Семинар 3: описательные статистики

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

Паблик с аниками категории Б

Упражнение 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	15	40
Мелисандра	ж	21	40
Эддард	м	20	80
Сандор	м	14	80
Миссандея	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 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?

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.

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

Ещё задачи!

Тут находится несколько задачек, о которых вам нужно подумать самостоятельно, в домашних условиях, за чашкой чая. Одна из этих задачек точно попадётся вам на самостоятельной работе. Вторая задачка на ней будет совсем новой. В ней надо будет посчитать что-то похожее на то, что было в первой части pdf-ки.

Упражнение 2

Ваня любит пить чай. Иногда он пьёт его с сахаром, иногда без. На этой неделе он помечал цифрой 1 дни, когда пил чай с сахаром. Получилось $1, 1, 0, 0, 1, 0$.

а) Найдите среднее значение сахарных дней в жизни Вани. Найдите дисперсию сахарных дней.

б) Правда ли, что среднее число сахарных дней совпало с долей сахарных дней? Почему так вышло? Всегда ли так будет происходить?

в) Между дисперсией и долей в случаях, когда переменная принимает значения 0 или 1 тоже есть связь. Сможете догадаться как будет выглядеть формула, описывающая эту связь?

Упражнение 3

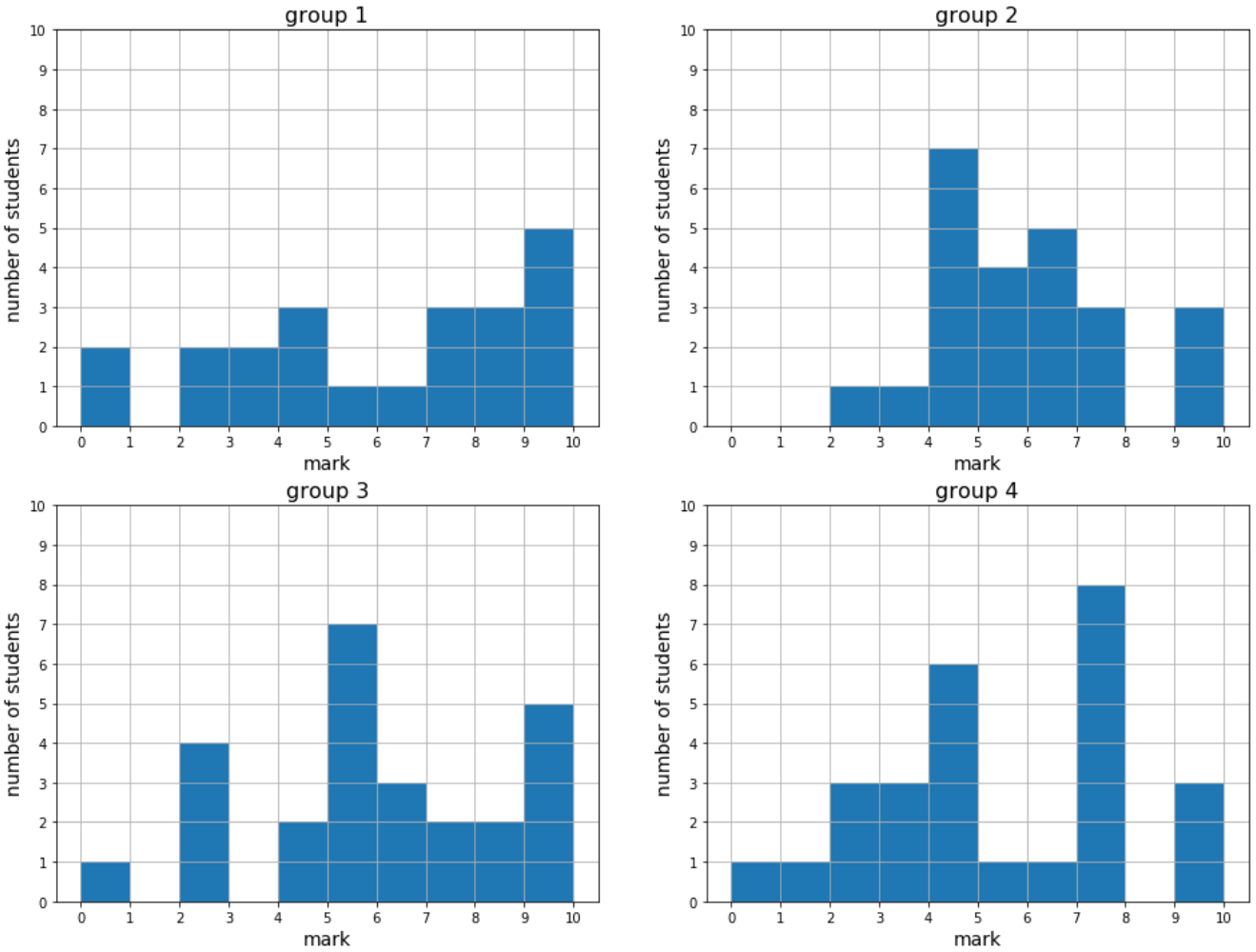
Имеется пять чисел: x, 9, 5, 4, 7. При каком значении x медиана будет равна среднему? А можно ли поставить такие цифры в условии задачи, чтобы x не существовал?

Упражнение 4

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?

Упражнение 5

Четыре группы на «соседнем» факультете написали контрольную работу по программированию, у каждой группы был свой проверяющий. Ниже приведены гистограммы оценок за эту работу с шагом один (в столбец включается значение *правой* границы). Посчитайте для каждой группы медиану и моду, сравните их между собой.



Упражнение 6

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

Упражнение 7

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.

Упражнение 8

Гостомысл прочитал, что средний класс вымирает. Американское исследовательское агентство с говорящим названием Пью Рисёч Центр объявило, что средний класс вымирает чуть менее, чем повсеместно. Большинство взрослых американцев более не являются средним классом. Десятилетия растущего расслоения и перевод производств в Азию сделали своё дело[1].

В голове Гостомысла возник вполне закономерный вопрос: а средний — это какой? Почитав чуть побольше он вычитал, что в США это от $^2/_3$ до 2 медианных зарплат. Для России чёткого определения Гостомысл так и не нашёл. Поэтому он решил придумать его сам.

[1] В книге Жлобология от Алексея Маркова вычитал

1. Объясните как может так получиться, что средний класс умирает. Ведь когда у людей меняются зарплаты, медиана тоже должна двигаться.
2. Гостомысл подумал, что если уж класс средний, то и определять его надо через средние. Поэтому именем себя самого он нарёк средним классом людей с доходом от $^{2}/_{3}$ до 2 средних зарплат. Как думаете, какой средний класс окажется более многочисленным? Измеренный с помощью средних или медиан? Почему в США для измерения используют именно медианы?