НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

ТЕХНОЛОГИИ АНАЛИЗА БОЛЬШИХ ДАННЫХ

Лекция 2

Маргарита Бурова

Москва, 2019

ГЕНЕРАЛЬНАЯ СОВОКУПНОСТЬ

Генеральная совокупность — совокупность всех объектов, относительно которых предполагается делать выводы при изучении конкретной задачи

ВЫБОРКА

Выборка или выборочная совокупность — множество случаев (испытуемых, объектов, событий, образцов), с помощью определённой процедуры выбранных из генеральной совокупности для анализа

ГЕНЕРАЛЬНАЯ СОВОКУПНОСТЬ И ВЫБОРКА

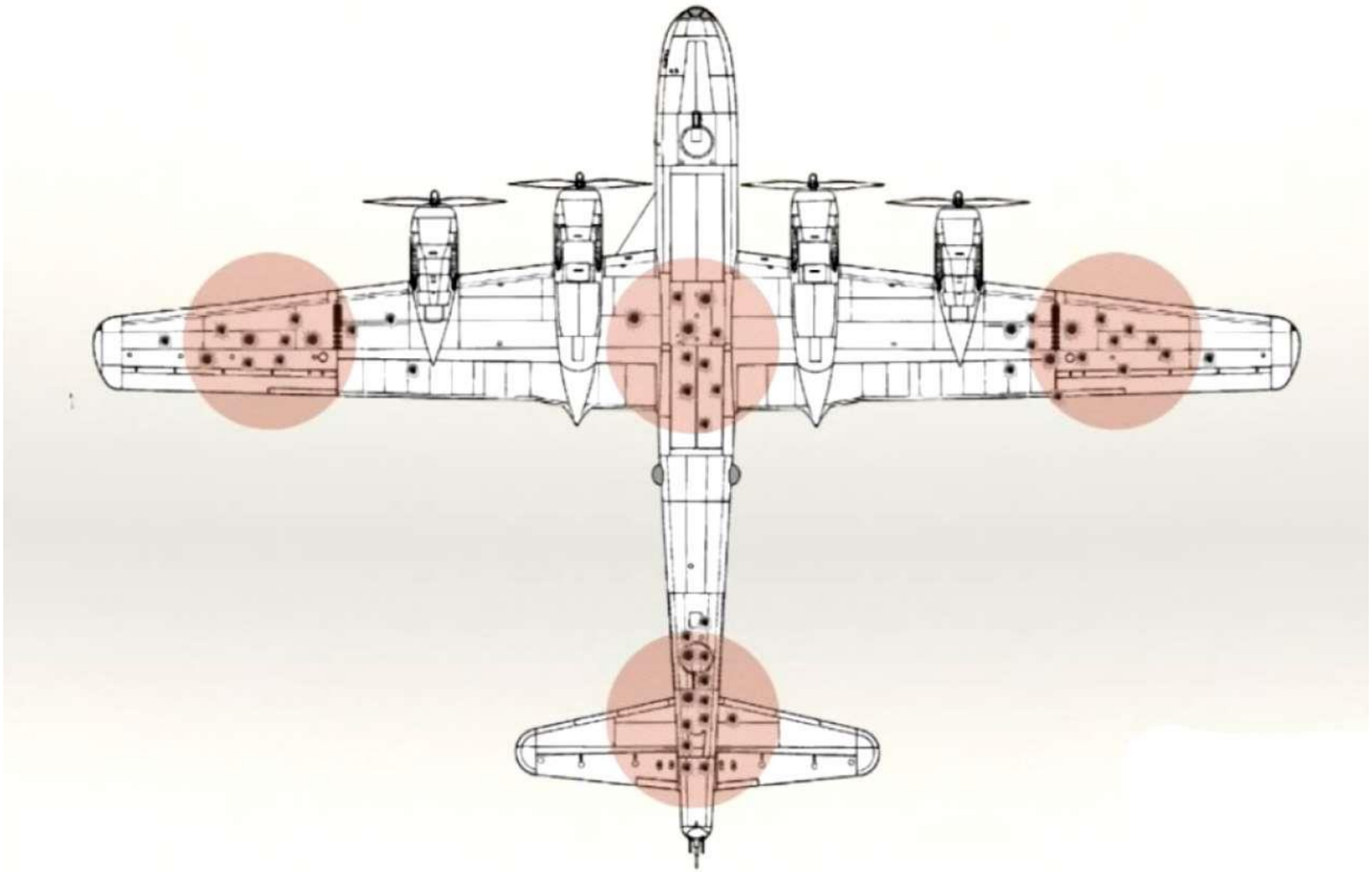
РЕПРЕЗЕНТАТИВНОСТЬ ВЫБОРКИ

Репрезентати́вность — соответствие характеристик выборки характеристикам популяции или генеральной совокупности в целом

РЕПРЕЗЕНТАТИВНОСТЬ ВЫБОРКИ

Рузвельт и Лэндон на выборах 1936 года

ОШИБКА ВЫЖИВШЕГО



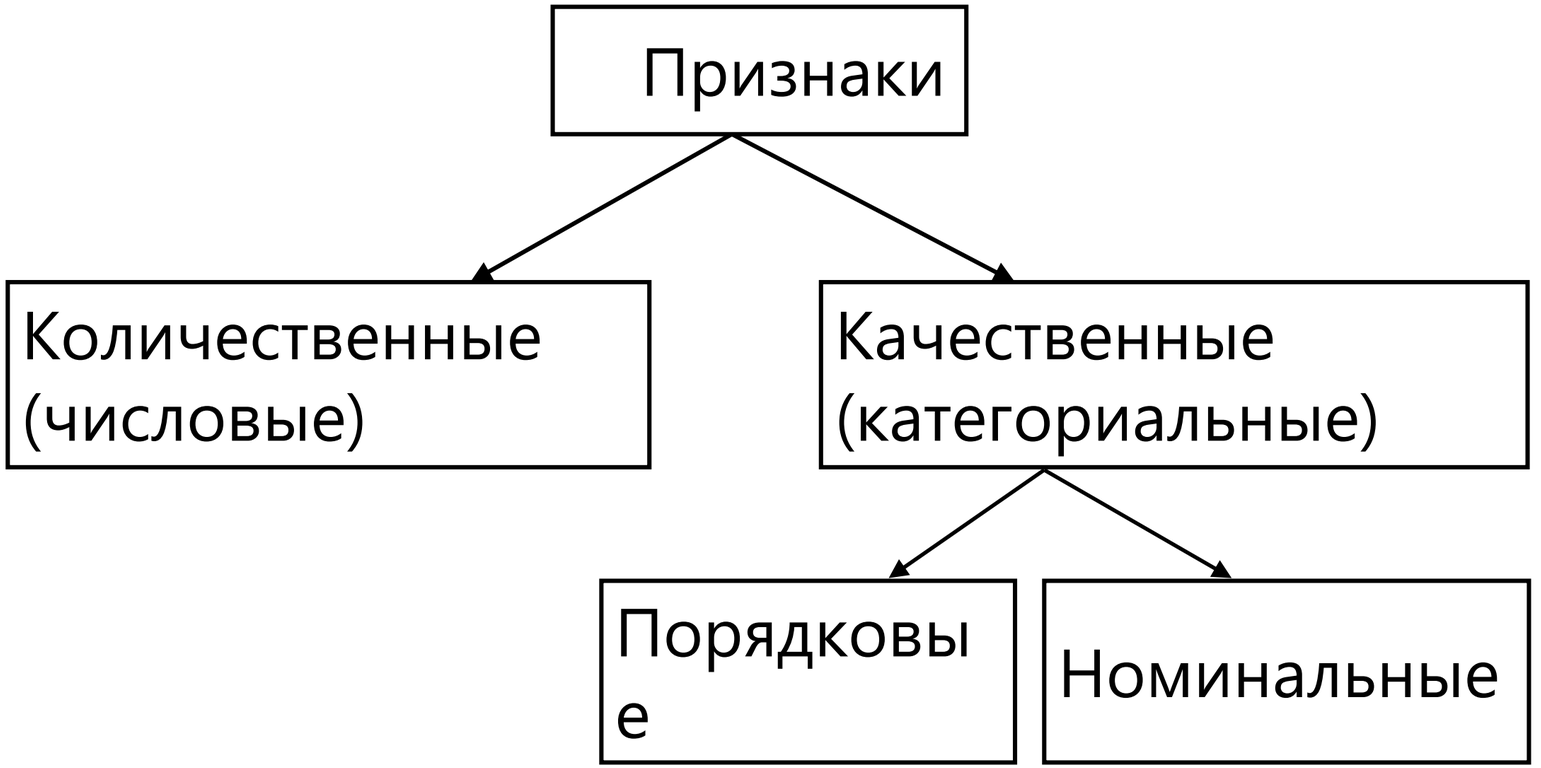
ОШИБКА ВЫЖИВШЕГО

ОБЪЕКТЫ И ПРИЗНАКИ

Выборка состоит из объектов, объекты характеризуются признаками

	Возраст	**Город**	**Уровень образования**
Иванов П.А.	24	Санкт-Петербург	Высшее
Петрова К.В.	35	Москва	Кандидат наук
Семенова Н.К.	31	Иваново	Среднее специальное
Сидоров С.О.	28	Сургут	Доктор наук

ТИПЫ ПРИЗНАКОВ



НОМИНАЛЬНЫЕ ПРИЗНАКИ

Качественные признаки, не подлежащие упорядочиванию

Примеры:

- Город
- Темперамент человека
- Группа крови
- Цвет предмета

ПОРЯДКОВЫЕ ПРИЗНАКИ

Качественные признаки, которые могут быть ранжированы в убывающем или восходящем порядке

Примеры:

- Уровень образования
- Степень ожога
- Социально-экономический статус
- Спортивный разряд

КОЛИЧЕСТВЕННЫЕ ПРИЗНАКИ

Признаки, измеряемые с помощью чисел, имеющих содержательный смысл

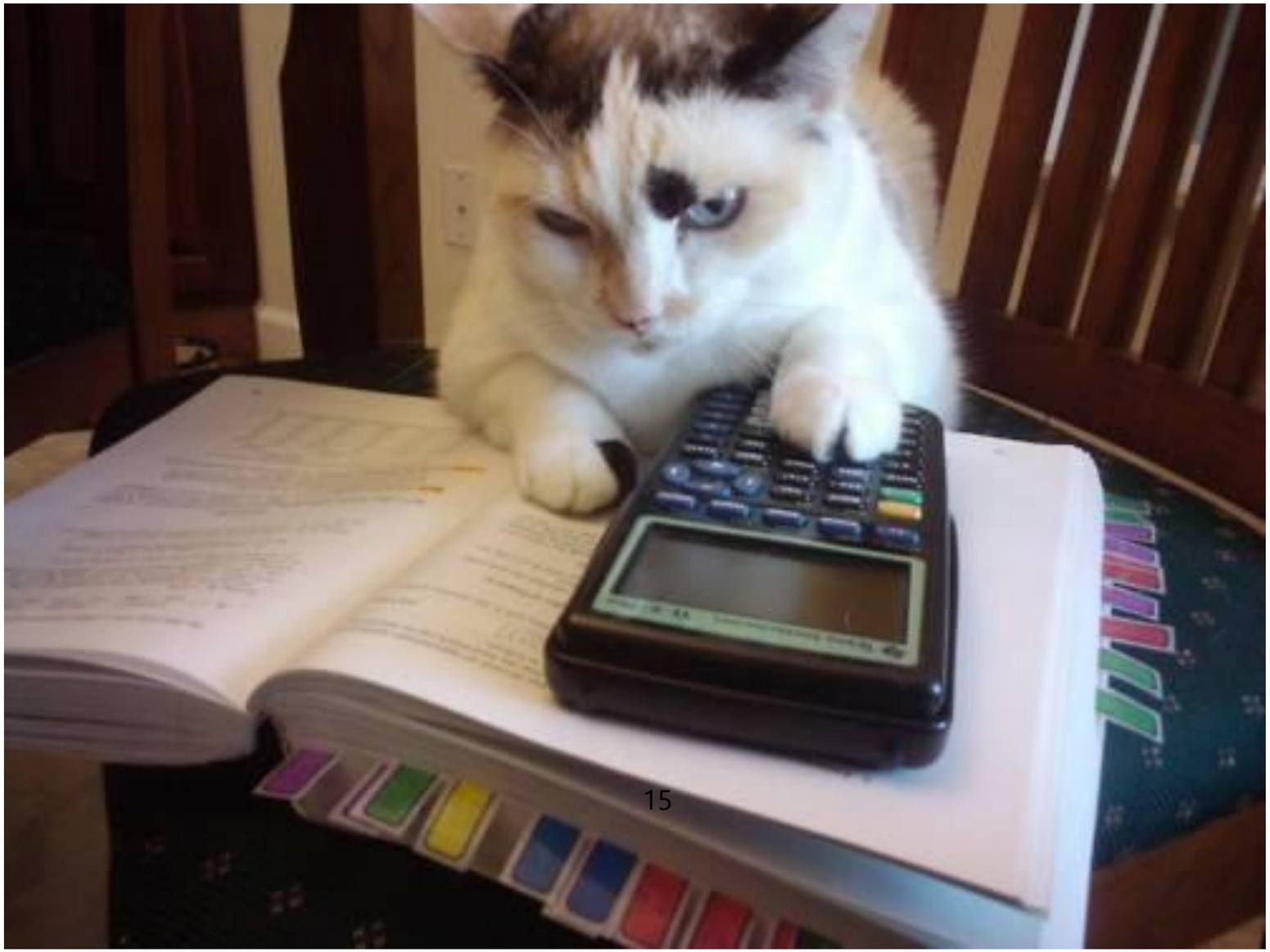
Примеры:

- Рост
- Вес
- Зарплата

ТИПЫ ПРИЗНАКОВ

	Возраст	**Город**	**Уровень образования**
Иванов П.А.	24	Санкт-Петербург	Высшее
Петрова К.В.	35	Москва	Кандидат наук
Семенова Н.К.	31	Иваново	Среднее специальное
Сидоров С.О.	28	Сургут	Доктор наук

ОПИСАТЕЛЬНЫЕ СТАТИСТИКИ: КОТИКИ



МЕРЫ ЦЕНТРАЛЬНОЙ ТЕНДЕНЦИИ

- Среднее арифметическое
- Медиана
- Мода

СРЕДНЕЕ АРИФМЕТИЧЕСКОЕ

$$\text{Среднее} = \frac{\text{СУММА ЭЛЕМЕНТОВ}}{\text{КОЛИЧЕСТВО ЭЛЕМЕНТОВ}}$$

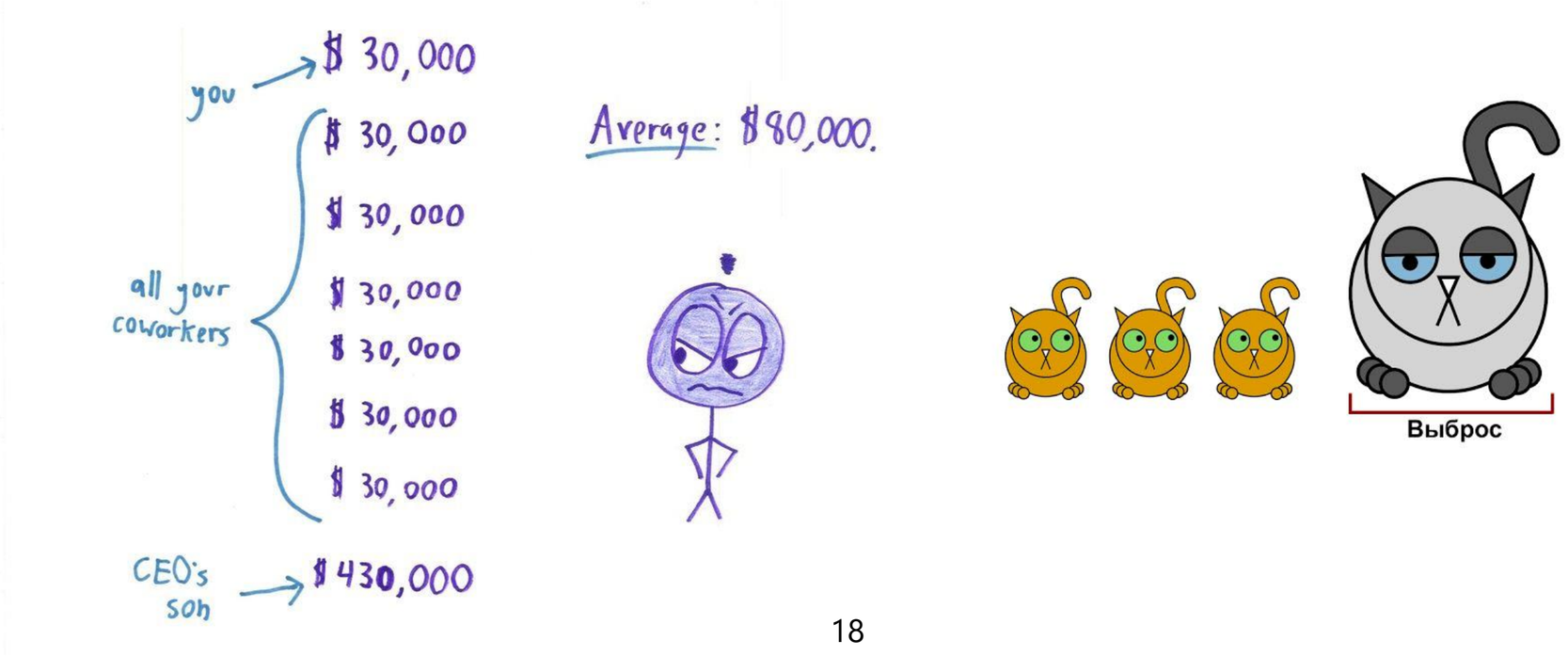
Пример: 1,2,6,6,7

$$\text{Среднее} = \frac{1+2+6+6+7}{5} = \frac{22}{5} = 4{,}4$$

*Иллюстрации взяты из книги «Статистика и котики»

СРЕДНЕЕ АРИФМЕТИЧЕСКОЕ

Минус данной МЦТ: чувствительность к выбросам



*Иллюстрации взяты из книги «Статистика и котики»

МЕДИАНА

Алгоритм нахождения медианы:

1. Расположить значения по возрастанию
2. Если количество значений нечетное, то медианой будет центральное значение в ряду
3. Если количество значений четное, то для вычисления медианы необходимо найти среднее арифметическое двух центральных значений

МЕДИАНА: ПРИМЕР

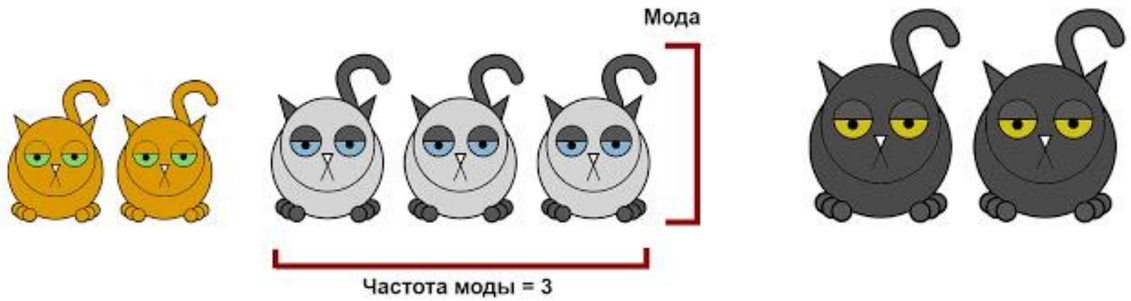
1. Дан числовой ряд: 1,5,3,9,11, 2, 14, 6
2. Расположим числа в порядке возрастания:

1,2,3,5,6,9,11,14

3. Найдем центральные числа: 5 и 6
4. Найдем их среднее арифметическое: (5+6):2
5. Получаем, что значение медианы равно 5,5

МОДА

Мода-наиболее часто встречающееся значение



*Иллюстрации взяты из книги «Статистика и котики»

МОДА

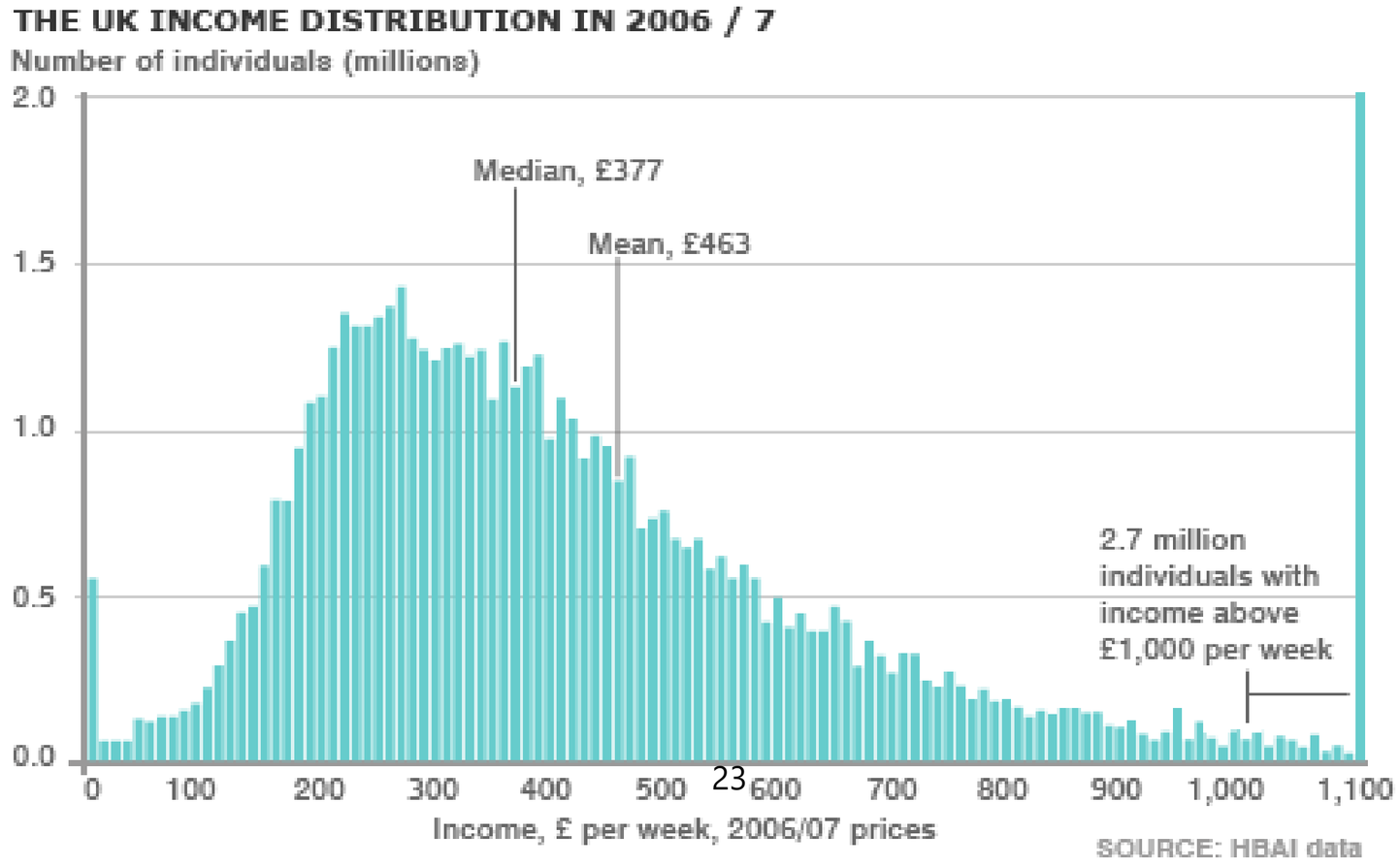
Пример вычисления моды:

1.Пусть дан числовой ряд 1,6,1,7,1,4,5,5

2.Чаще всего в нем встречается единица

3.Получается, что мода данного ряда равна одному

ДОЛЖНЫ ЛИ СОВПАДАТЬ МЦТ?



МЕРЫ РАЗБРОСА

- Размах
- Межквартильный размах
- Стандартное отклонение
- Дисперсия

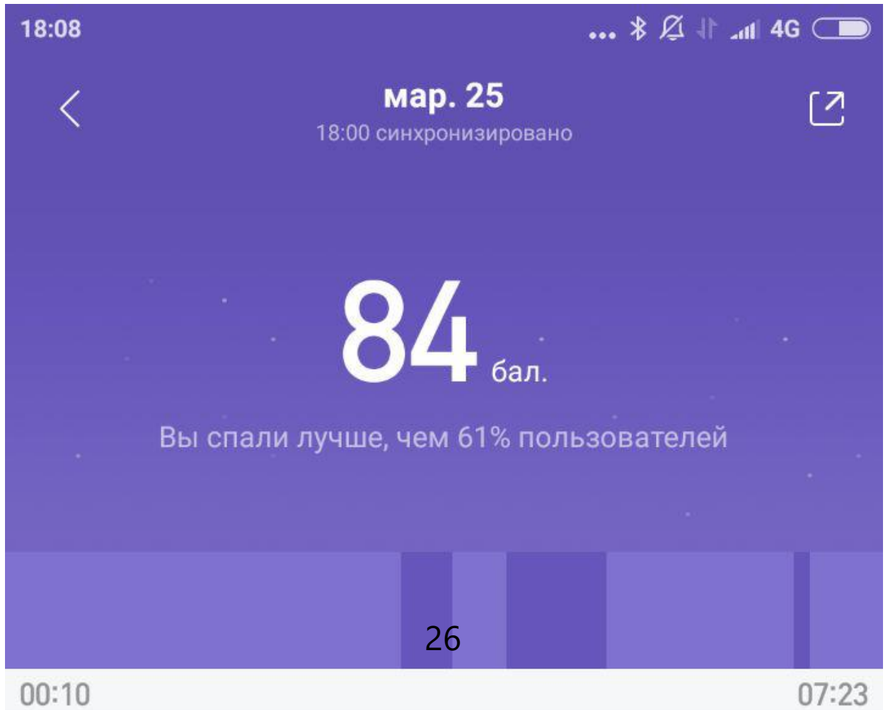
КВАНТИЛИ

Квантиль в математической статистике — значение, которое заданная случайная величина не превышает с фиксированной вероятностью. Если вероятность задана в процентах, то квантиль называется процентилем или перцентилем

- 0,25-квантиль называется первым (или нижним) квартилем (от лат. *quarta* — четверть);
- 0,5-квантиль называется медианой (от лат. *mediana* — середина) или вторым квартилем;
- 0,75-квантиль называется третьим (или верхним) квартилем.

ПЕРЦЕНТИЛИ

Я спала лучше, чем 61% пользователей.
Значит, 25 марта я находилась в 61-ом процентиле



МЕРЫ РАЗБРОСА: СТАНДАРТНОЕ ОТКЛОНЕНИЕ

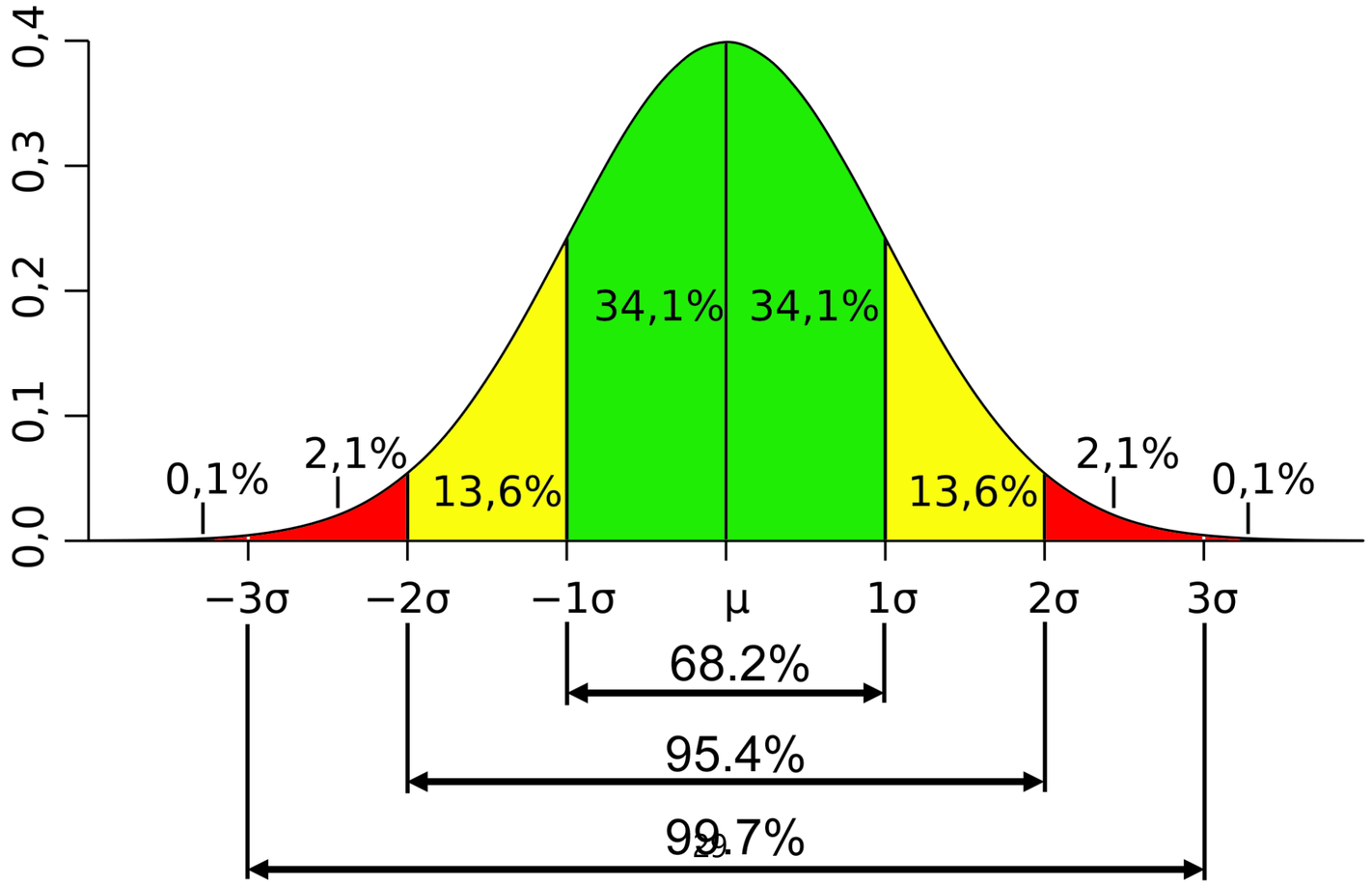
Стандартное отклонение- показатель рассеивания значений случайной величины относительно её математического ожидания.

*Иллюстрации взяты из книги «Статистика и котики»

МЕРЫ РАЗБРОСА: ДИСПЕРСИЯ

*Иллюстрации взяты из книги «Статистика и котики»

НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ



МЕРЫ И ТИПЫ ПРИЗНАКОВ

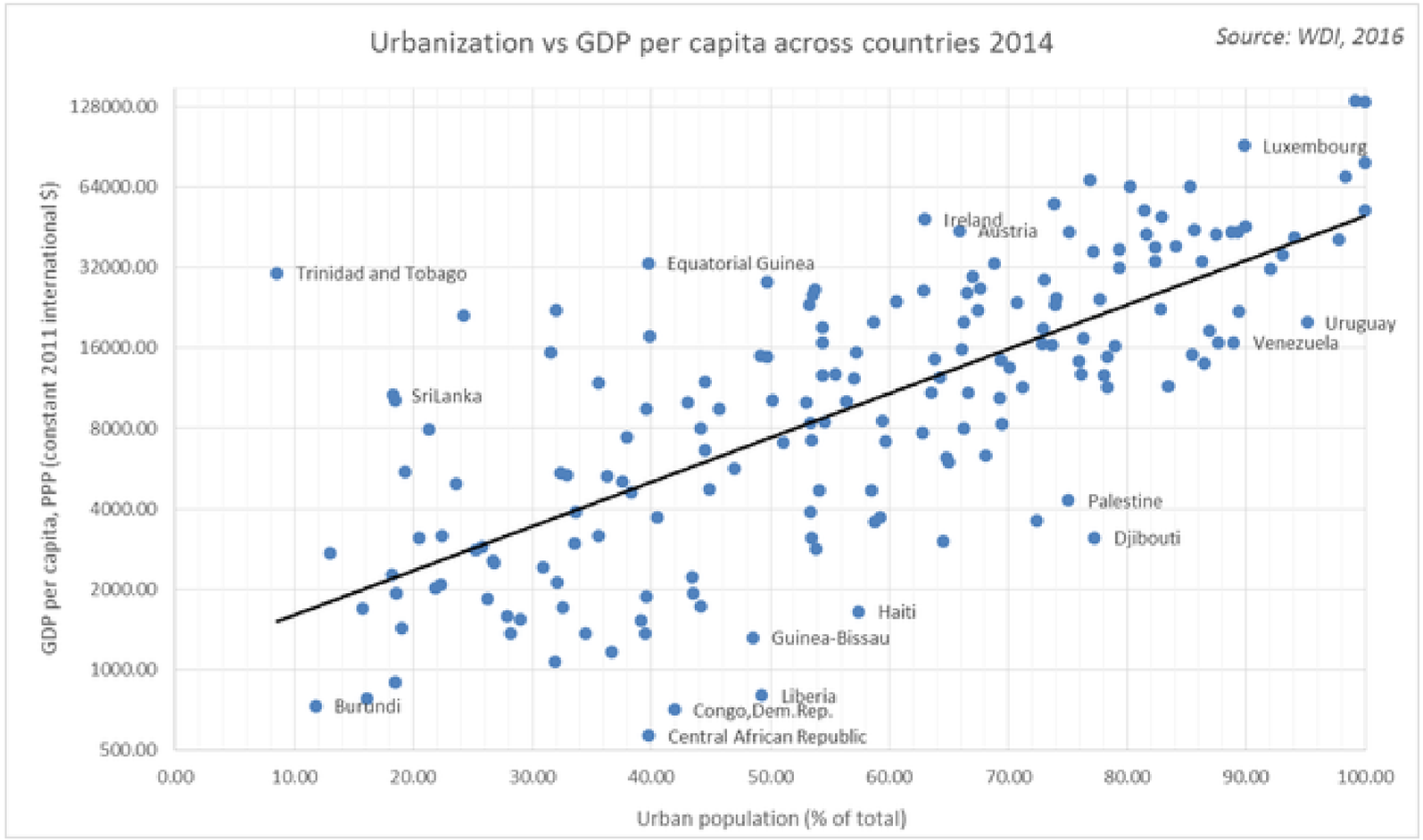
Типы данных	Меры центра			Меры разброса		
	Мода	Медиана	Среднее	Размах	Q-Q	Ст.Откл.
Номинальные	✓	✗	✗	✗	✗	✗
Порядковые	✓	✓	✗	✓	✓	✗
Количественные	✓	✓	✓	✓	✓	✓

КОРРЕЛЯЦИЯ

Корреляция – мера взаимосвязи двух величин

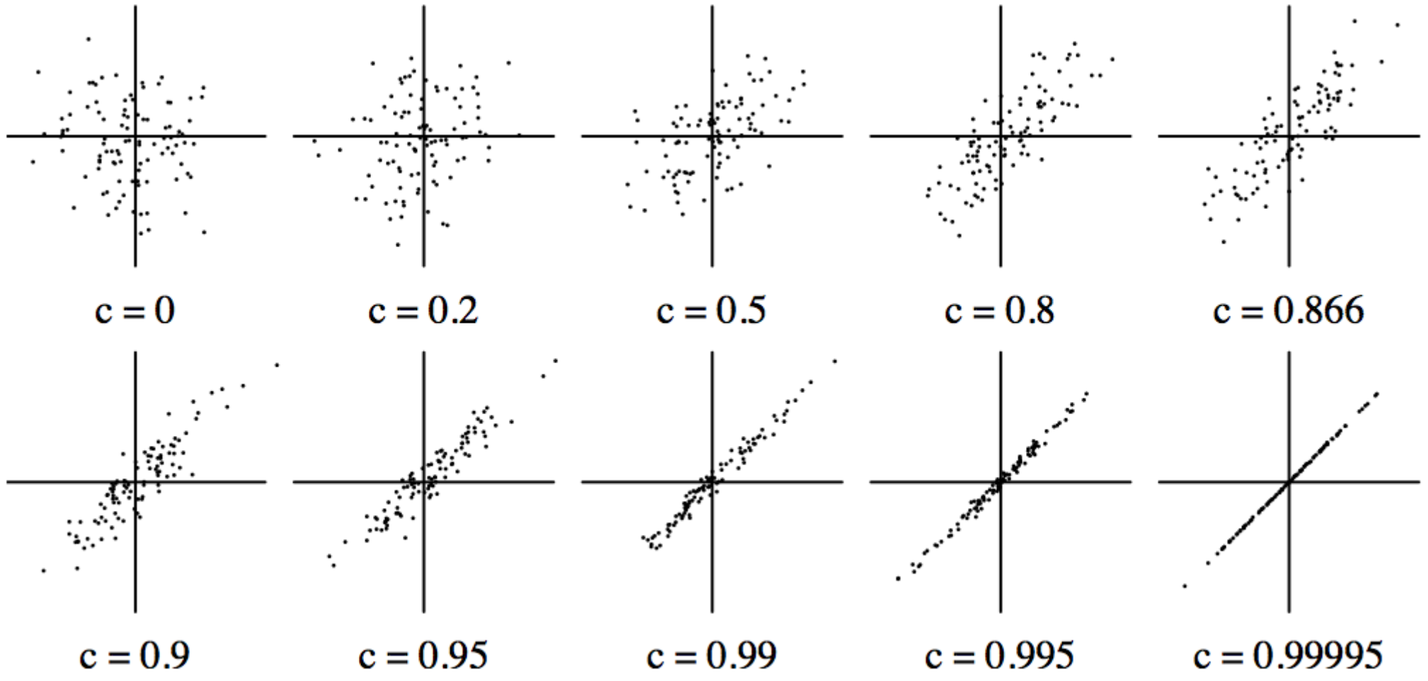
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

КОРРЕЛЯЦИЯ


Urbanization vs GDP per capita across countries 2014
Source: WDI, 2016
GDP per capita, PPP (constant 2011 international $)
128000.00
64000.00
32000.00
16000.00
8000.00
4000.00
2000.00
1000.00
500.00
0.00
10.00
20.00
30.00
40.00
50.00
60.00
70.00
80.00
90.00
100.00
Urban population (% of total)
Luxembourg
Ireland
Austria
Trinidad and Tobago
Equatorial Guinea
Uruguay
Venezuela
SriLanka
Palestine
Djibouti
Haiti
Guinea-Bissau
Liberia
Burundi
Congo,Dem.Rep.
Central African Republic

КОРРЕЛЯЦИЯ

Корреляция – мера взаимосвязи двух величин

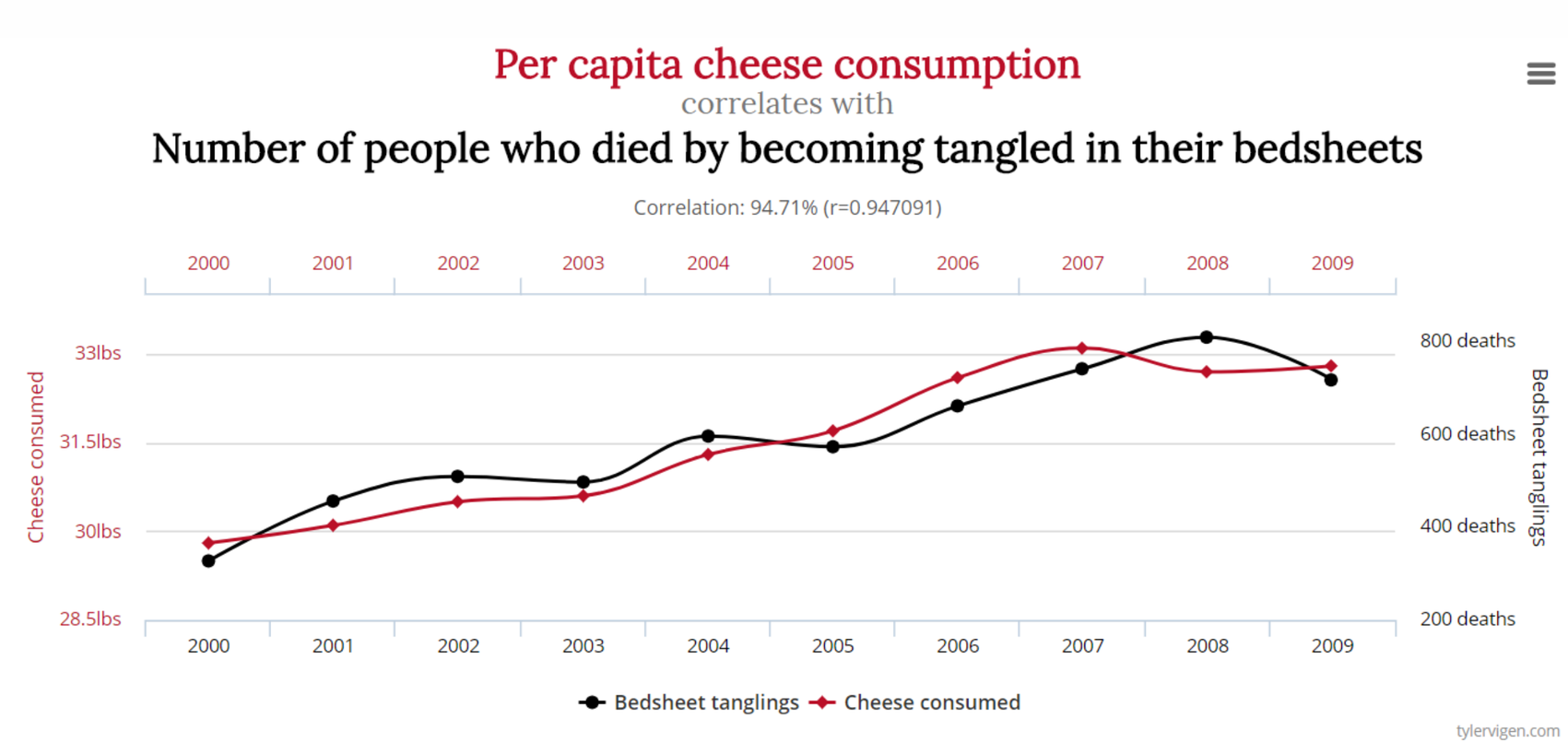


КОРРЕЛЯЦИЯ

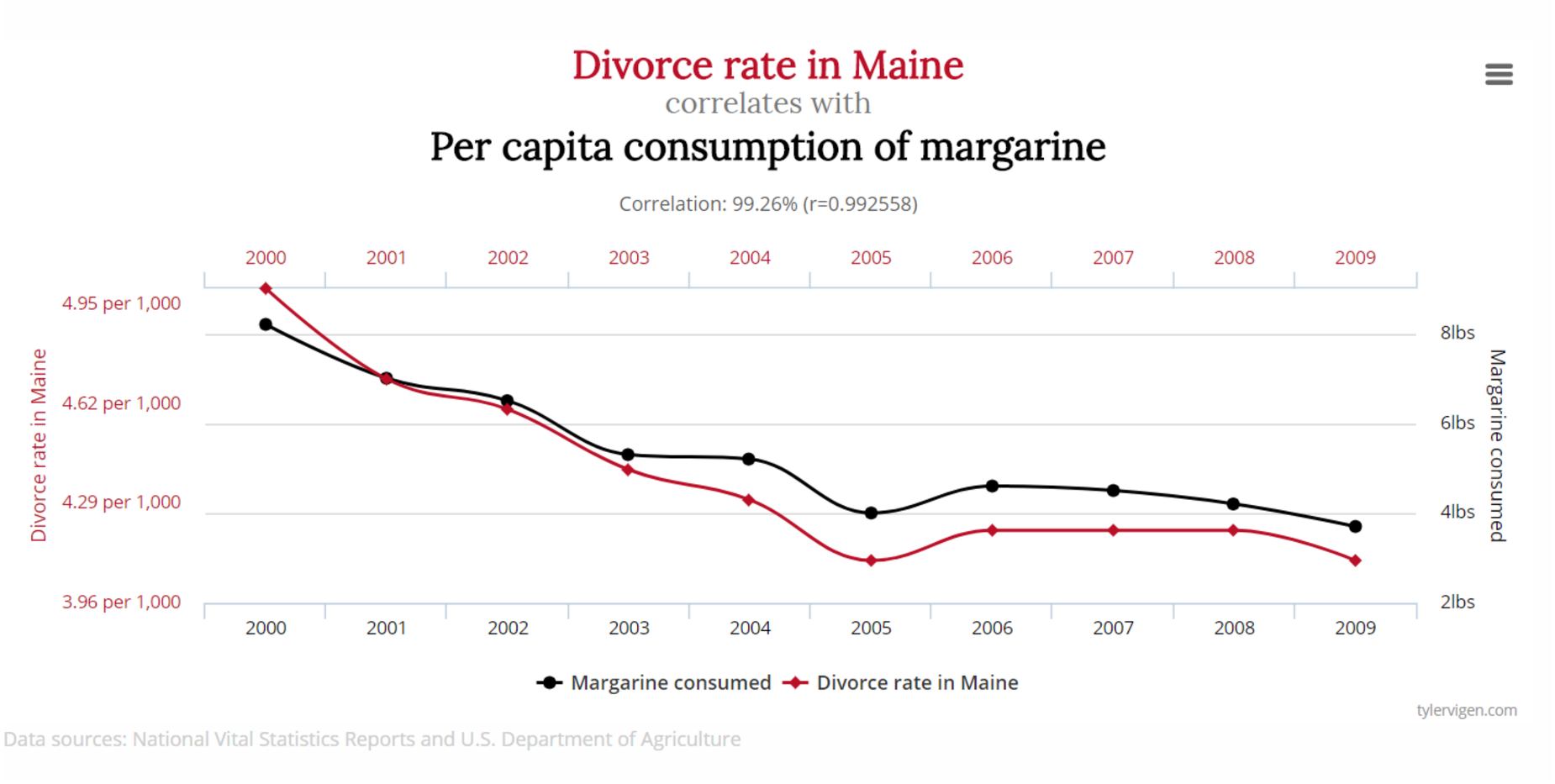
Свойства корреляции:

- Всегда принимает значения от -1 до 1
- Положительный коэффициент свидетельствует о прямой зависимости
- Отрицательный коэффициент свидетельствует об обратной зависимости

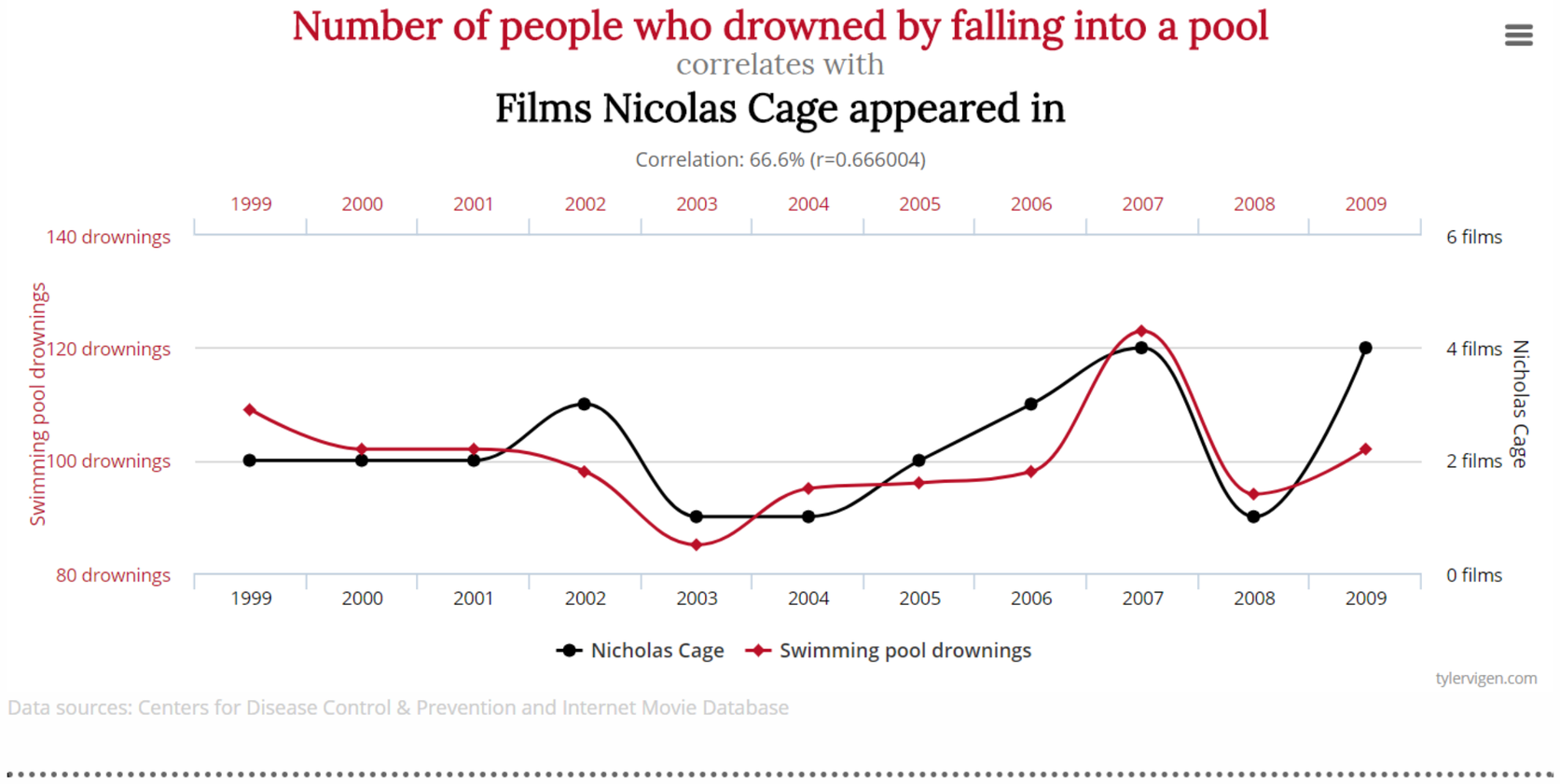
КОРРЕЛЯЦИЯ


Per capita cheese consumption
correlates with
Number of people who died by becoming tangled in their bedsheets
Correlation: 94.71% (r=0.947091)
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
33lbs
31.5lbs
30lbs
28.5lbs
Cheese consumed
800 deaths
600 deaths
400 deaths
200 deaths
Bedsheet tanglings
Bedsheet tanglings
Cheese consumed
tylervigen.com

КОРРЕЛЯЦИЯ

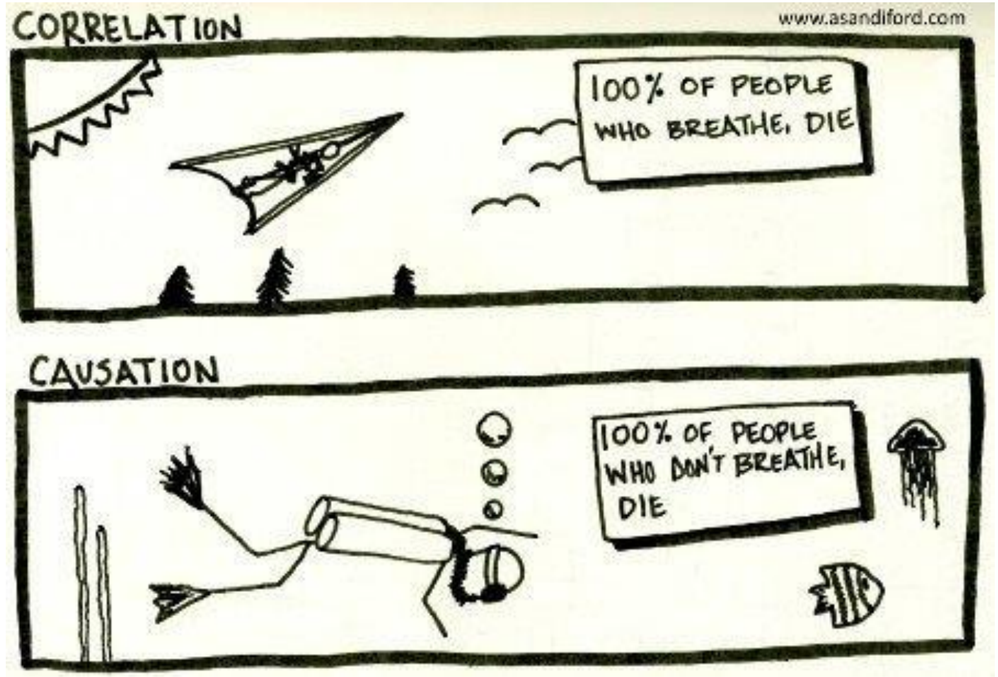

Divorce rate in Maine
correlates with
Per capita consumption of margarine
Correlation: 99.26% (r=0.992558)
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
4.95 per 1,000
4.62 per 1,000
4.29 per 1,000
3.96 per 1,000
Divorce rate in Maine
8lbs
6lbs
4lbs
2lbs
Margarine consumed
Margarine consumed
Divorce rate in Maine
tylervigen.com
Data sources: National Vital Statistics Reports and U.S. Department of Agriculture

КОРРЕЛЯЦИЯ


Number of people who drowned by falling into a pool
correlates with
Films Nicolas Cage appeared in
Correlation: 66.6% (r=0.666004)
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
140 drownings
120 drownings
100 drownings
80 drownings
Swimming pool drownings
6 films
4 films
2 films
0 films
Nicholas Cage
Nicholas Cage
Swimming pool drownings
tylervigen.com
Data sources: Centers for Disease Control & Prevention and Internet Movie Database

КОРРЕЛЯЦИЯ

ВАЖНО: корреляция – не является поводом для того, чтобы делать выводы о причинно-следственных связях



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ