

Основы анализа и визуализации данных для медиа

Семинар 2
12 сентября 2019

Рогович Татьяна
rogovich@gmail.com

Меры центральной
тенденции.
Работа с табличными
данными

Структура семинара

- Поговорим, что такое дескриптивные статистики и какую информацию мы узнаем благодаря им
- Узнаем, какие бывают меры центральной тенденции и в каких случаях их применять
- Покрутим данные в Excel и проверим гипотезы “на коленке”

Меры центральной тенденции

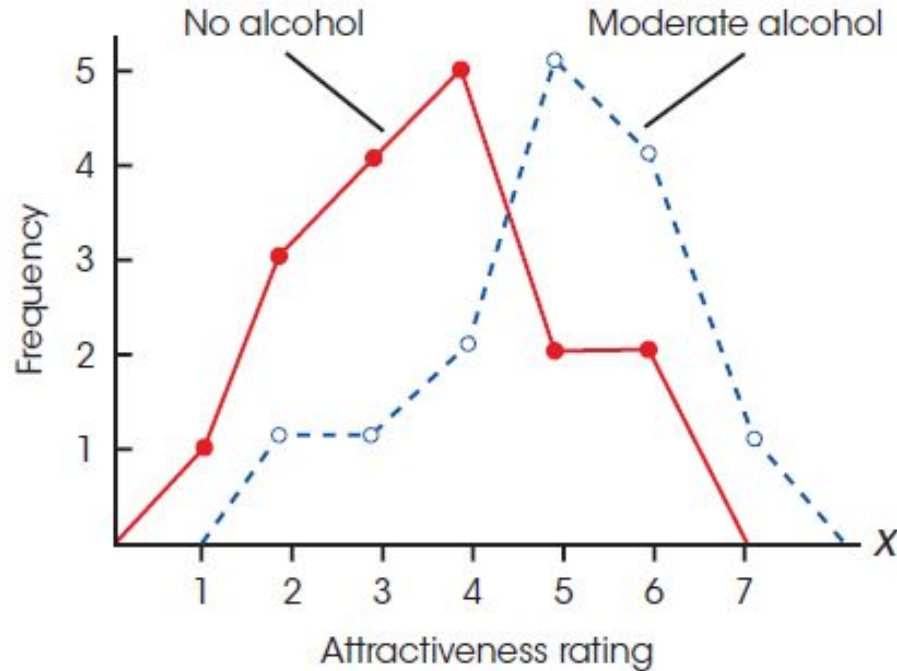
Central Tendency

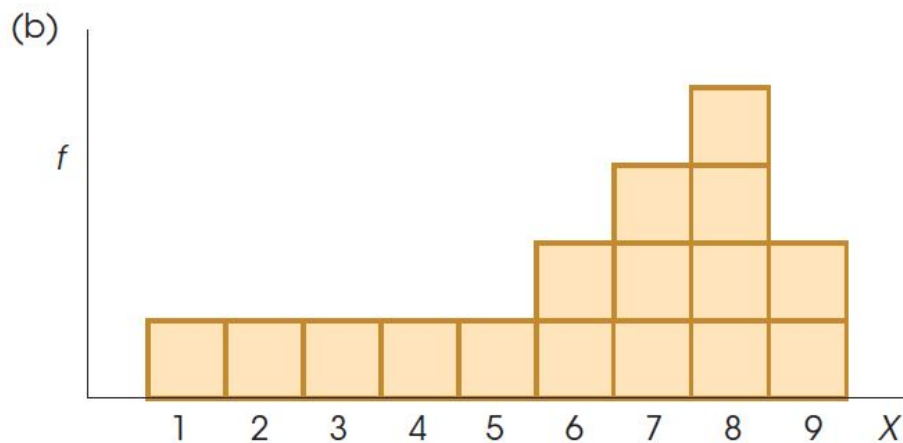
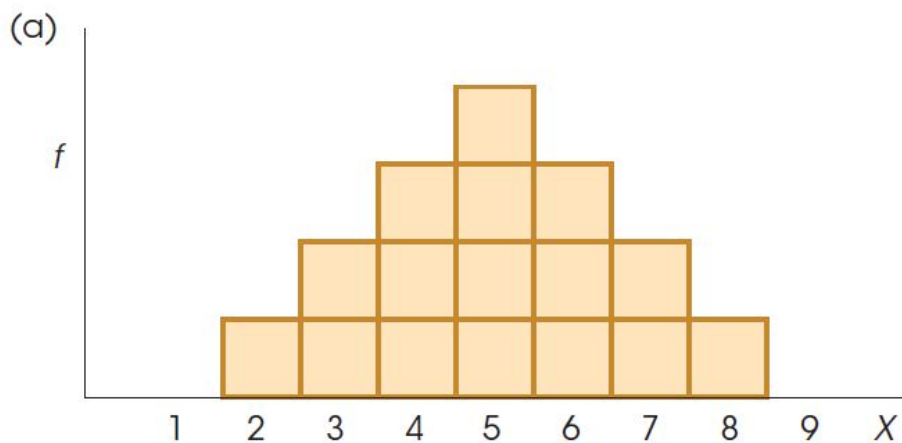
Цель их расчета - дать представление о типичном или часто встречающемся значении в данной переменной.

Среднее, медиана, мода

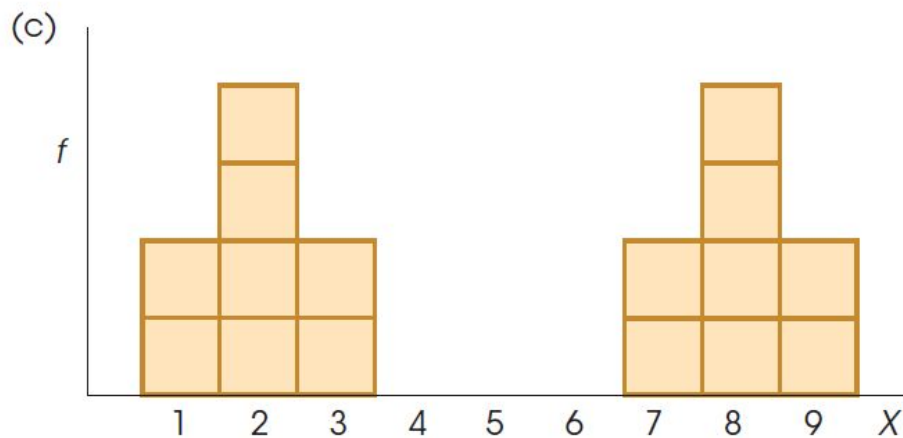
Меры центральной тенденции для проверки гипотезы

Research has now confirmed what you already suspected to be true—alcohol consumption increases the attractiveness of opposite-sex individuals (Jones, Jones, Thomas, and Piper, 2003). In the study, college-age participants were recruited from bars and restaurants near campus and asked to participate in a “market research” study. During the introductory conversation, they were asked to report their alcohol consumption for the day and were told that moderate consumption would not prevent them from taking part in the study. Participants were then shown a series of photographs of male and female faces and asked to rate the attractiveness of each face on a 1–7 scale. Figure 3.1 shows the general pattern of results obtained in the study. The two polygons in the figure show the distributions of attractiveness ratings for one male photograph obtained from two groups of females: those who had no alcohol and those with moderate alcohol consumption. Note that the attractiveness ratings from the alcohol group are noticeably higher than the ratings from the no-alcohol group. Incidentally, the same pattern of results was obtained for the male’s ratings of female photographs.





Какую
меру
выбрать?



Среднее Mean

Среднее арифметическое.

Подходит для переменных:

- Количественные непрерывные
- Бинарные 1/0(узнаем пропорцию единиц).

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\mu = \frac{\Sigma X}{N}$$

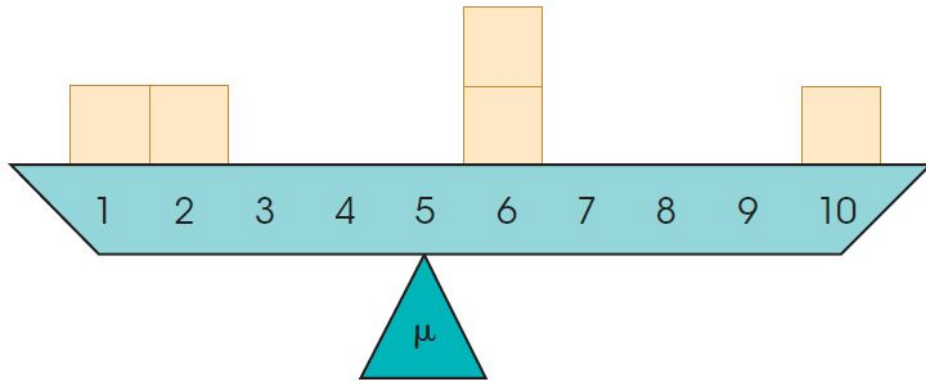
Плюсы, минусы, подводные камни

Среднее

Не всегда хороший показатель

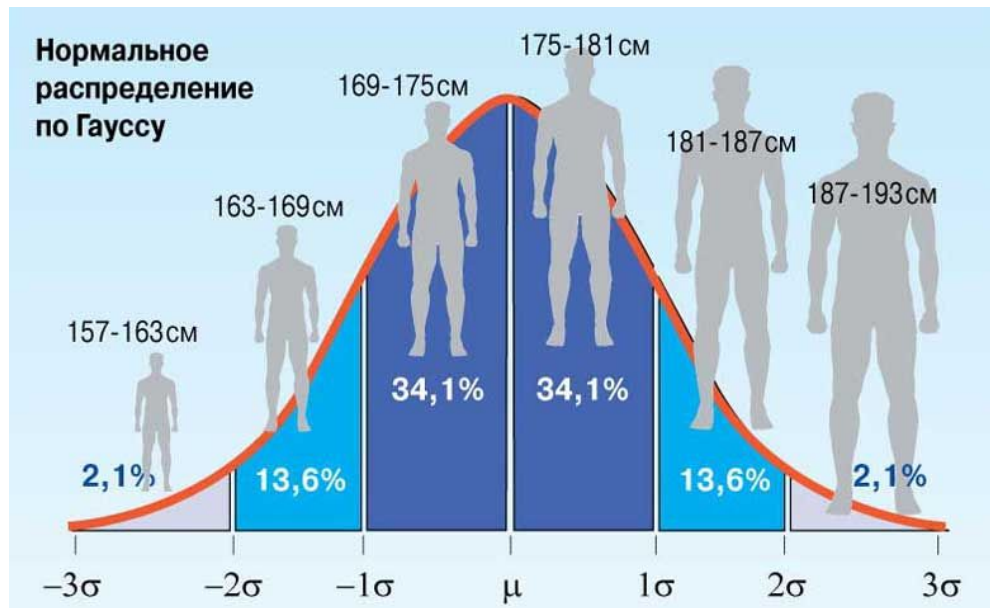
- Чувствителен к выбросам (например, несколько человек с очень большим уровнем дохода могут завысить средний уровень).
 - Можно использовать *винсоризованное среднее* (отсекаем определенный процент наблюдений с обоих концов распределения)

Пример использования



Score	Distance from the Mean
$X = 1$	4 points below the mean
$X = 2$	3 points below the mean
$X = 6$	1 point above the mean
$X = 6$	1 point above the mean
$X = 10$	5 points above the mean

Когда использовать среднее?



Когда нет выбросов
и распределение близко
к нормальному

Медиана

Median

Это срединное значение, если данные отсортировать по возрастанию или убыванию.

Если есть n значений, то медиана формально определяется как значение с порядковым номером $(n + 1)/2$.

- Не чувствительна к выбросам!

Подходит для переменных:

- Порядковые и дискретные

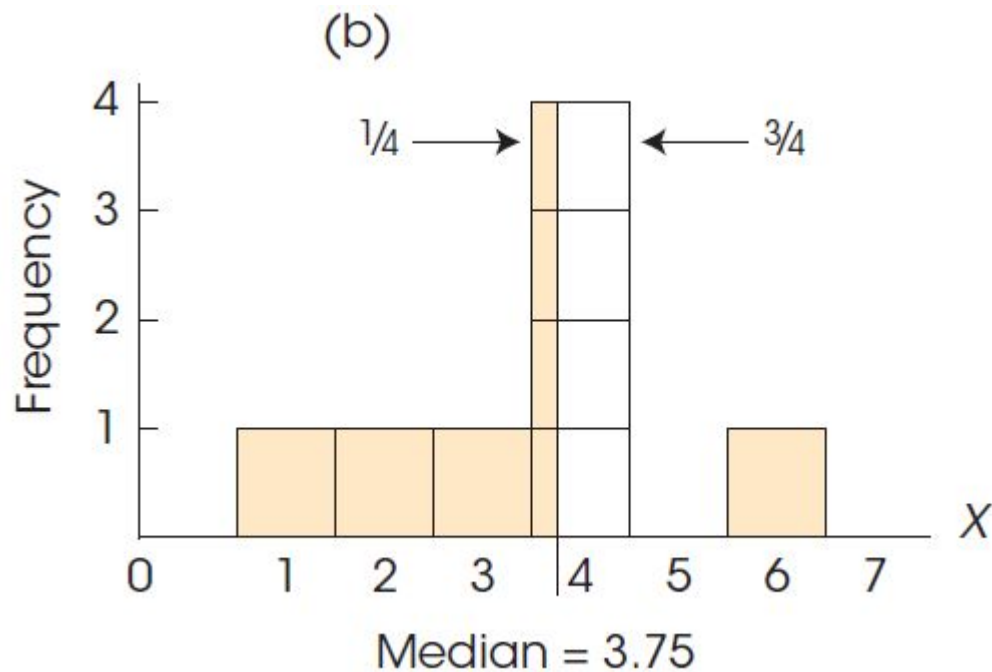
Плюсы, минусы, подводные камни

Медиана

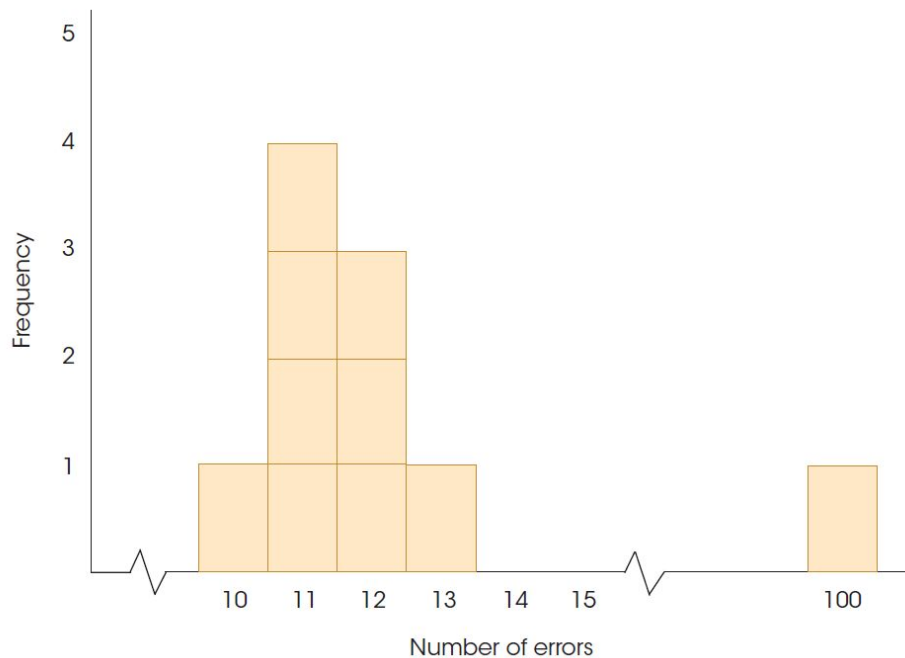
Медиана лучше среднего в качестве меры центральной тенденции для симметричных данных или данных с выбросами.

- Это связано с тем, что медиана основана на рангах, а не на самих значениях, и по определению половина значений лежит ниже медианы, а половина – выше, вне зависимости от конкретных чисел.

Пример использования



Когда использовать медиану?



Когда есть выбросы
или распределение
смещено

Когда использовать медиану?

Person	Time (Min.)
1	8
2	11
3	12
4	13
5	17
6	Never finished

Number of Pizzas (X)	<i>f</i>
5 or more	3
4	2
3	2
2	3
1	6
0	4

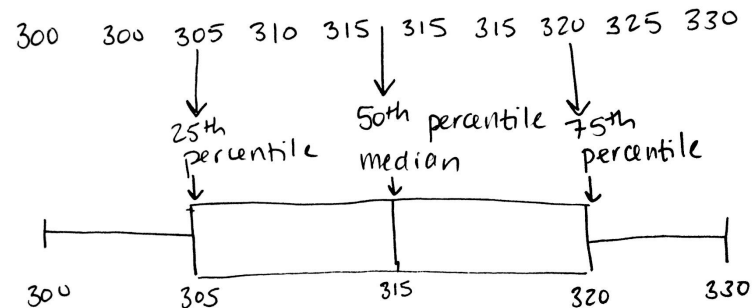
- Когда есть неопределенные значения или открытые распределения

Перцентиль Percentile

Значение, выделяющее одну из 100 частей распределения, выстроенных в ряд по их величине.

Так медиана это 50-я перцентиль.
Отделяет 50% наблюдений.

10 Data Points: GRE Scores



Упражнение

В каком случае лучше использовать медиану?

распределение А:

1, 1, 3, 4, 5, 6, 7;

распределение Б:

0.01, 3, 3, 4, 5, 5, 5;

распределение В:

1, 1, 2, 4, 5, 100, 2000.

Мода Mode

Самое частое значение в данных.

Подходит для переменных:

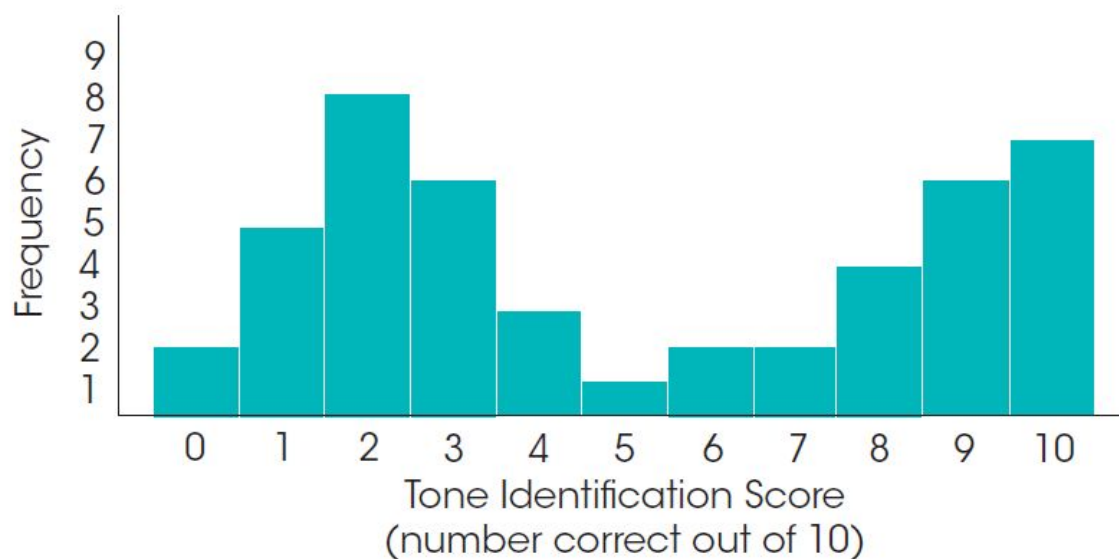
- Количественные
- Категориальные

Когда использовать моду?

Restaurant	f
College Grill	5
George & Harry's	16
Luigi's	42
Oasis Diner	18
Roxbury Inn	7
Sutter's Mill	12

- Когда используем номинальную шкалу
- Когда используем дискретные переменные
- Когда описываем форму распределения (не требует сложных вычислений, простая статистика)

Когда использовать моду?

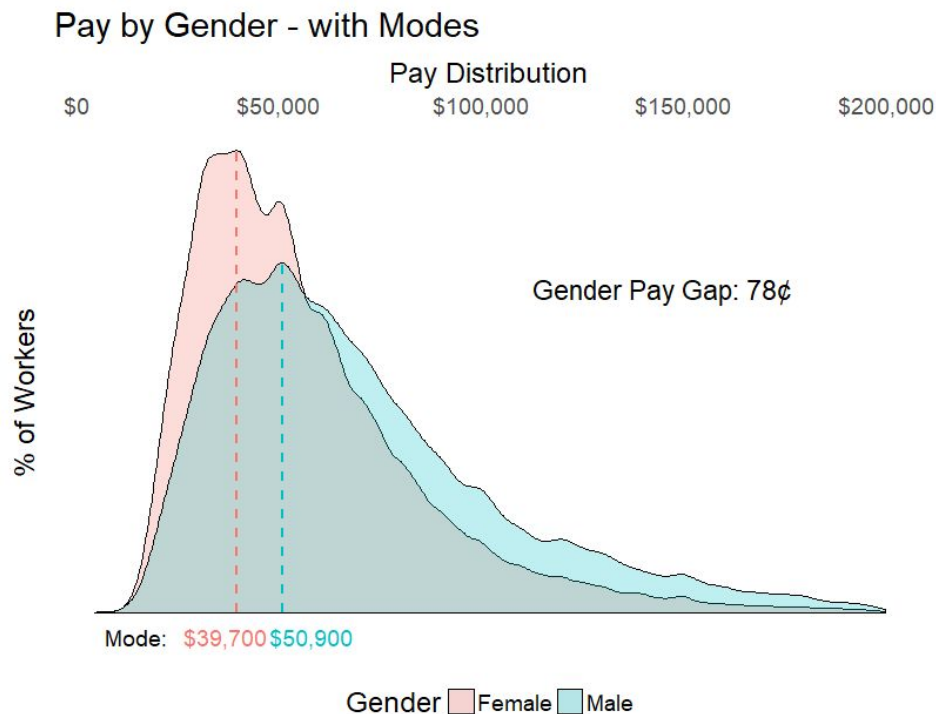


- Для би- и мультимодальных распределений (самые частые значения). Возможно, за этим скрывается дополнительная информация.

Когда какую меру использовать

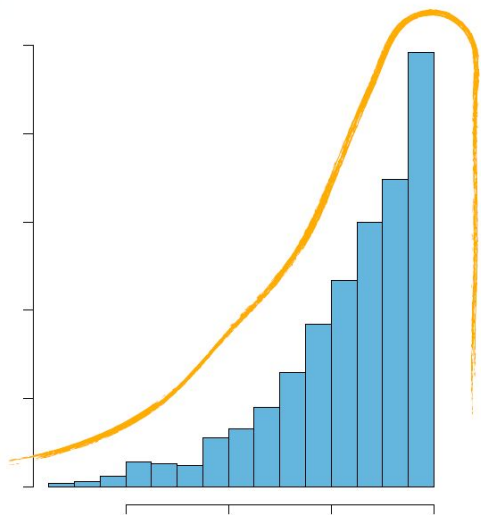
	Среднее (Mean)	Медиана (Median)	Мода (Mode)
Данные категориальные	Нет	Нет	Да
Есть выбросы (аномально большие или маленькие значения)	Нет	Да	Да
Распределение симметричное	Да (осторожно!)	Да	Да
Распределение скошено	Нет	Да	Нет

Что зависит от того, какую меру выбрать?

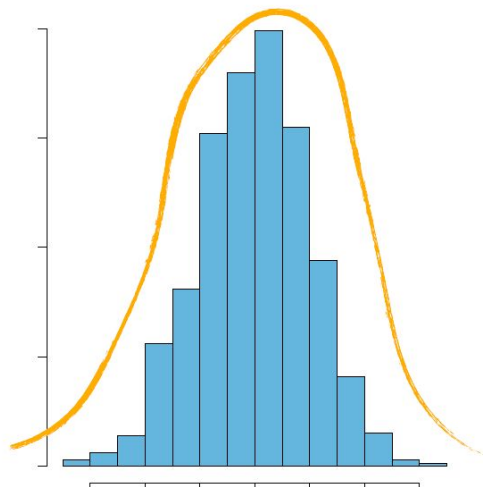


<https://www.payscale.com/data/average-mean-median-mode>

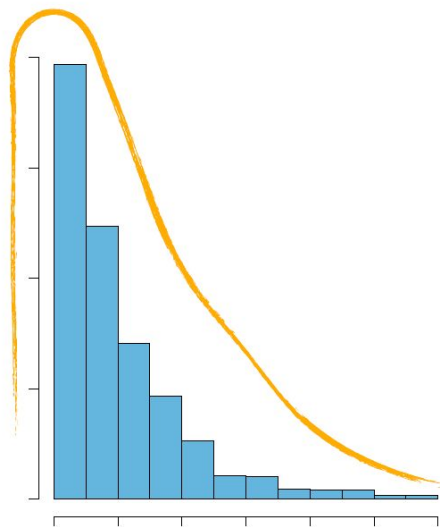
Центральная тенденция и форма распределения



Данные с левым плечом -
медиана больше среднего.

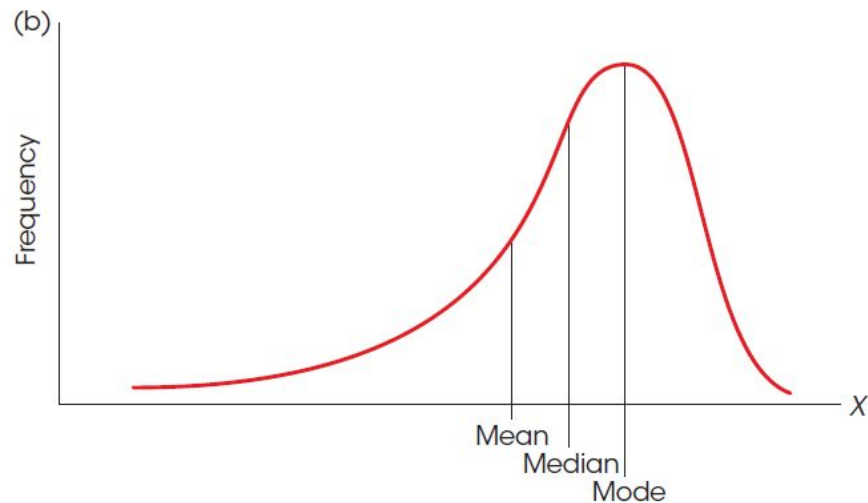
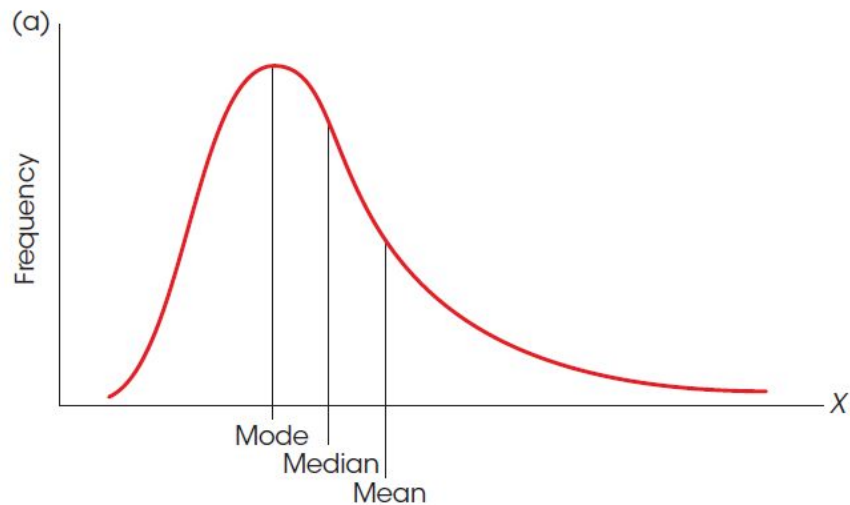


Для нормального
распределения мода,
медиана и среднее совпадают!

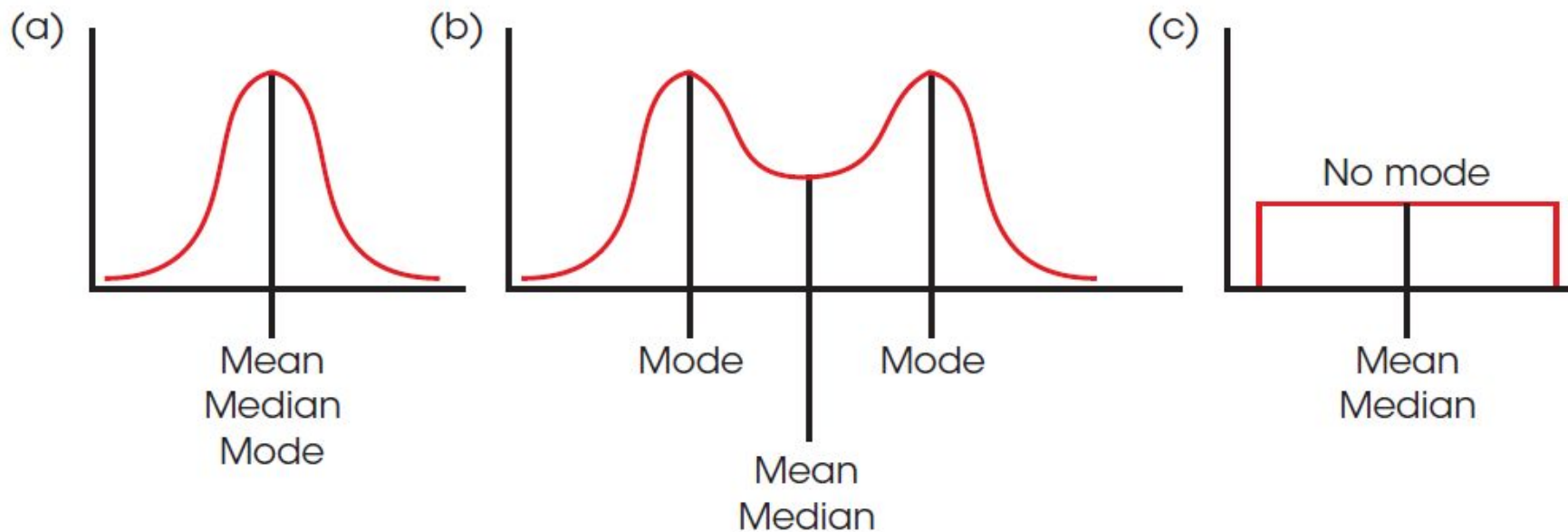


Данные с правым плечом -
медиана меньше среднего.

Центральная тенденция и форма распределения



Центральная тенденция и форма распределения



Меры центральной тенденции в научных текстах

The treatment group showed fewer errors ($M = 2.56$) on the task than the control group ($M = 11.76$).

TABLE 3.6

The mean number of errors made on the task for treatment and control groups according to gender.

	Treatment	Control
Females	1.45	8.36
Males	3.83	14.77

The median number of errors for the treatment group was 8.5, compared to a median of 13 for the control group.