Семинар 4: основы статистики!

Задача 1

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
Кхал	м	14	80
Санса	ж	16	40
Мелисандра	ж	20	40
Эддард	м	20	80
Сандор	м	14	80
Миссаедея	ж	25	40
Якен	м	30	80
Теон	ж	23	40
Тирион	м	22	80
Станис	м	16	440

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 измеряются эта величина? Зачем обычно ищут среднее квадратическое 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?

ж) Чувствительна ли дисперсия к выбросам?

з) Что такое мода? Почему использовать её для непрерывных переменных не очень хорошая идея? Найдите моду для имени, пола и возраста. (уточнить что это дискретная мода и тп)

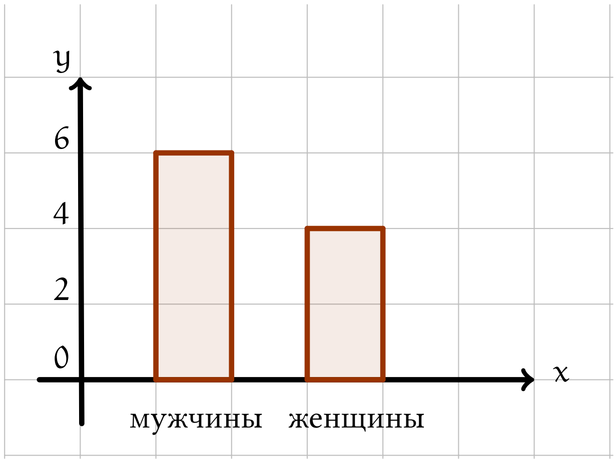
и) Что такое квантиль? Предложите способ, борьбы с выбросами, основанный на знании того что такое квантиль...

Решение:

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

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

Средний возраст это 20 лет. Формула для подсчёта среднего выглядела как

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

Привыкайте к формулам. Они будут часто встречаться вам по жизни. Чтобы найти медиану нам нужно упорядочить всех людей из выборки по возрасту и посмотреть на середину получившегося ряда.

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

У нас в середине находятся сразу два человека. Медианой будет их среднее, то есть 20 лет. Грубо говоря, половина нашей выборки оказывается слева от этого числа, а вторая справа. Медиана находится в серединке. Оба числа 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. Чтобы найти её, нужно посмотреть насколько сильно каждый представитель в выборке отличается от текущего. Величина такого отличия называется отклонением. Предположим, что Алёне 18 лет. Карине 22 года. Тогда отклонением для Алёны от среднего возраста будет $18 - 20 = -2$ года. Для Карины отклонением будет $22 - 20 = 2$ года.

Если просуммировать эти отклонения, мы получим $-2 + 2 = 0$. То есть в выборке нет никакого разброса. Все 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:

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

Сначала из каждого вычитаем среднее. Это даст нам вектор

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

Теперь возводим все отклонения в квадрат

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

Складываем их! Получается 242. Остаётся разделить это число на 10 (количество наблюдений). Получается, что дисперсия составит 24.2 квадратных года. Из-за того, что мы каждое слагаемое возводили в квадрат, дисперсия измеряется в квадратных годах.

Когда мы умножаем одну сторону квадрата на другую, мы получаем его площадь. Она измеряется в квадратных метрах. Тут похожая ситуация. Мы бы хотели вернутся назад, к обычным годам. Для этого из дисперсии извлекают корень и получают штуку под названием стандартное отклонение. В нашем случае получится 4.9 года.

Здесь нам осталось обсудить пару 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, 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

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

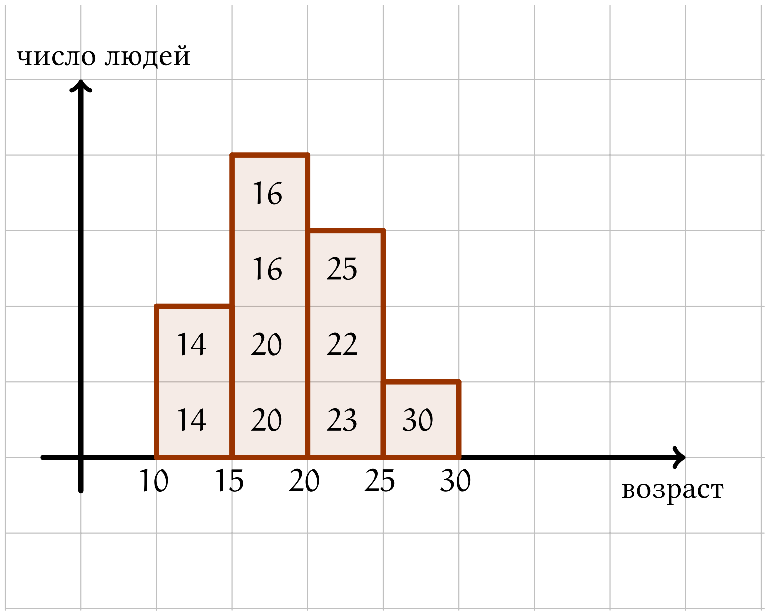
Вторая формула на самом деле корректнее, чем первая. В питоне используется именно она. У этого есть глубокие причины. В полной мере их вы узнаете в курсе по математической статистике. Мы вкратце скажем об этом ближе к концу курса, когда будем говорить про АБ-тесты. Пока держите это в голове, как вопрос, на который у вас нет ответа. Надеюсь, что это будет как следует мучать вас по ночам и стимулировать ботать.

- Если распределение у данных нормальное (что такое нормальное распределение — отдельный и очень важный вопрос), тогда большая часть выборки, а именно 69% кучкуется в диапазоне между $\bar{x} - \hat{\sigma}$ и $\bar{x} + \hat{\sigma}$.

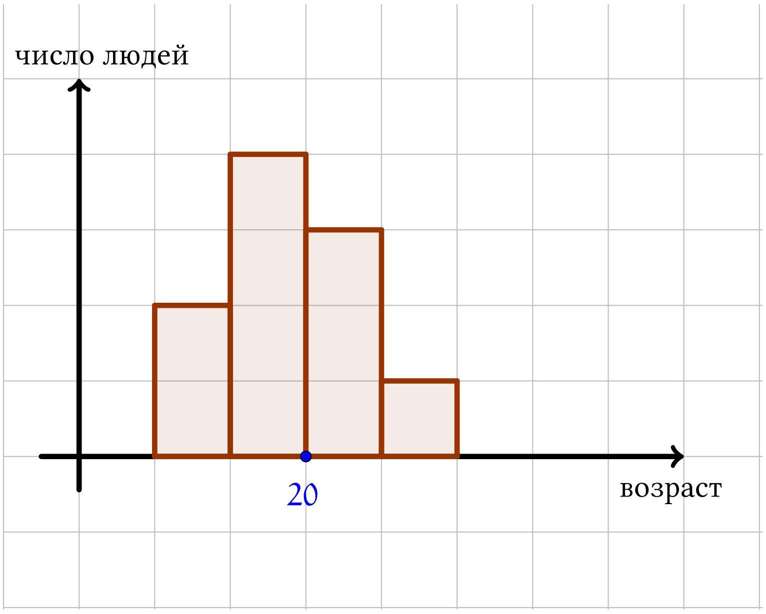
  При этом 95% выборки находится между $\bar{x} - 2 \cdot \hat{\sigma}$ и $\bar{x} + 2 \cdot \hat{\sigma}$, а 99.9% выборки находятся между $\bar{x} - 3 \cdot \hat{\sigma}$ и $\bar{x} + 3 \cdot \hat{\sigma}$.

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

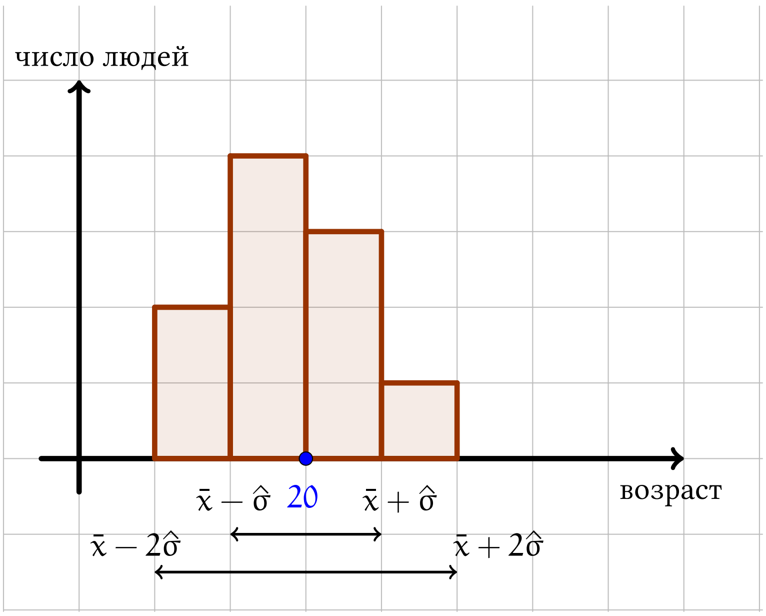
д) Отмечаем по оси x каждые 5 лет, как сказано в условии задачи. Для всех людей, попавших в этот отрезок рисуем столбик высоты равной количеству людей, попавших в отрезок. Если человек попадает в **правую** границу отрезка, он попадает и в столбик. Например, 20 — это правая граница второго отрезка. Все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



Отлично! Гистограмма готова. Каждого человека, которого мы внесли в тот или иной столбец, мы подписали. Давайте отметим на гистограмме медиану и среднее значение. Как это не 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 $\frac{1000}{10} = 100$. 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. Мы посмотрим на такие гистограммы на копах. .

ж) К несчастью, да. Когда мы считаем её, мы возводим все разности в квадрат. Грубо говоря, разница между средним и выбросом будет большой. Когда мы возведем её в квадрат и прибавим к дисперсии, она очень сильно увеличится.

з) 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 либо 20 либо 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. И после моду ищут по плотности распределения (грубо говоря, по гистограмме), пытаясь понять какому числу соответствует её самая высокая точка. Но об этом вы узнаете на теории вероятностей.

и) 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.

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

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:

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

Ровно 30% меньше 16 и 70% больше 16. И наоборот в случае 23. Среднее и медиана помогают понять какие представители типичны для середины распределения. Квантили помогают понять какие представители типичны для разных кусков распределения.

Как мы выяснили выше, выбросы могут существенным образом искажать наши представления о выборке. От них нужно выборку очищать. Один из способов: отрубить все наблюдения, которые находятся выше 99% квантиля и все наблюдения, которы находятся ниже 1% квантиля. Все выбросы такой процедурой будут убиты и мы сможем спокойно работать с выборкой.

Ещё задачи!

Тут находится несколько задачек, о которых вам нужно подумать самостоятельно. Возможно, что похожие задачи попадутся вам на самостоятельной работе.

Задача 2

Имеется пять чисел: x, 9, 5, 4, 7. При каком значении x медиана будет равна среднему? А можно ли поставить такие цифры в условии задачи, чтобы x не существовал?

Решение:

Расположим числа в порядке возрастания: $4, 5, 7, 9$. В зависимости от расположения x меняется медиана. Так, если мы воткнём x перед или сразу после 4, медианой будет 5. Если воткнуть x после 5, то сам x будет медианой. Если воткнуть x в конце или перед 9, то медианой окажется 7.

Составим три уравнения:

$$\frac{x+4+5+7+9}{5} = 5 \Rightarrow x = 0$$
$$\frac{4+5+x+7+9}{5} = x \Rightarrow x = 6.25$$
$$\frac{4+5+7+9+x}{5} = 7 \Rightarrow x = 10$$

Задача 3

Измерен рост 25 человек. Средний рост оказался равным 160 см. Медиана оказалась равной 155 см. Машин рост в 163 см был ошибочно внесен как 173 см. Как изменится медиана и среднее после исправления ошибки? А как могут измениться медиана и среднее, если рост Маши равен 153?

Решение:

Если рост Маши ошибочно был внесен как 173 см вместо 163, то при исправлении ошибки изменения никак не отразятся на медиане, потому что ошибочно внесенный рост и ее рост больше медианы. Средний рост уменьшится. В случае 153 изменения могут коснуться как среднего, так и медианы.

Задача 4

Деканат утверждает, что если студента N перевести из группы A в группу B,то средний рейтинг каждой группы возрастет. Возможно ли такое?

Решение:

Да, возможно. Если средняя оценка N ниже средней оценки группы A, но выше средней в группе B, то после смены студентом группы средняя оценка каждой группы и ее рейтинг возрастут.

Например, группы и состоят из 3 человек, которые имеют оценки 8, 9, 10 и 1, 2, 3 соответственно. Студент N, имеющий оценку 8, желает перейти в группу . Тогда изменения оценки группы A: $\frac{(9+10)}{2} - \frac{(9+10+8)}{3} = 0.5$, то есть рейтинг группы повысится. Изменения для группы B: $\frac{(1+2+3+8)}{4} - \frac{(1+2+3)}{3} = 1.5$, а значит рейтинг группы B тоже повысится.

Задача 5

Иногда в качестве меры разброса используют размах. Находят максимальное значение в выборке, минимальное значение выборке, а после вычитают из максимума минимум. Как думаете, такая мера чувствительна к выбросам? Предложите способ сделать её устойчивой к ним.

Решение:

Да, будет. Для того, чтобы сделать эту мену устойчивой к выбросам, можно считать интерквантильный размах, то есть вычитать из 75% квантиля 25% квантиль.