Описательные статистики)

Коллекционер Настя собрала целых 10 наблюдений и записала их в табличку. Теперь Настя хочет стать аналитиком и проанализировать таблицу. Помогите ей.

имя	пол	возраст	вес
Ария	ж	14	50
Санса	ж	15	50
Джон	м	21	50
Дэнни	ж	20	50
Бран	м	14	80
Сандор	м	25	80
Гора	м	30	440
Тирион	м	23	80
Теон	ж	22	80
Якен	м	16	80

а) Что такое непрерывная переменная? Что такое категориальная переменная? Какие переменные в табличке относятся к непрерывным? Какие к категориальным? Приведите ещё примеров непрерывных и категориальных переменных!

б) Найдите долю мужчин и женщин в выборке. Постройте для пола гистограмму.

в) Найдите средний возраст и медианный возраст. Что означают эти числа. В чём они измеряются?

г) Найдите дисперсию возраста. В чём измеряется эта величина? Зачем обычно ищут среднее квадратичное отклонение? Найдите его.

д) Постройте гистограмму для возраста. Считайте, что ширина одного столбца — 5 лет. Если человек попадает на правую границу отрезка, он попадает в текущий столбец. Изобразите на гистограмме среднее, медиану. Как бы вы нарисовали на гистограмме стандартное отклонение?

е) Что такое выброс? Есть ли выбросы в возрасте? Есть ли выбросы в весе? Как выглядит выброс на гистограмме? Найдите средний вес и медианный вес. Чем медиана в данном случае лучше, чем среднее?

ё) Чувствительна ли дисперсия к выбросам?

ж) Что такое мода? Почему использовать её для непрерывных переменных не очень хорошая идея? Найдите моду для имени, пола и возраста.

з) Что такое квантиль? Предложите способ борьбы с выбросами, основанный на знании того, что такое квантиль.

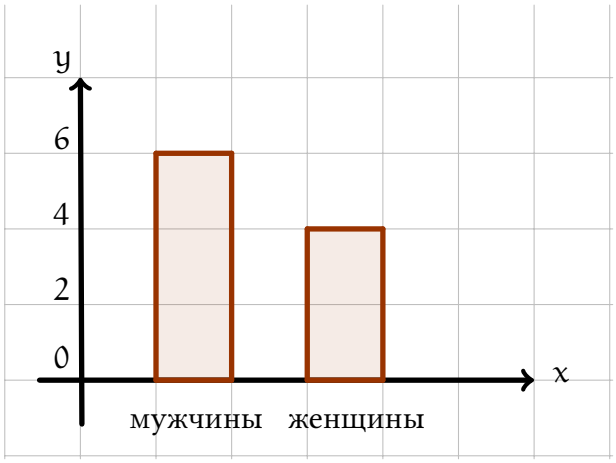
и) Постройте для возраста эмпирическую функцию распределения.

Решение:

а) **Непрерывная переменная** не ограничена каким-то конечным набором значений и может принимать любые числовые значения. Например: цена на квартиру, валютный курс, возраст, число лайков под фото и т.п.

Категориальная переменная принимает значения из какого-то фиксированного конечного множества. Например: пол, марка машины и тп.

б) В выборке 6 мужчин и 4 женщины. Всего 10 человек. Значит доля мужчин $\frac{6}{10} = 0.6$, доля женщин $\frac{4}{10} = 0.4$. Нарисуем гистограмму. По оси x будем откладывать возможные значения для нашей переменной, по оси y насколько часто это значение наблюдается в выборке.



в) Найдём **средний возраст**. Для этого сложим все числа и поделим их на количество наблюдений

$$\frac{1}{10} \cdot (14 + 15 + 20 + 21 + 14 + 25 + 30 + 23 + 22 + 16) = 20.$$

Средний возраст это 20 лет. Формула для подсчёта среднего выглядела так:

$$\bar{x} = \frac{1}{n}(x_1 + ... + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Привыкайте к формулам. Они будут часто встречаться вам по жизни.

Чтобы найти медиану, нам нужно упорядочить всех людей из выборки по возрасту и посмотреть какое число оказалось в середине:

14 14 15 16 20 21 22 23 25 30

У нас в середине находятся сразу два числа. В случаях, когда такое происходит в качестве медианы берут среднее этих двух чисел. В нашем случае это $^1/_2 \cdot (20 + 21) = 20.5$. Медиана это число, взятое посередине. Половина выборки оказывается слева от него, а вторая половина справа. Среднее и медиана в нашей задачке измеряются в годах и обозначают типичный возраст, который присущ людям из выборки.

г) **Дисперсия** — это мера разброса. Она показывает насколько разнообразными могут быть элементы в выборке, насколько сильно они могут отклоняться от своего типичного значения.

Чтобы найти её, нужно посмотреть насколько сильно каждый представитель в выборке отличается от текущего. Величина такого отличия называется отклонением.

Предположим, что Алёне и Карине обеим по 18 лет. При рассмотрении этих двух человек средний возраст вычисляется как средняя сумма двух слагаемых: $\frac{18+18}{2} = 18$. Тогда отклонением для Алёны и Карины от среднего возраста будет $18 - 18 = 0$ года, а среднее отклонение нулевое: $\frac{0+0}{2}$. То есть в нашей выборке из двух человек нет неоднородности, у всех людей одинаковый возраст.

Предположим теперь, что Алёне 18 лет, а Карине 22 года. Средний возраст всё ещё вычисляется как средняя сумма двух слагаемых: $\frac{18+22}{2} = 20$. Тогда отклонением для Алёны от среднего возраста будет $18 - 20 = -2$ года. Для Карины отклонением будет $22 - 20 = 2$ года. Если посчитать среднее этих отклонений, мы получим $\frac{-2+2}{2} = 0$. То есть в выборке нет никакого разброса. Все люди не отличаются от среднего. Это неправда.

Для того, чтобы избежать неправды и жить по правде, отклонения возводят в квадрат, тогда мы получаем, что суммарное отклонение будет $(-2)^2 + 2^2 = 4 + 4 = 8$. Посмотрев на такое число мы сразу же поймём, что в выборке есть неоднородность.

Среднее значение квадратов отклонений от среднего и называется дисперсией. Давайте найдём её. Ещё раз выпишем наши наблюдения:

$$14 \quad 14 \quad 15 \quad 16 \quad 20 \quad 21 \quad 22 \quad 23 \quad 25 \quad 30$$

Сначала из каждого вычитаем среднее. Это даст нам

$$-6 \quad -6 \quad -5 \quad -4 \quad 0 \quad 1 \quad 2 \quad 3 \quad 5 \quad 10.$$

Теперь возводим все отклонения в квадрат

$$36 \quad 36 \quad 25 \quad 16 \quad 0 \quad 1 \quad 4 \quad 9 \quad 25 \quad 100.$$

Складываем их. Получается 252. Остаётся разделить это число на 10 (количество наблюдений). Получается, что дисперсия составит 25.2 квадратных года. Из-за того, что мы каждое слагаемое возводили в квадрат, **дисперсия измеряется в квадратных годах.**

Когда мы умножаем одну сторону квадрата, измеренную в метрах, на другую, мы получаем его площадь. Она измеряется в квадратных метрах. Тут похожая ситуация. Мы бы хотели вернуться назад, к обычным годам. Для этого из дисперсии извлекают корень и получают штуку под названием стандартное отклонение. В нашем случае получится $\sqrt{25.2} \approx 5.02$ года.

Можно найти дисперсию проще и быстрее. Для этого есть специальная формула:

$$\hat{\sigma}^2 = \overline{x^2} - (\bar{x})^2,$$

то есть дисперсия это среднее квадратов минус квадрат среднего. Эту формулу довольно просто доказать. На матстате вы её докажете. А пока просто воспользуемся ей. Найдём квадрат среднего:

$$(\bar{x})^2 = 20^2 = 400$$

Теперь среднее квадратов:

$$\overline{x^2} = \frac{1}{10} \cdot (14^2 + 15^2 + 21^2 + 20^2 + 14^2 + 25^2 + 30^2 + 23^2 + 22^2 + 16^2) = 425.2.$$

Остался последний штрих:

$$\hat{\sigma}^2 = 425.2 - 400 = 25.2.$$

Пользуйтесь тем способом, который вам больше нравится. Про дисперсию давайте обсудим ещё пару дополнительных полезных нюансов:

- Мы возводим отклонения в квадрат не только для того, чтобы сделать все числа положительными. Попутно мы подчёркиваем, что чем больше отклоняется возраст от среднего, тем это хуже. Так штраф за отклонение в два года составит 4, а за отклонение в три года штраф будет 9. С подобной логикой мы ещё встретимся, когда будем обсуждать различные метрики, используемые в машинном обучении.

- Часто при подсчёте дисперсии вместо формулы

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

которую использовали мы, используют

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

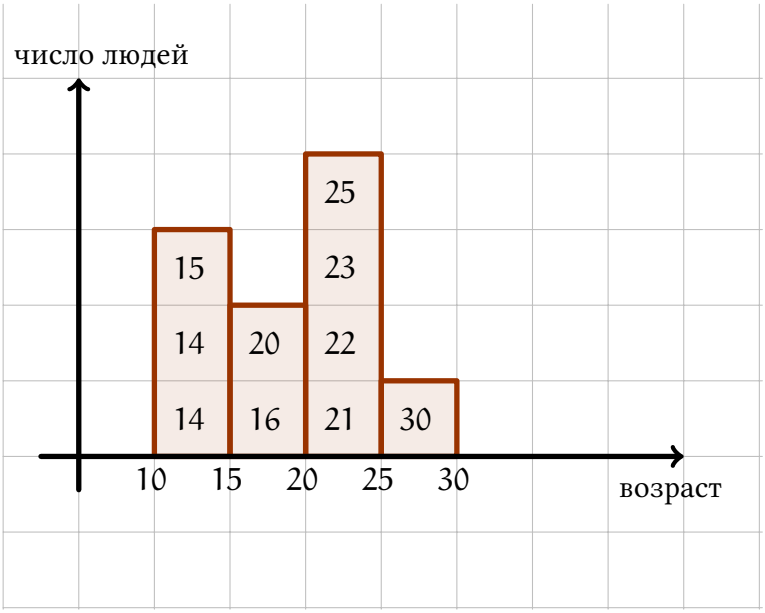
Вторая формула, на самом деле, корректнее, чем первая. В *pandas* используется именно она. У этого есть глубокая причина, которая называется **несмещённостью оценки.** О том, что это мы поговорим позже.

- Если распределение у данных *нормальное*, тогда большая часть выборки, а именно 69% кучкуется в диапазоне между $\bar{x} - \hat{\sigma}$ и $\bar{x} + \hat{\sigma}$.

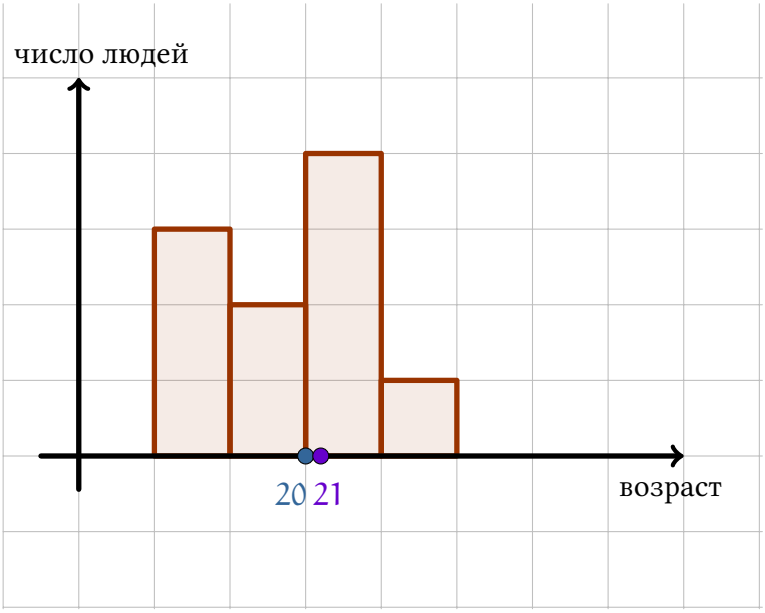
  При этом 95% выборки находится между $\bar{x} - 2\cdot\hat{\sigma}$ и $\bar{x} + 2\cdot\hat{\sigma}$, а 99.9% выборки находятся между $\bar{x} - 3\cdot\hat{\sigma}$ и $\bar{x} + 3\cdot\hat{\sigma}$.

  Правила таких кучкований называют правилом одной, двух и трёх сигм. Их часто используют для проведения АБ-тестов. Об этом мы тоже поговорим ближе к концу курса.

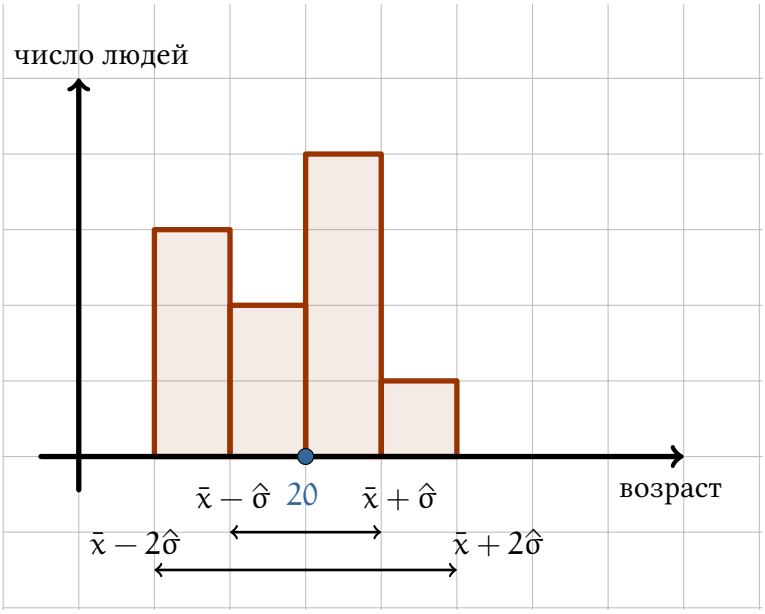
д) Отмечаем по оси x каждые 5 лет, как сказано в условии задачи. Для всех людей, попавших в этот отрезок рисуем столбик высоты равной количеству людей, попавших в отрезок. Если человек попадает в правую границу отрезка, он попадает и в столбик. Например, 20 — это правая граница второго отрезка. Все люди, которым 20 лет попадают во второй столбик. Это просто договорённость о том, что делать на границе, в спорной ситуации. Не более того.



Отлично! Гистограмма готова. Возраст каждого человека, которого мы внесли в тот или иной столбец, мы подписали. Давайте отметим на гистограмме медиану и среднее значение. Как это ни странно, они оказываются в ”центре”распределения.



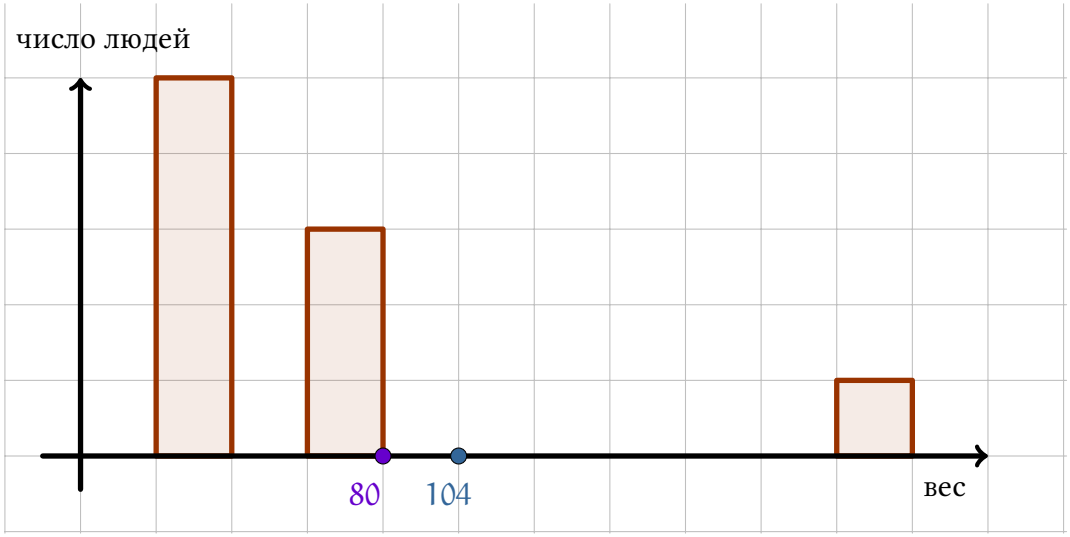
Выше мы обсудили, что стандартное отклонение — величина, которая описывает вариацию выборки вокруг среднего значения и поговорили про правила сигм. Давайте нарисуем от среднего отступы на сигмы вправо и влево.



e) В возрасте всё хорошо. В весе есть выброс, кто-то слишком много ест. Давайте найдём среднее и медиану. Среднее окажется равно $\frac{1040}{10} = 104$. Медиана окажется равна 80. Видим, что выброс существенно сдвинул среднее значение веса в большую сторону. Из-за этого оно перестало отражать типичный вес человека из выборки. Наше представление о людях оказалось искажено.

Медиана, в отличие от среднего, оказывается нечувствительна к выбросам. Это происходит из-за способа её поиска. Мы упорядочиваем наблюдения по порядку и смотрим на то, какое в середине. Значение выброса никак не участвует в подсчёте медианы и именно из-за этого не искажает её.

На гистограмме переменным, в которых есть выбросы соответствуют очень длинные хвосты. В нашем случае именно так и произошло:



ё) К несчастью, **дисперсия чувствительна к выбросам.** Когда мы считаем её, мы возводим все разности в квадрат. Разница между средним и выбросом будет большой. Когда мы возведем её в квадрат и прибавим к дисперсии, она очень сильно увеличится.

ж) Мы с вами определили моду как самое часто встречаемое значение признака в выборке. Для пола модной будут мужчины. Для веса модой будет 80. Для возраста модой будет 14.

Для непрерывных переменных использовать моду в качестве меры типичности довольно глупо. Часто бывает так, что непрерывные признаки довольно близки друг к другу, но немного различаются. **Чаще всего моду используют, чтобы охарактеризовать именно категориальные переменные. Смотрят на пару: мода, её частота.**

На самом деле моду можно определить так, чтобы она была корректна и для непрерывных признаков. Обычно говорят, что мода это самое вероятное значение в выборке. Это позволяет найти её по плотности распределения (грубо говоря, по гистограмме): моде соответствует самая высокая точка плотности. Подробнее об этом вы узнаете на теории вероятностей.

з) На вопрос что такое квантиль, нам поможет ответить медиана. Мы сказали с вами, что если отсортировать выборку по возрастанию, то в середине у неё окажется медиана.

$$14 \quad 14 \quad 15 \quad 16 \quad 20 \quad 21 \quad 22 \quad 23 \quad 25 \quad 30$$

Получается, что 50% выборки больше медианы, и 50% выборки меньше медианы. Медиана — это 50% квантиль. По аналогии можно придумать другие квантили. Например, ниже красным отмечены 30% и 70% квантили:

$$14 \quad 14 \quad 15 \quad 16 \quad 20 \quad 21 \quad 22 \quad 23 \quad 25 \quad 30$$

Ровно 30% выборки $\leqslant 15$ и 70% больше 15. И наоборот в случае 22. Среднее и медиана помогают понять какие представители типичны для середины распределения. Квантили помогают понять какие представители типичны для разных кусков распределения. Например, если мы имеем дела со стоимостью недвижимости, мы можем понять какая стоимость квартир типична для элитных районов.

Как мы выяснили выше, **выбросы могут существенным образом искажать наши пред-**

ставления о выборке. От них нужно выборку очищать. Один из способов: отрубить все наблюдения, которые находятся выше 99% квантиля и все наблюдения, которые находятся ниже 1% квантиля. Все выбросы такой процедурой будут убиты и мы сможем спокойно работать с выборкой. Иногда берут 95% и 5% квантили.

и) Снова для удобства выпишем все наши наблюдения в порядке возрастания.

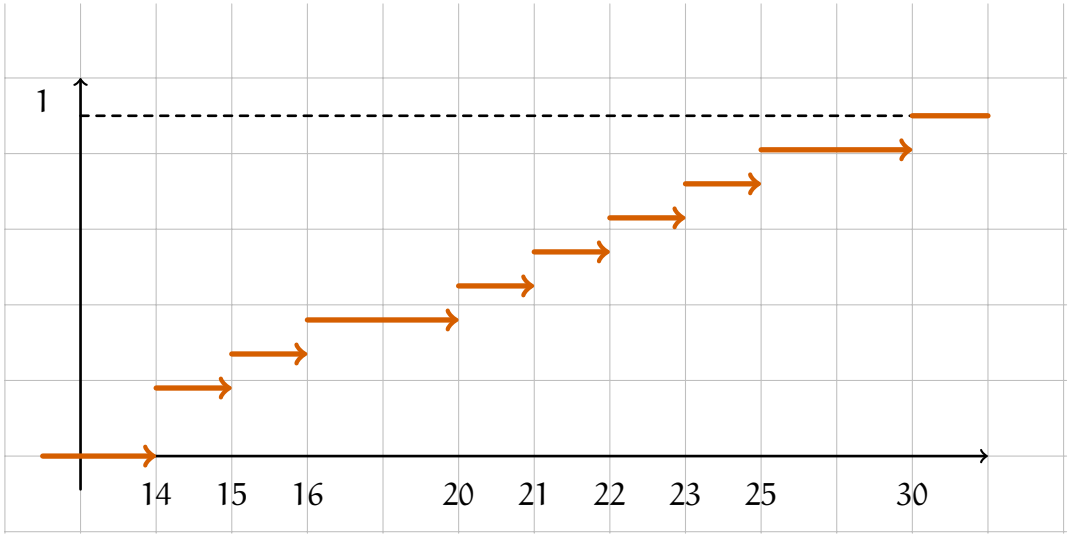
$$14 \quad 14 \quad 15 \quad 16 \quad 20 \quad 21 \quad 22 \quad 23 \quad 25 \quad 30$$

Функция распределения описывает то, как накапливается вероятность события $P(X \leqslant x)$ по мере увеличения x. Эмпирическая функция распределения представляет из себя частоты таких событий

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n}[x_i \leqslant x]$$

Будем перебирать x от маленьких значений к большим и оценивать такие частоты. Когда x левее 14, ни один элемент выборки x_i не оказывается меньше x. Частота интересующего нас события равна нулю.

Когда x между 14 и 15, два значения x_i удовлетворяют условию. Частота прыгает на 0.2. Вторая ступенька появляется в точке 15. Мы поднимаемся до 0.3. По аналогии скачки происходят в следующих точках выборки. В итоге мы получаем следующую кусочно-линейную функцию:



Постепенно, по ступенькам, мы дошли от 0 до 1. Мы пытаемся с помощью ступенек описать какую-то теоретическую функцию распределения. Мы это делаем из-за того, что на практике довольно удобно описывать случайную величину через её функцию распределения.