Семинар 4: основы статистики!

Обсудили до задачки:

- Что такое распределение, что такое гистограмма
- Меры центральности: среднее, медиана, мода
- Меры разброса: размах, дисперсия, среднее квадратичное отклонение
-

Задача 1

В табличке ниже дано целых 10 наблюдений.

имя	пол	возраст	вес	рост
	м	14		
	ж	16		
	ж	20		
	м	20		
	м	14		
	ж	25		
	м	30		
	ж	23		
	м	22		
	м	16		

Что нужно сделать :

а) Что такое непрерывная переменная? Что такое категориальная переменная? Какие переменные в табличке относятся к непрерывным? Какие к категориальным?

б) Постройте для пола гистограмму. Ченить спизданите про неё

б) Постройте для возраста гистограмму. Ченить спизданите про неё

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

е) 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. (уточнить что это дискретная мода и тп)

з) ченить в стиле как измерть среднее для категориальных переменных. Вопрос про доли, моду.

и) Что такое квантиль? Предложите способ, который помог бы избавиться от выбросов.

Решение:

а) Что такое непрерывная переменная? Что такое категориальная переменная? Какие переменные в табличке относятся к непрерывным? Какие к категориальным?

б) Посчитаем с какой частотой в выборке встречаются мужчины и с какой женщины. В выборке 6 мужчин и 4 женщины.

б) Постройте для возраста гистограмму. Ченить спизданите про неё

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 14 16 16 20 20 22 23 25 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.

Отметим оба значения на гистограмме.

Отметить

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

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 количество наблюдений, 10.

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 подобное логикой мы ещё встретимся, когда будем обсуждать различные метрики, используемые в машинном обучении.
- Часто при подсчёте дисперсии вместо формулы

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

  которую использовали мы, используют

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

  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.

- Как правило, большая часть выборки, а именно 69% кучкуется в диапазоне между $\bar{x} - \sigma$ и $\bar{x} + \sigma$.

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

  Так обычно происходит, если признак имеет нормальное распределение. 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.

  картинка с правилом 3-х сигм

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

рассчёты! картинка гистограммы с выбросами

Среди наших элементов оказался выброс. Он существенно сдвинул среднее значение 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.

.

е) Чувствительна ли дисперсия к выбросам?

К несчастью, да.

ж) 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)

Чтобы найти моду, мы должны найти самое часто встречаемое значение. Для непрерывной переменной это не очень хорошая затея.

про возраст

з) ченить в стиле как измерть среднее для категориальных переменных. Вопрос про доли, моду.

и) Что такое квантиль? Предложите способ, который помог бы избавиться от выбросов.

Объяснения про квантиль!

Как мы выяснили выше, выбросы могут существенным образом искажать наши представления о выборке. От них нужно выборку очищать. Один из способов: отрубить все наблю-

дения, которые находятся выше 99% квантиля и все наблюдения, которы находятся ниже 1% квантиля. На гистограмме такой подходит выглядит как-то вот так:

Борьбу с выбросами нарисовать

Задача 1

Ещё задачи!

Задача 1

Имеется пять чисел: x, 9, 5, 4, 7. При каком значении x медиана будет равна среднему?

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:

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

А можно ли поставить такие цифры в условии задачи, чтобы x не существовал?

Задача 2

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.

Задача 3

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

Задача 4

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? Как можно сделать её устойчивой к ним? (интерквантильный размах).

Семинар про АБ-тесты

Задача 1

Скарлет, Сьюлин и Кэррин исследуют рост людей в Атланте.

Аристарх, Маша и Паша исследуют рост людей.

Три исследователя, все исследуют рост людей. Сделали три выборки.

Илья занимается баскетболом, поэтому он переписал из журнала тренера рост членов команды

Маша измеряет рост людей на остановке, пока те ждут автобус.

Паша ещё чет делает не оч стабильное. Рост девушек пока они спят в его кровати - ???

У кого из исследователей получится репрезентативная выборка?

Задача 2

По всем 4 наблюдениям реально равно.

4 наблюдения всего в генеральной совокупности. Мы хотим проверить равно ли оно чему-то. Для этого делаем два наблюдения и смотрим равно ли.

мы делаем два наблюдения и считаем среднее. Если равно да, если не равно нет.

В реальности равно. Как нам получить правильные выводы, если все 4 наблюдения мы никогда не можем собрать.

Построить распределение среднего. Отсюда родить статистический критерий для среднего. Типо если не сильно отклоняется ок, сильно не ок. Плавно перейти к ЦПТ отсюда.

что-нибудь с монеткой

Задача 3

Проверить какую-нибудь гипотезу через правило трёх сигм и нормальное распределение про среднее.

Задача 4

Задача 5

Задача 6

Во время Второй Мировой войны американские военные собрали статистику попаданий пуль в фюзеляж самолёта. По самолётам, вернувшимся из полёта на базу, была составлена карта повреждений среднестатистического самолёта. С этими данными военные обратились к статистику Абрахаму Вальду с вопросом, в каких местах следует увеличить броню самолёта. Что посоветовал Абрахам Вальд и почему?

Профессор, я решал эту задачу 3 часа, а ответ не совпадает. Где я сделал ошибку? — Предположив, что ответы к задачнику верные.

Задачка 0 (в которой мы разгоняем сомнения)

- Как Сергею понять где кофе любят больше?
- Как Акулине правильно провести АБ-тест своего снадобья?

Задачка 0 (в которой сомнение селится в наших головах)

- Хипстер Сергей пристаёт на улице к людям со странными вопросами. Он опросил в Питере и Москве по сто человек. Каждому он задавал вопрос: ”Кофе любишь?”В Москве ”Да”сказали 50 человек, в Питере 55 человек. Можно ли исходя из этого сделать вывод, что в Питере кофе любят больше?
- Знахарка Акулина смешала в тазике ”доктор Мом”с соком редьки. Этот настой она дала простудившейся внучке. Внучка выздоровела. Означает ли это, что лекарство работает? Предположим, что знахарка дала своё ”лекарство”пяти простуженным внучкам. Три из них выздоровели, две нет. Означает ли это, что лекарство в большей части случаев помогает?