

Основы анализа и визуализации данных для медиа

Семинар 3
19 сентября 2019

Рогович Татьяна
rogovich@gmail.com

Меры разброса. Работа в Tableau

Структура семинара

- Поговорим про меры разброса и какую дополнительную информацию они нам сообщают
- Обсудим, что такое выбросы и как их находить
- Попробуем уже что-то наконец-то сделать в Tableau (ура!)

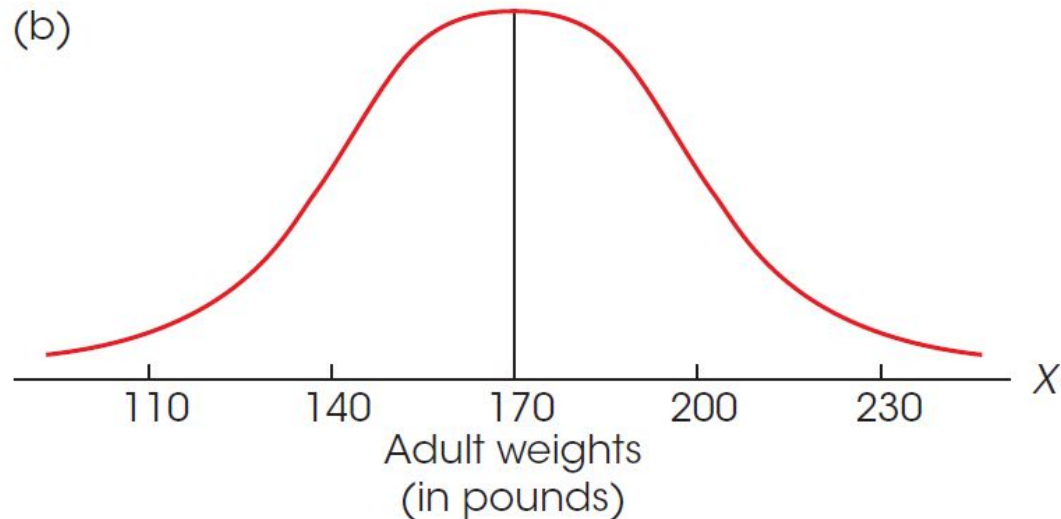
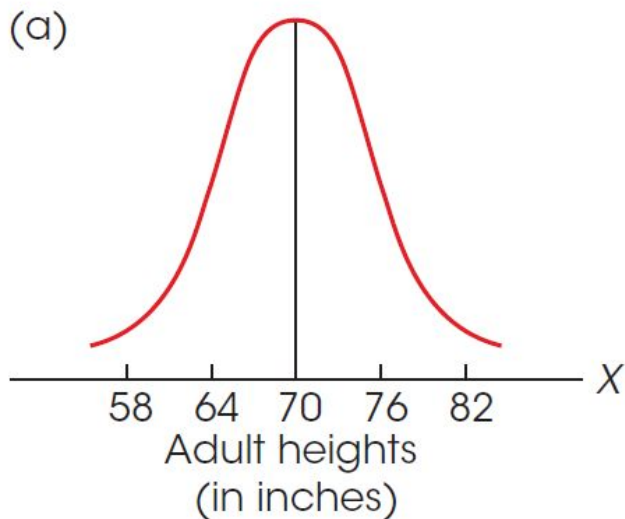
Меры разброса

Variability

Говорят о том, насколько сильно рассеяны значения в данных.

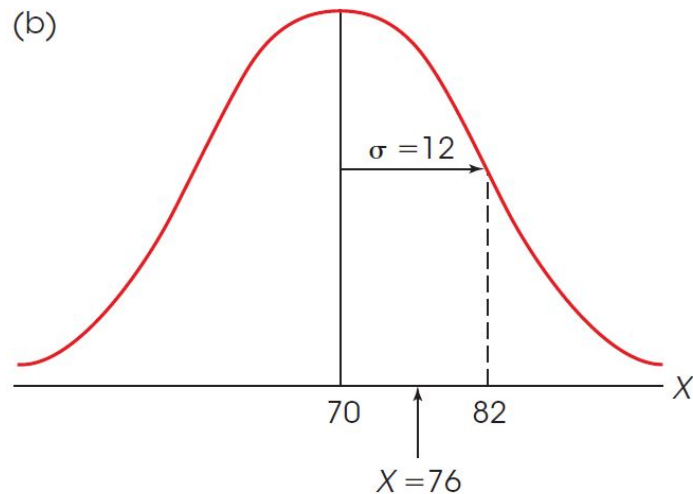
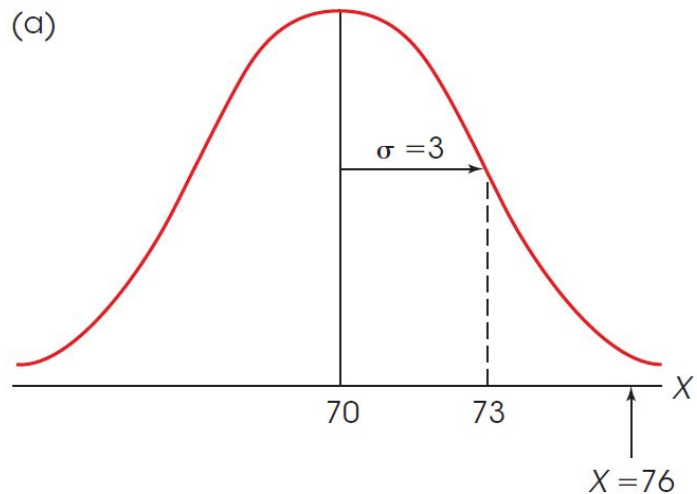
Размах, дисперсия, стандартное отклонение.

Зачем нужны меры разброса?



Чтобы описывать распределения более полно. Так мера разброса говорит нам, например, насколько плотно значения сконцентрированы вокруг среднего.

Зачем нужны меры разброса?



Так, например, один и тот же результат теста может быть как очень хорошим, так и “средненьким” в распределениях с разной плотностью.

Зачем нужны меры разброса

						Среднее
33	31	32	36	31	31	32,3
22	34	58	52	10	21	32,8

Представьте, что вы хотите сделать некий вывод про людей. И в том и в другом случае вы скажите, что ваши выводы корректны для людей 32 лет.

Но разве это корректно?

Размах Range

Разность между самым большим и самым маленьким значениями в выборке.

- Плохо подходит, когда в данных есть выбросы.
=> необычно большой размах или крайне экстремальные минимальное или максимальное значения могут быть поводом для дальнейшего исследования.

$$\text{range} = X_{\max} - X_{\min}$$

Интер- квартильный размах Interquartile range

Разность 75-й и 25-й перцентилей (медианы половинок данных до медианы и после медианы (которая сама по себе 50-я перцентиль)).

Применяется в паре с медианой!

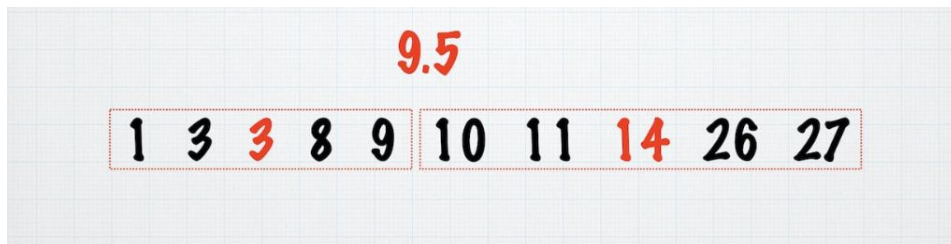
<https://medium.com/@congyuzhou/%D0%B8%D0%BD%D1%82%D0%B5%D1%80%D0%BA%D0%B2%D0%B0%D1%80%D1%82%D0%B8%D0%BB%D1%8C%D0%BD%D1%8B%D0%B9-%D1%80%D0%B0%D0%B7%D0%BC%D0%B0%D1%85-28604de365a0>

Интерквартильный размах

Interquartile range



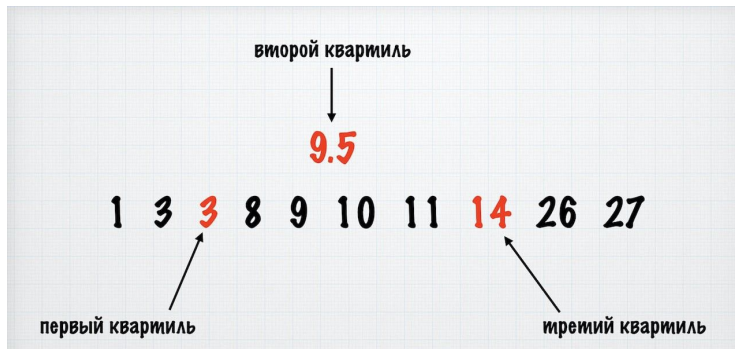
- Находим медиану (срединное значение выборки, она же 50-ю перцентиль (справа и слева от нее по 50% выборки))



- Теперь, чтобы найти 25-ю и 75-ю перцентиль - находим “медианы” левой и правой половин выборок (слева и справа от медианы).

Интерквартильный размах

Interquartile range



- Медиана всей выборки — это второй квартиль, медианы левой и правой половин — это, соответственно первый (или нижний) и третий (или верхний) квартили. Они же 25-я и 75-я перцентили.



- Разница между 75-й и 25-й перцентилем и будет интерквартильным размахом.

Интерквартильный размах

Interquartile range



- В данном случае, интерквартильный размах равен

$$14 - 3 = 11$$

- Это значит, что 50% наших данных с медианой 9.5 находятся в интервале 11 значений (5.5 влево от медианы и 5.5 вправо от медианы). Эти данные достаточно рассеяны.

Интерквартильный размах

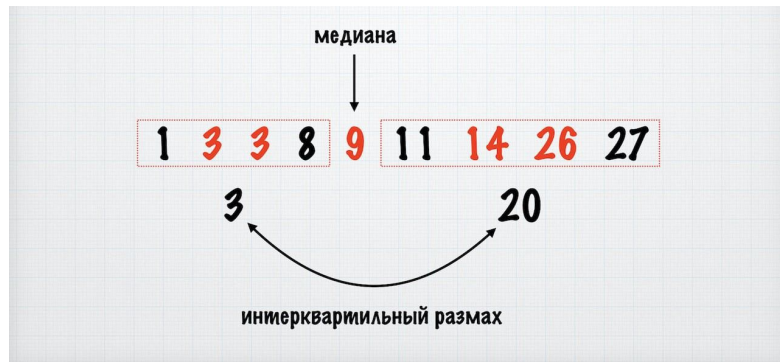
Interquartile range



- Представьте, что мы опросили 10 человек, о том, сколько они тратят в месяц на ребенка и спросили о возрасте их детей. Например, у нас получилось, что они тратят, в среднем, 30 000 рублей в месяц. Медиана нашей выборки 9.5 и мы бы хотели сделать вывод, что “опрошенные в среднем тратят 30 000 рублей в месяц на детей 9.5 лет”.
- Но если мы учтем интерквартильный размах (11, 50% детей наших опрошенных находятся в интервале от 3 до 14 лет), то увидим, что делать вывод, что наша выборка отвечает на вопрос про детей 9.5 лет - неверно. Корректно указать и медиану, и интерквартильный размах.

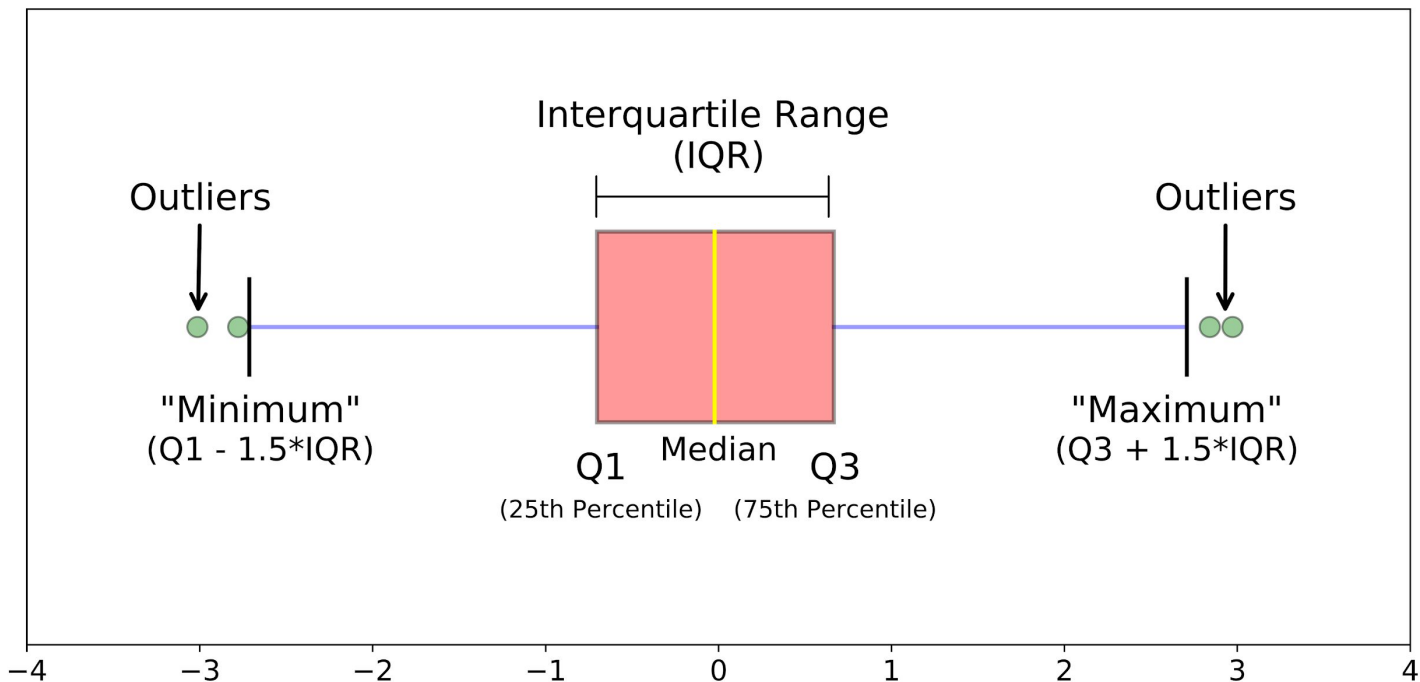
Интерквартильный размах

Interquartile range



- Если в выборке нечетное количество элементов, то половинки выборки будут содержать четное количество элементов. Действуем так же как в случае с медианой (находим серединки между двумя элементами, складываем их и делим на два).

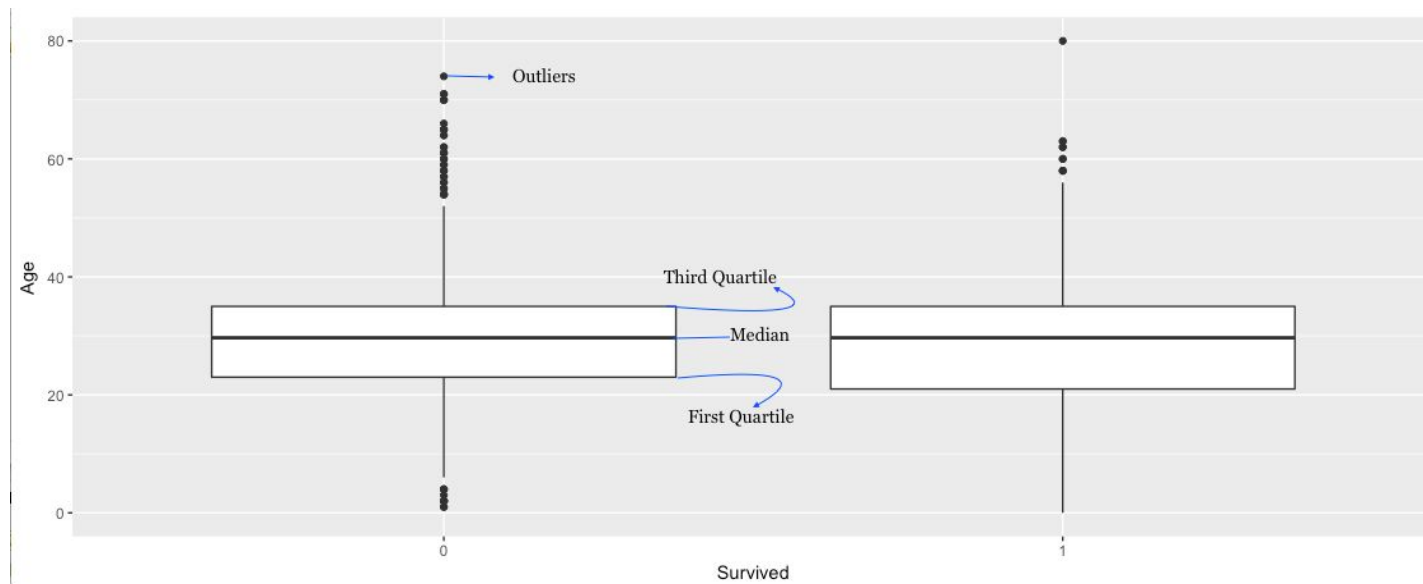
Межквартильный размах на графиках



Box
and whiskers
Plot

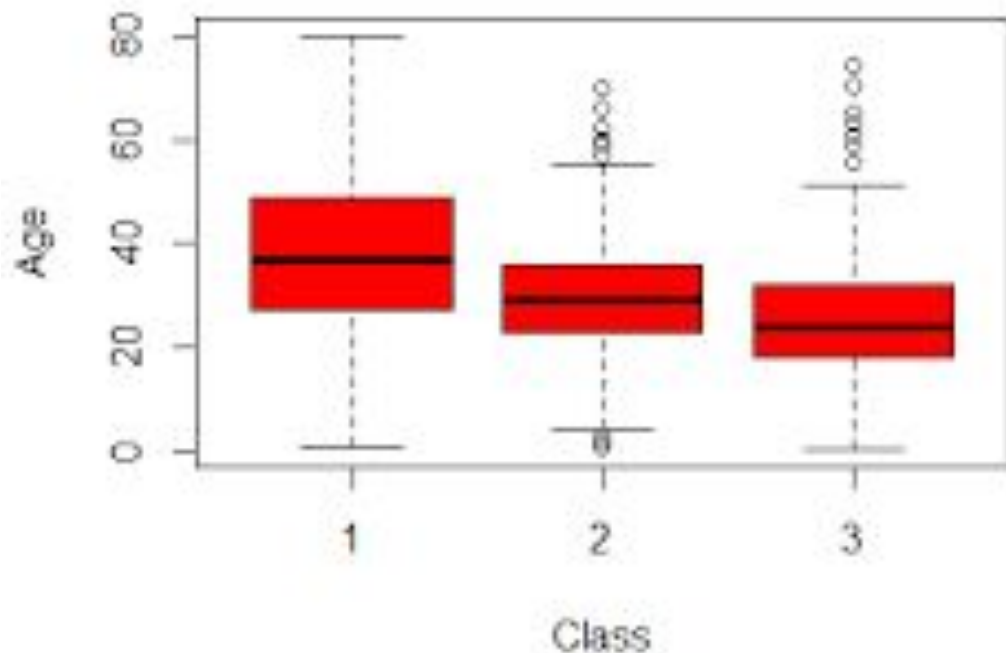
Ящик-с-усами

Интерквартильный размах на графиках



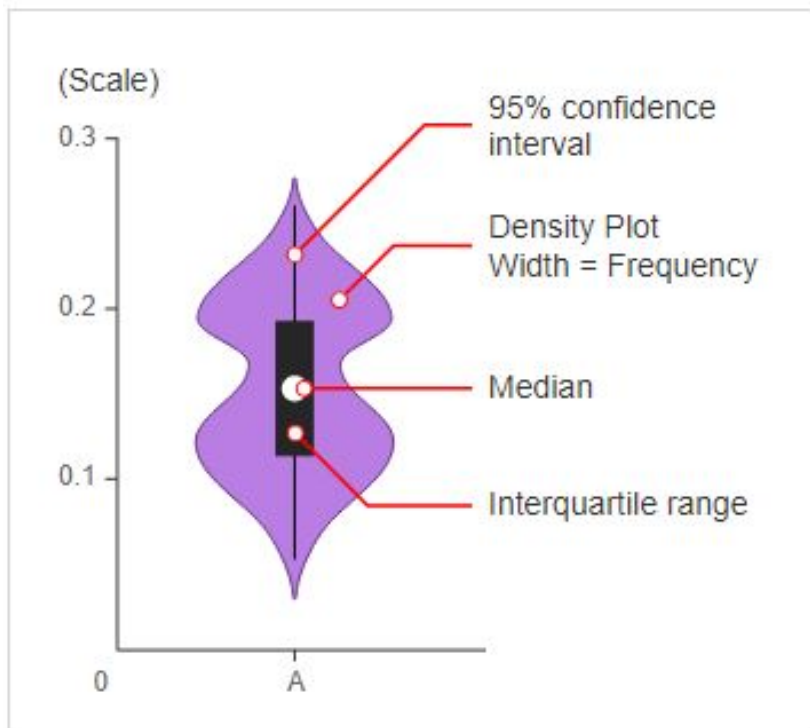
Он же,
но вертикальный

Интерквартильный размах на графиках



Ящик-с-усами,
показывающий
распределение
возрастов
пассажиров
“Титаника” по
классам билетов

Межквартильный размах на графиках



Violin plot

Экзотика,
но знайте, что он
существует

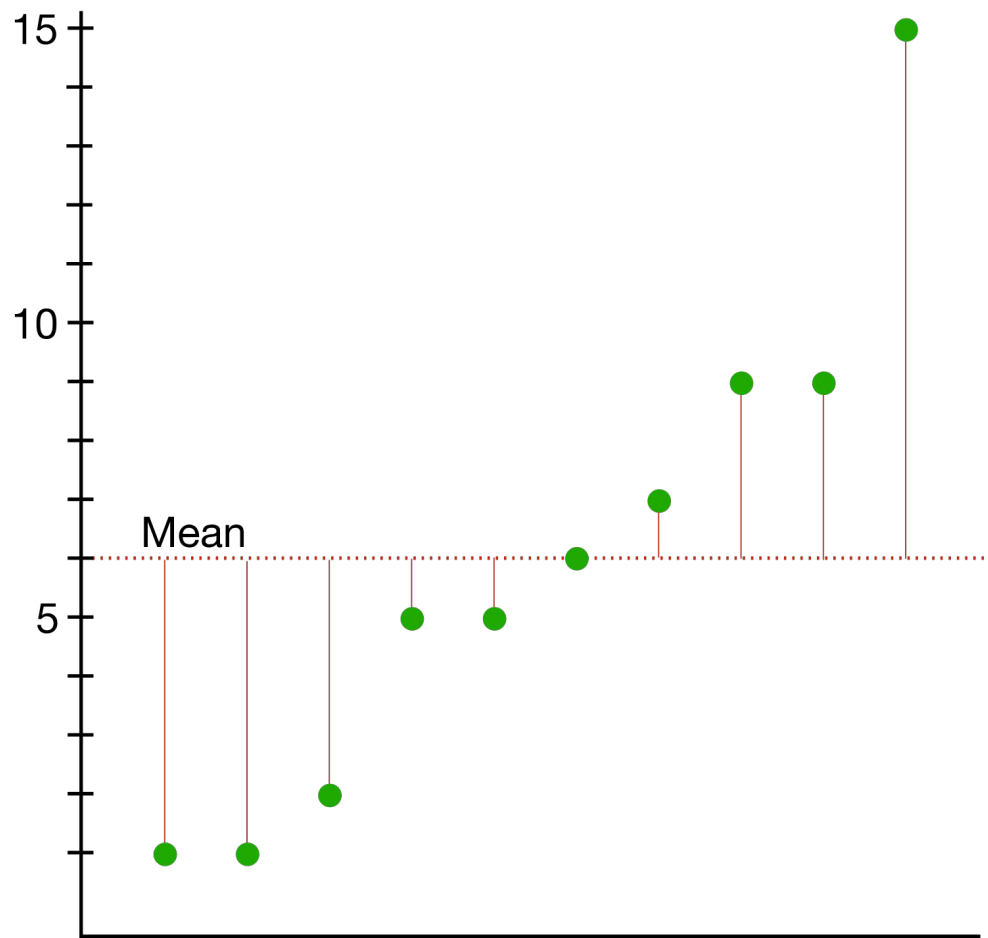
Дисперсия и стандартное отклонение Variance and standard deviation

Описывают насколько отдельные значения в данных в среднем отличаются от среднего значения.

Дисперсия – это средний квадрат отклонения от среднего. Стандартное отклонение – это квадратный корень из дисперсии.

Дисперсию генеральной совокупности обозначают как σ^2 (произносится как «сигма в квадрате»), а стандартное отклонение – σ (греческая буква «сигма»), тогда как в случае выборки дисперсия и стандартное отклонение обозначаются как s^2 и s соответственно

Как выглядит
отклонение от
среднего?



Зачем возводим в квадрат?

$$\sum_{i=1}^n (x_i - \mu)$$

Отклонение от среднего. Всегда будет равно 0.
Чтобы обойти эту проблему, мы работаем с квадратами отклонений, которые по определению всегда положительны.

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

Формула дисперсии генеральной совокупности.

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

Формула дисперсии выборки.
(обратите внимание на n-1)

Как получаем стандартное отклонение?

=> Чтобы вернуться к изначальным единицам, мы извлекаем квадратный корень из дисперсии.

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2} \qquad s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

Альтернативная формула (те же результаты)

Для генеральной совокупности

$$SS = \sum X^2 - \frac{(\sum X)^2}{N}$$

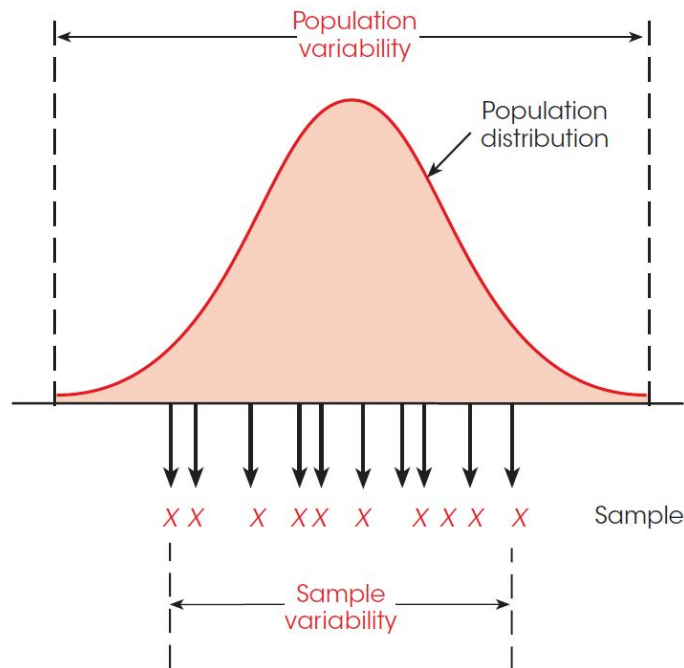
$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{SS}{N}}$$

Для выборки

$$SS = \sum X^2 - \frac{(\sum X)^2}{n}$$

$$s = \sqrt{s^2} = \sqrt{\frac{SS}{n - 1}}$$

Для тех, кому интересно, почему делим на $n-1$ для выборки



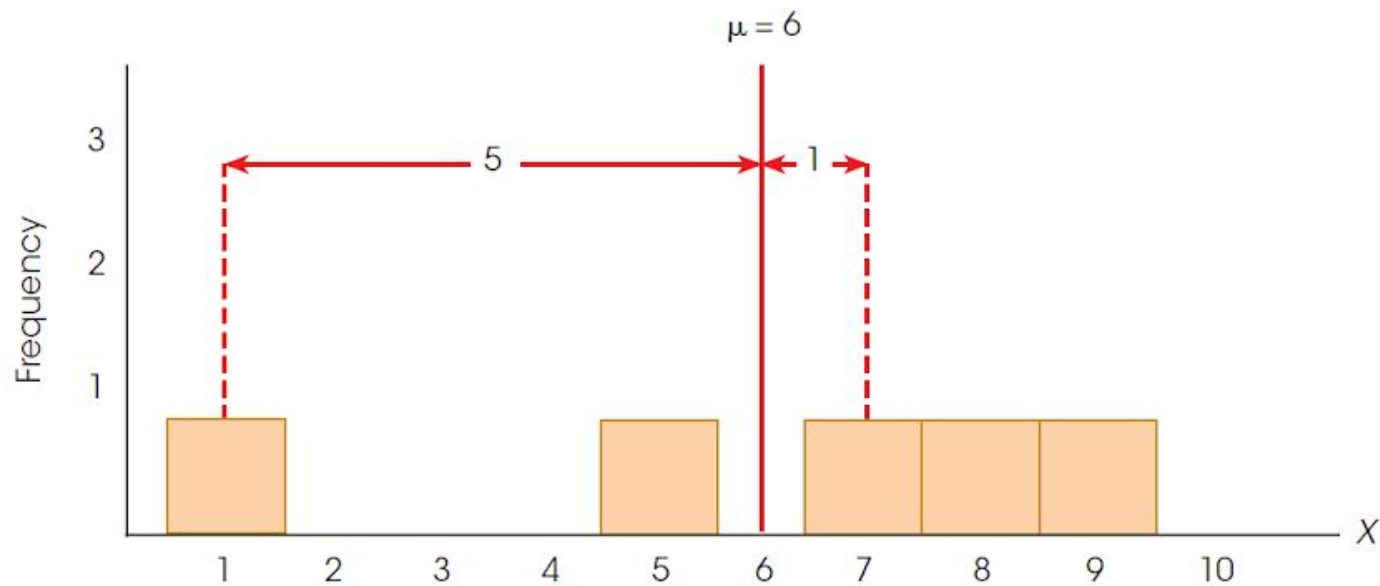
Цель дескриптивных статистик - сказать что-то о генеральной совокупности. Но мы знаем, что наши наблюдения скорее всего из самой популярной части генеральной совокупности и в них меньше не стандартных значений (меньше вариативности). И через $n-1$ мы пытаемся сделать небольшую корректировку этого факта.

Упражнение

Посчитать дисперсию и стандартное отклонение для следующего ряда чисел:

1, 9, 5, 8, 7

Упражнение

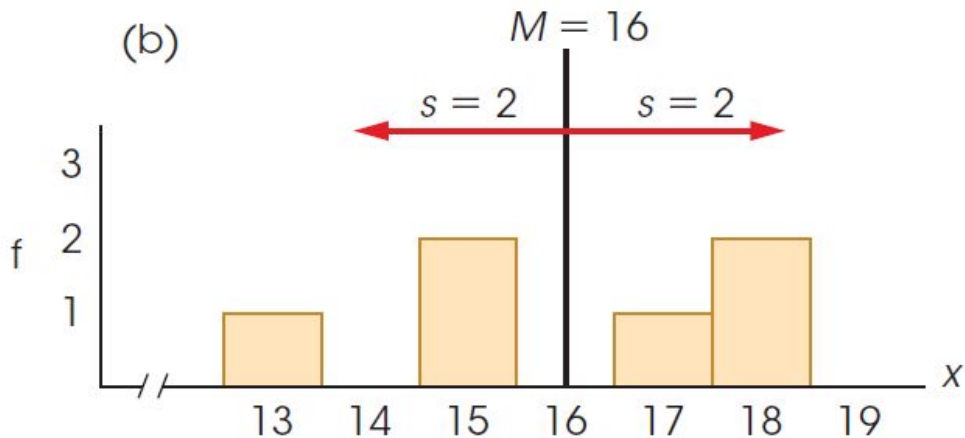
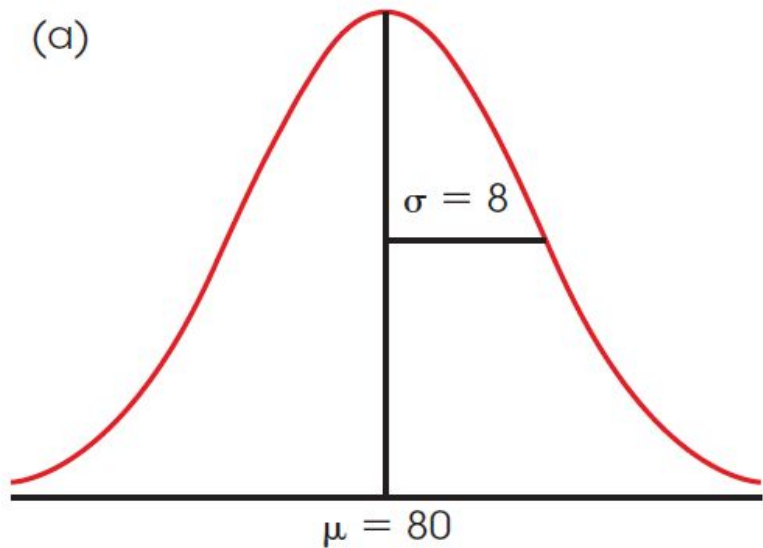


Упражнение

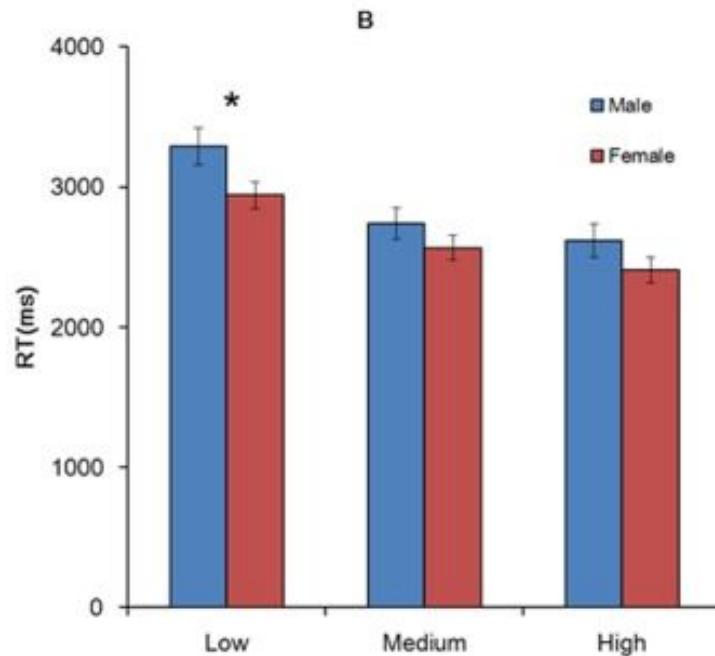
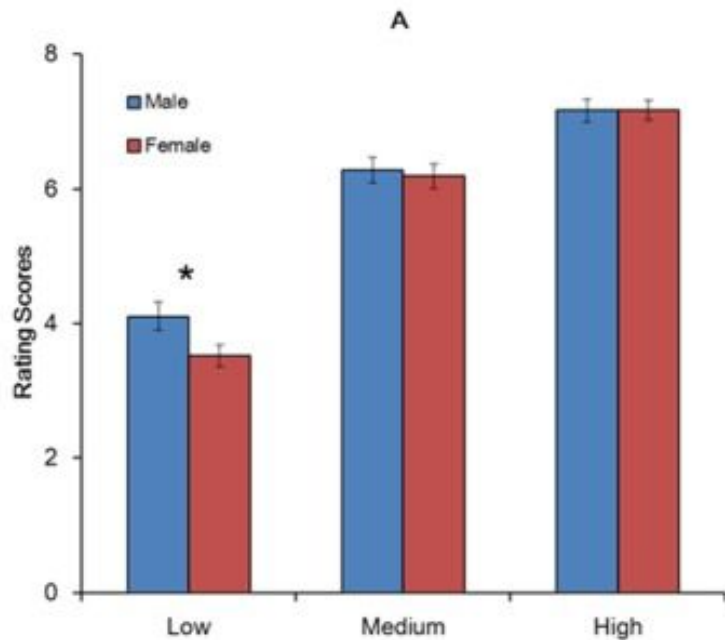
						Среднее
33	31	32	36	31	31	32,3
22	34	58	52	10	21	32,8

Посчитайте стандартное отклонение и скорректируйте суждение про людей возраста 32 лет.

Стандартное отклонение на графиках



Стандартное отклонение на графиках



Меры разброса в научных текстах

Children who viewed the violent cartoon displayed more aggressive responses ($M = 12.45$, $SD = 3.7$) than those who viewed the control cartoon ($M = 4.22$, $SD = 1.04$).

Type of Video Game		
	Violent	Control
Males	$M = 7.72$	$M = 4.34$
	$SD = 2.43$	$SD = 2.16$
Females	$M = 2.47$	$M = 1.61$
	$SD = 0.92$	$SD = 0.68$

Выбросы Outliers

это наблюдение в анализируемых данных, значение которого сильно отличается от других.

Как опознать выброс? Эвристики.
Например, расстояние три стандартных отклонения от среднего.

С точки зрения исследователя:
принять решение, как их
“нейтрализовать”.

С точки зрения журналиста:
что-то, что нужно расследовать!

Сила выбросов

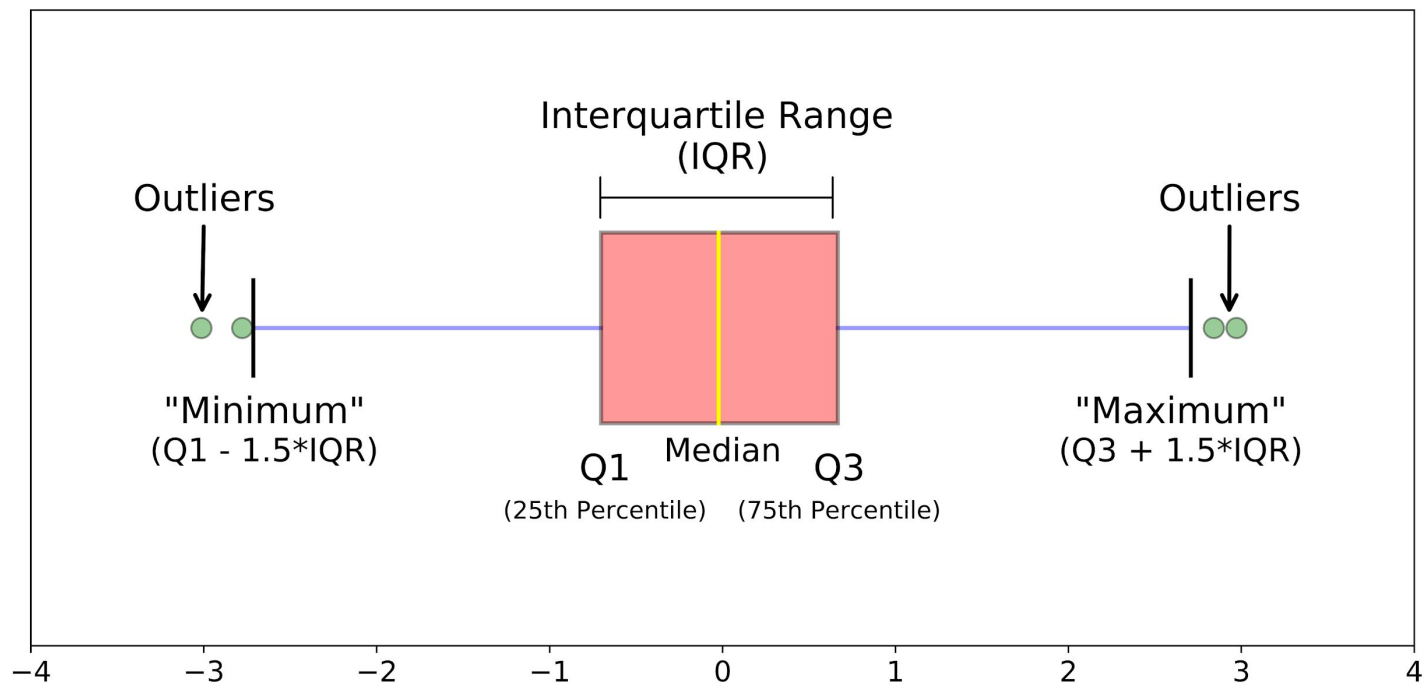
Слабые выбросы

- Те значения, которые меньше 25% перцентиля минус $1,5 \cdot \text{ИКР}$ (интерквартильный размах) или больше 75% персентиля плюс $1,5 \cdot \text{ИКР}$.
- Или находятся от среднего на расстоянии от двух до трех стандартных отклонений.

Сильные выбросы

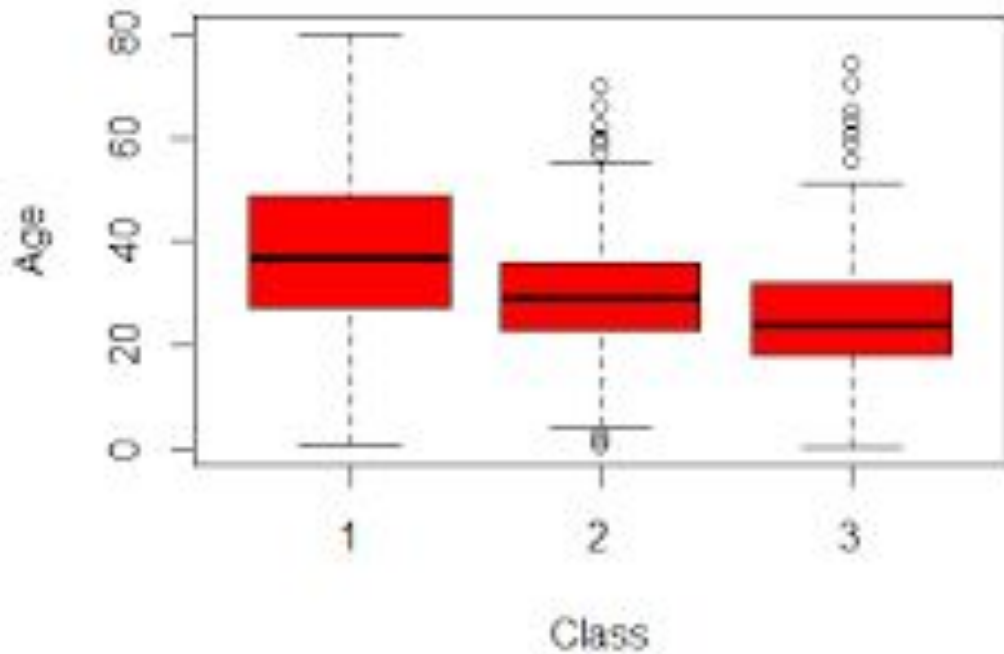
- Те значения, которые меньше 25% перцентиля минус $3 \cdot \text{ИКР}$ (межквартильный размах) или больше 75% персентиля плюс $3 \cdot \text{ИКР}$.
- Или дальше от среднего чем три стандартных отклонения.

Выбросы



В ящике-с-усами, усы как раз отграничивают выборку от выбросов.

Выбросы



Найдите выбросы
на графике.