



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

Научно-исследовательский семинар 1 курса
ОП «Политология»

ОСНОВЫ АНАЛИЗА ДАННЫХ В МЕЖДУНАРОДНЫХ ОТНОШЕНИЯХ

Лекция 1

Маргарита Бурова

Москва, 2018



ОРГАНИЗАЦИОННЫЕ ВОПРОСЫ

Формула оценки:

- Индивидуальный проект – 15%
- Групповой проект – 15%
- Две самостоятельные работы по 15%
- Экзамен – 40%

Где найти основную информацию по курсу:

- [http://wiki.cs.hse.ru/Основы анализа данных в международных отношениях](http://wiki.cs.hse.ru/Основы_анализа_данных_в_международных_отношениях)
- Телеграм канал (инвайт был выслан в чат курса)



ПРЕПОДАВАТЕЛИ

Лекции:

Бурова Маргарита Борисовна

Семинары:

БМО1 : Бурова Маргарита Борисовна

БМО2,БМО3: Попенова Полина Сергеевна

БМО4: Петросян Артур Тигранович

Учебные ассистенты:

БМО1:Артемов Максим

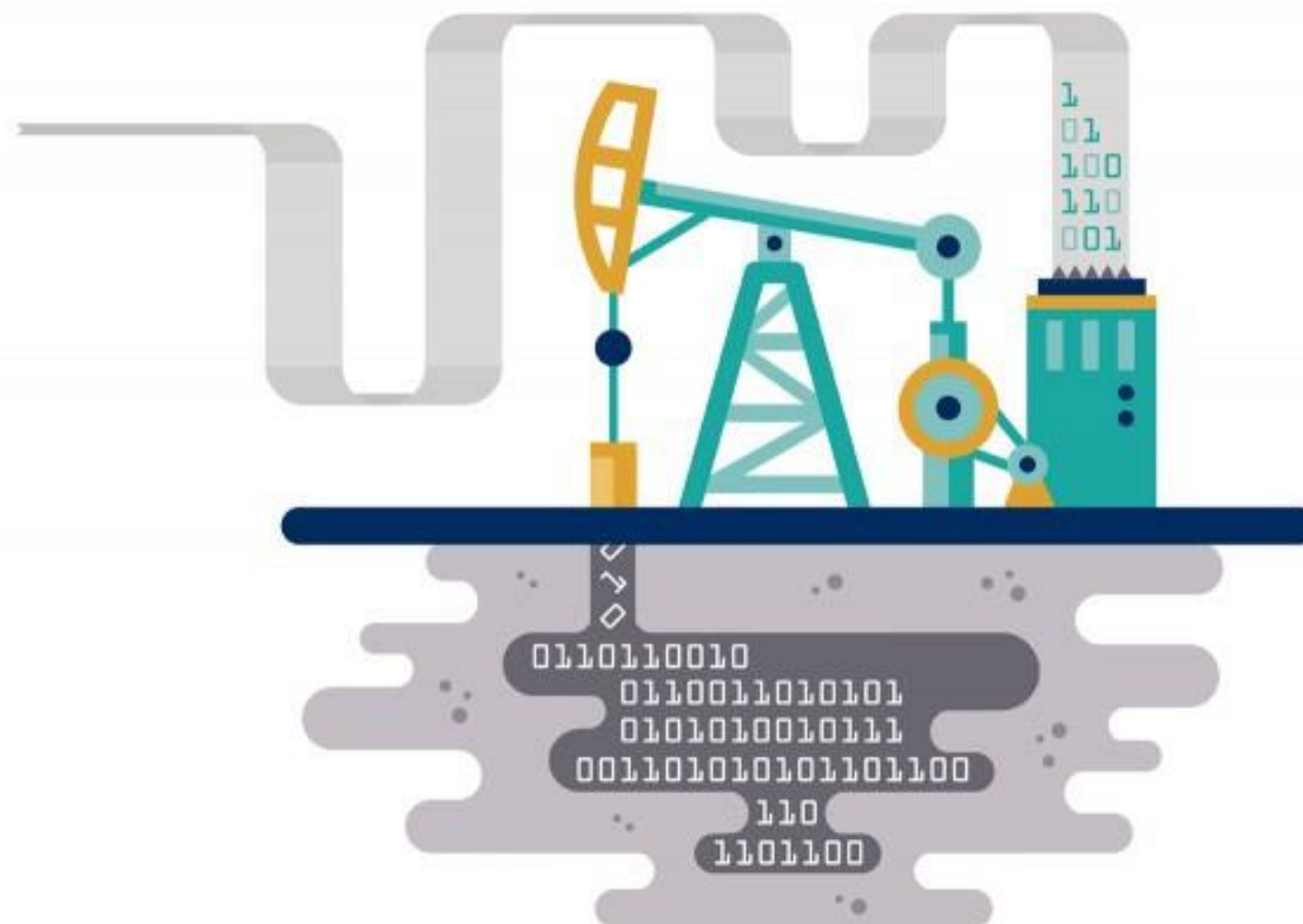
БМО2: Турышев Арсений

БМО3:Пузырев Дмитрий

БМО4:Ксения Закирова



Данные-нефть 21 века(с)





2018 This is What Happens in an **INTERNET MINUTE**





НЕСКОЛЬКО ИНТЕРЕСНЫХ КЕЙСОВ

В 2018 году математик Бен Зозмер предсказал победителей Оскара во всех номинациях





НЕСКОЛЬКО ИНТЕРЕСНЫХ КЕЙСОВ

Интернет-магазин может узнать о Вашей беременности раньше Ваших близких





НЕСКОЛЬКО ИНТЕРЕСНЫХ КЕЙСОВ

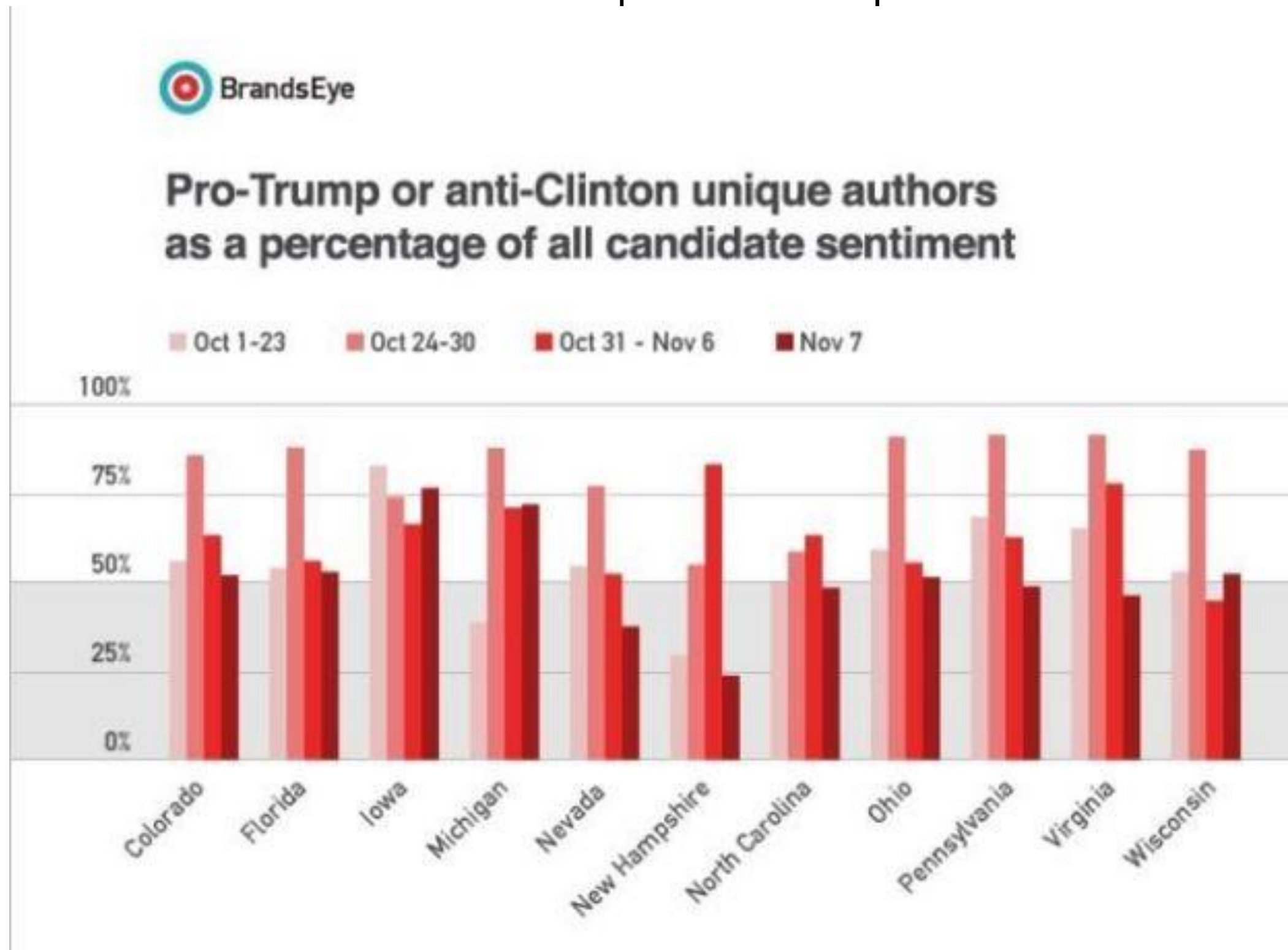
Фитнес-браслеты: данные с трекера сна как источник дорогостоящей информации





ПРЕДСКАЗАНИЕ РЕЗУЛЬТАТОВ ВЫБОРОВ

С помощью твиттера:





ЧЕМУ МЫ НАУЧИМСЯ НА ЭТОМ КУРСЕ?

- Основы статистического анализа данных
- Анализ социальных сетей
- Анализ текстов
- Представление результатов исследования





ЧЕМ МЫ БУДЕМ ПОЛЬЗОВАТЬСЯ?





ОСНОВНЫЕ ПОНЯТИЯ СТАТИСТИКИ

Генеральной называют совокупность всех объектов, которые подвергаются обследованию или изучению.

Выборкой или выборочной совокупностью называется часть отобранных элементов из всей совокупности.





РЕПРЕЗЕНТАТИВНОСТЬ ВЫБОРКИ

Рузвельт и Лэндон на выборах 1936 года





МЕРЫ ЦЕНТРАЛЬНОЙ ТЕНДЕНЦИИ

- Среднее арифметическое
- Медиана
- Мода



СРЕДНЕЕ АРИФМЕТИЧЕСКОЕ

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



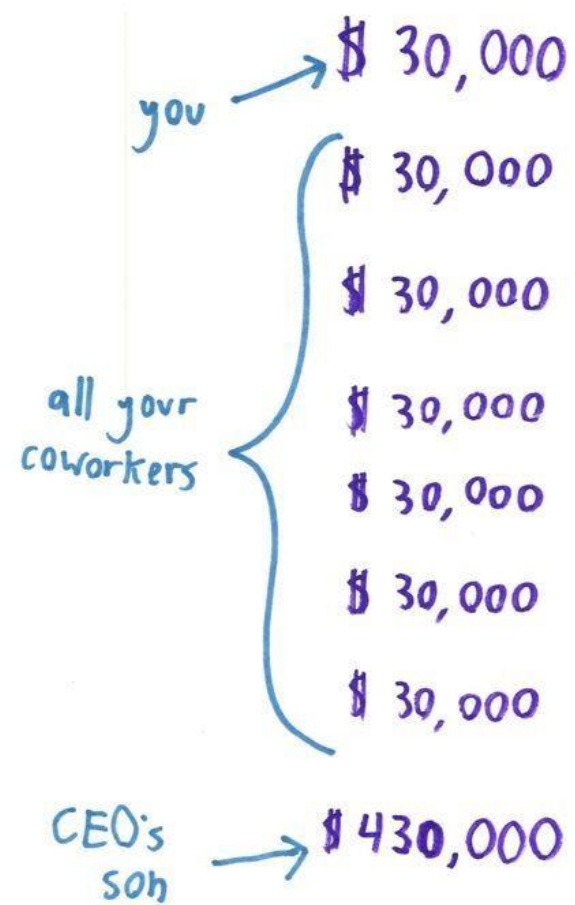
Среднее значение \bar{x}

Пример: 1,2,6,6,7

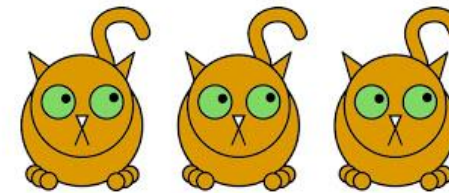
$$\text{Среднее} = \frac{1+2+6+6+7}{5} = \frac{22}{5} = 4,4$$

СРЕДНЕЕ АРИФМЕТИЧЕСКОЕ

Минус данной МЦТ: чувствительность к выбросам



Average: \$80,000.

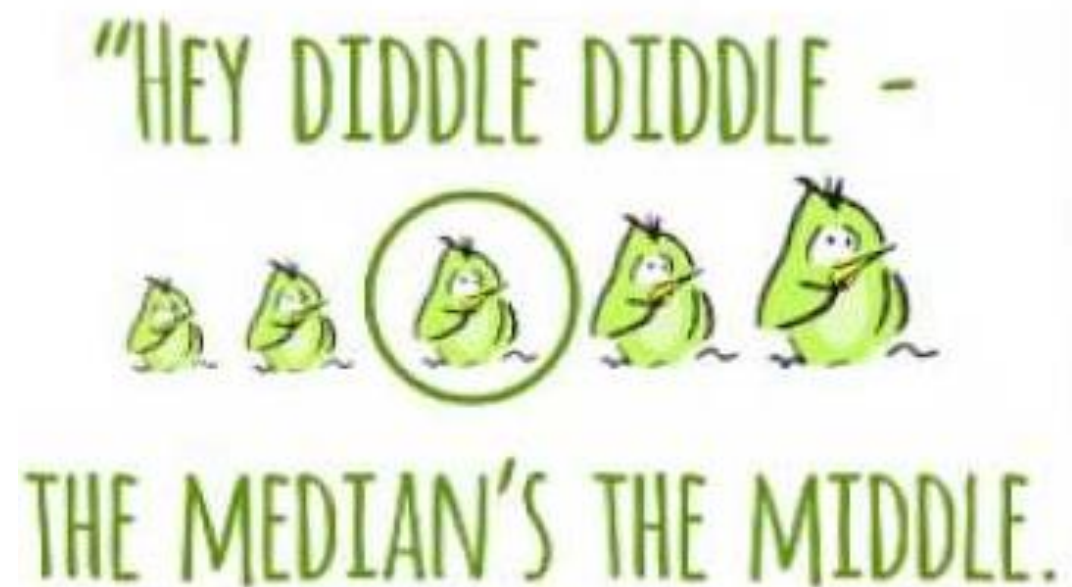




МЕДИАНА

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





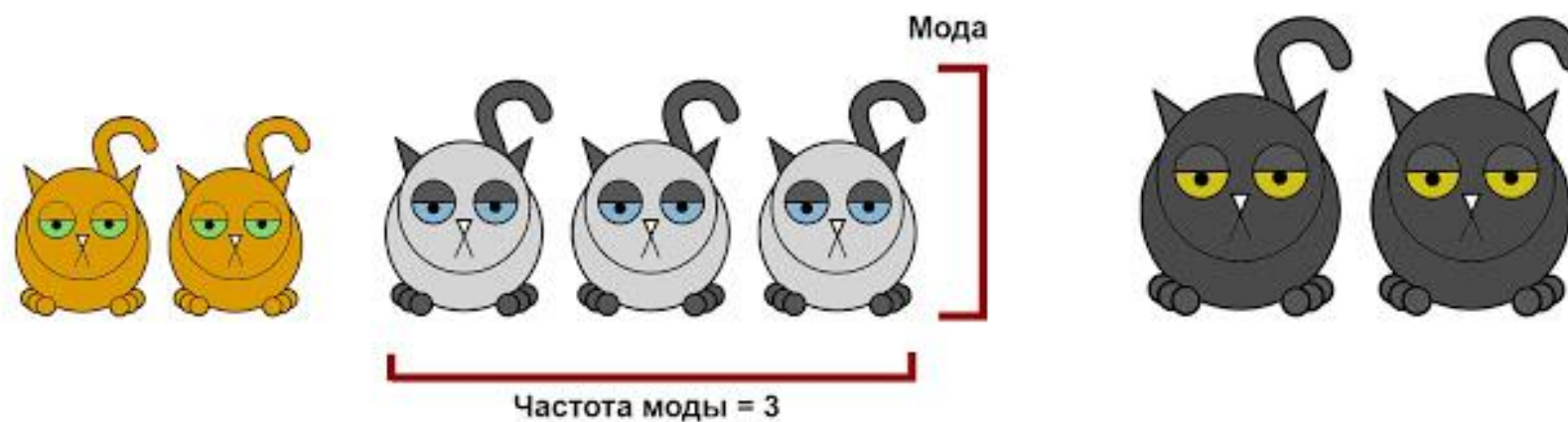
МЕДИАНА: ПРИМЕР

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





МОДА

Пример вычисления моды:

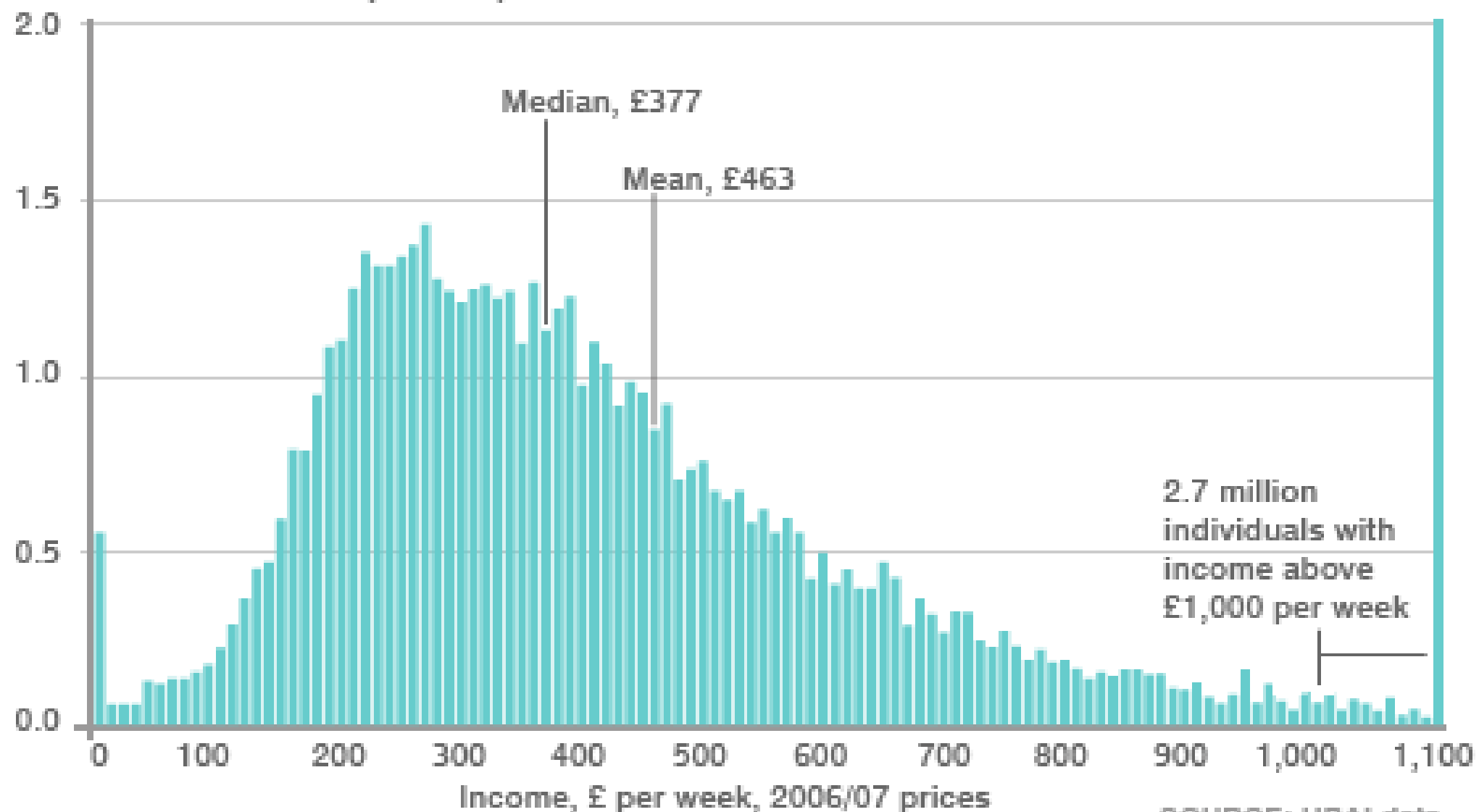
1. Пусть дан числовой ряд 1,6,1,7,1,4,5,5
2. Чаще всего в нем встречается единица
3. Получается, что мода данного ряда равна одному



МЦТ: ПРИМЕР

THE UK INCOME DISTRIBUTION IN 2006 / 7

Number of individuals (millions)





КВАНТИЛИ

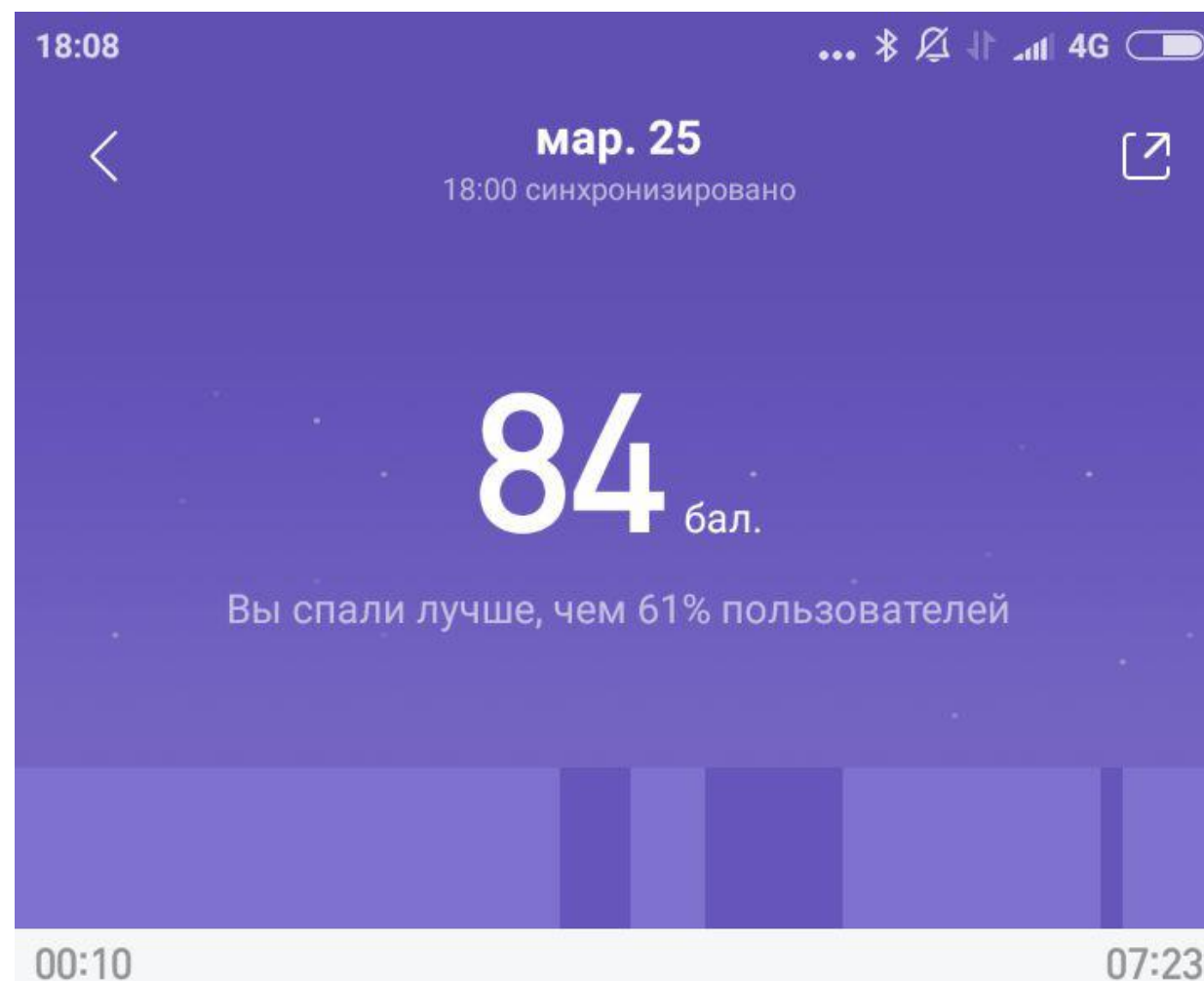
Кванти́ль в математической статистике — значение, которое заданная случайная величина не превышает с фиксированной вероятностью. Если вероятность задана в процентах, то квантиль называется проценти́лем или перценти́лем

- 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 И ДИСПЕРСИЯ

Дисперсия и стандартное отклонение – показатели рассеивания значений случайной величины относительно её математического ожидания.





КОРРЕЛЯЦИЯ

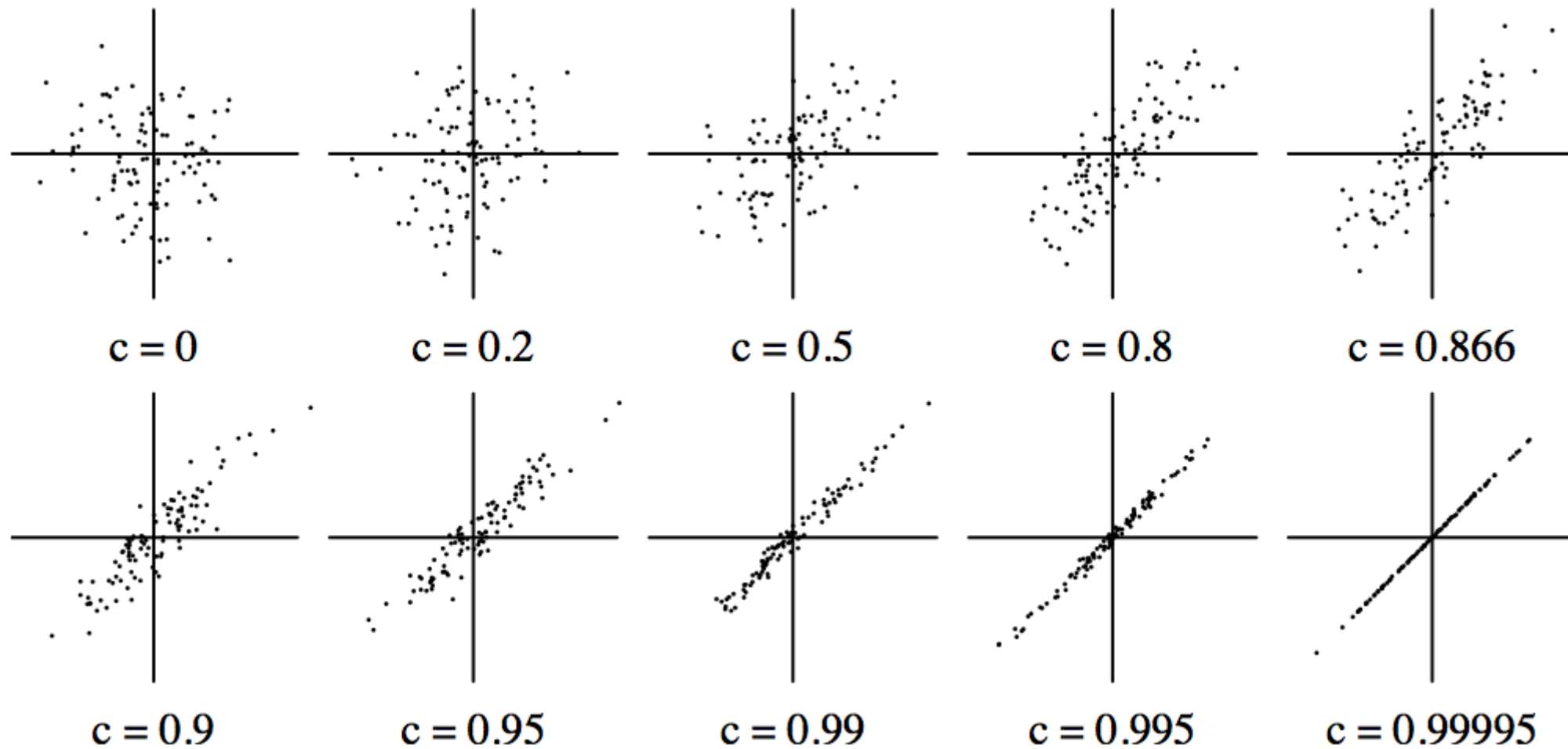
Корреляция – мера взаимосвязи двух величин

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



КОРРЕЛЯЦИЯ

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

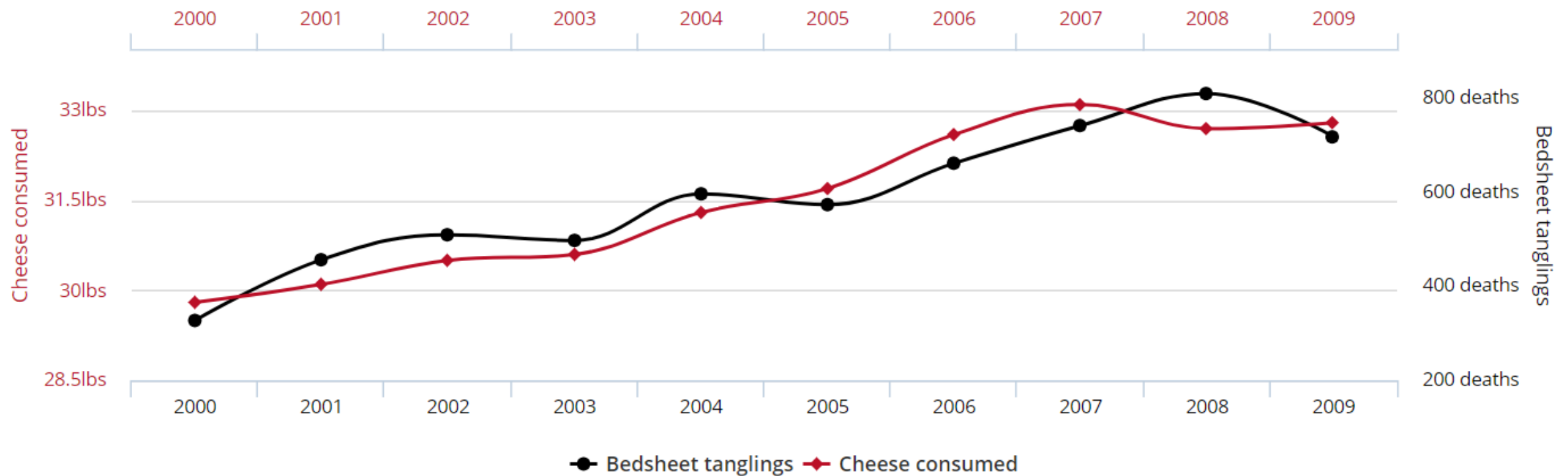


КОРРЕЛЯЦИЯ

Per capita cheese consumption
correlates with

Number of people who died by becoming tangled in their bedsheets

Correlation: 94.71% ($r=0.947091$)

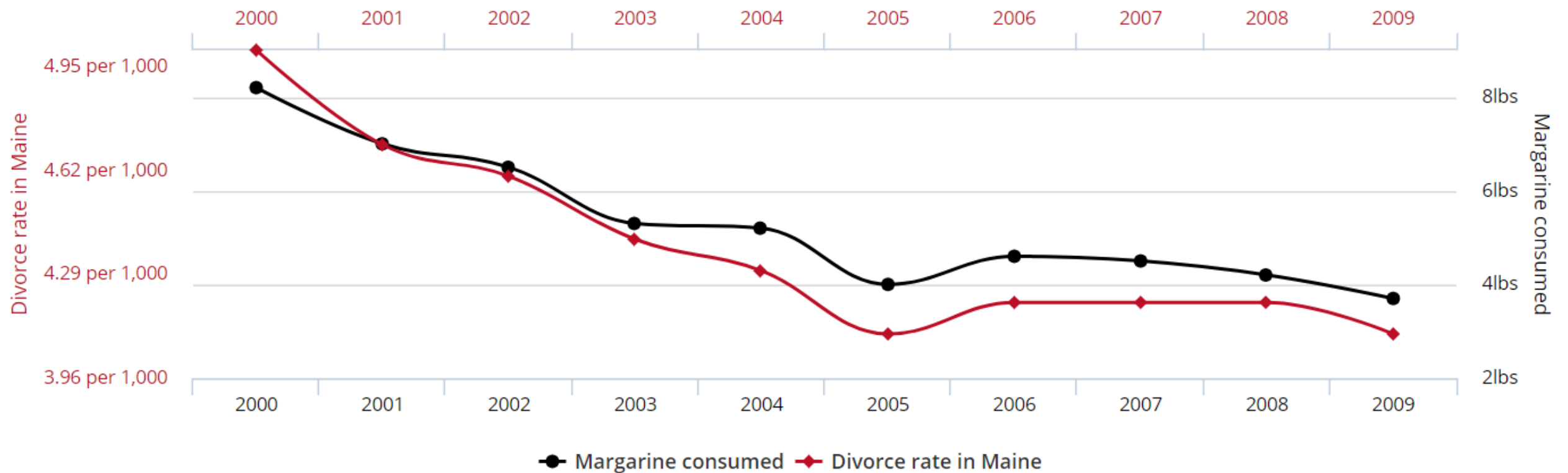




КОРРЕЛЯЦИЯ

Divorce rate in Maine correlates with Per capita consumption of margarine

Correlation: 99.26% ($r=0.992558$)



tylervigen.com

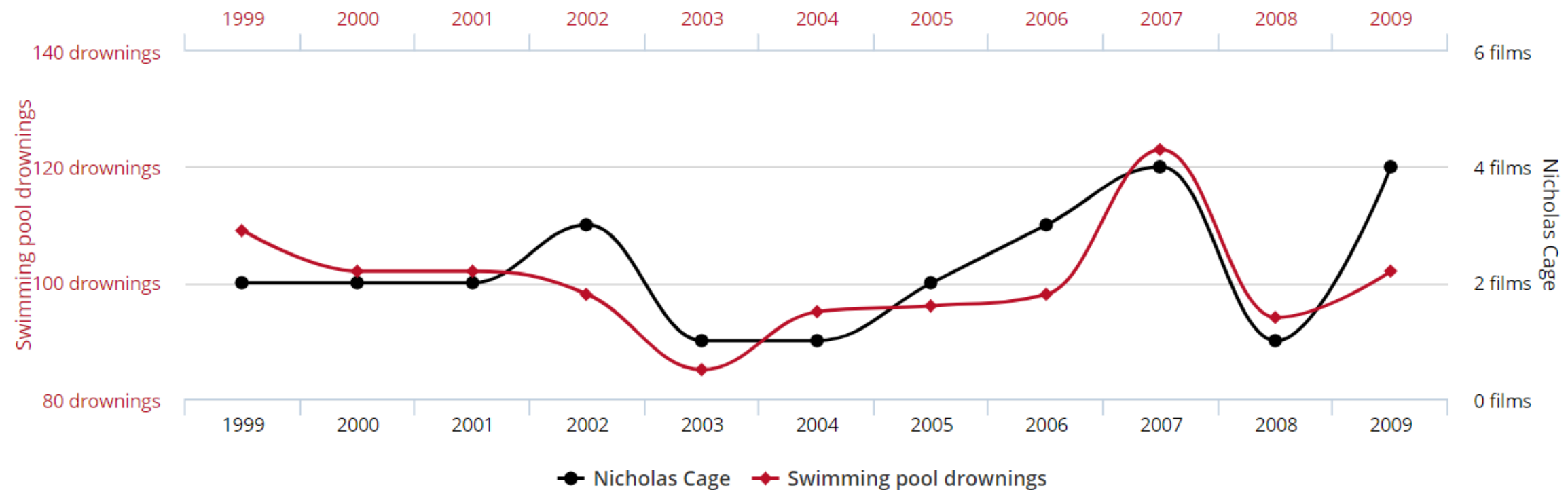
Data sources: National Vital Statistics Reports and U.S. Department of Agriculture



КОРРЕЛЯЦИЯ

Number of people who drowned by falling into a pool correlates with Films Nicolas Cage appeared in

Correlation: 66.6% ($r=0.666004$)



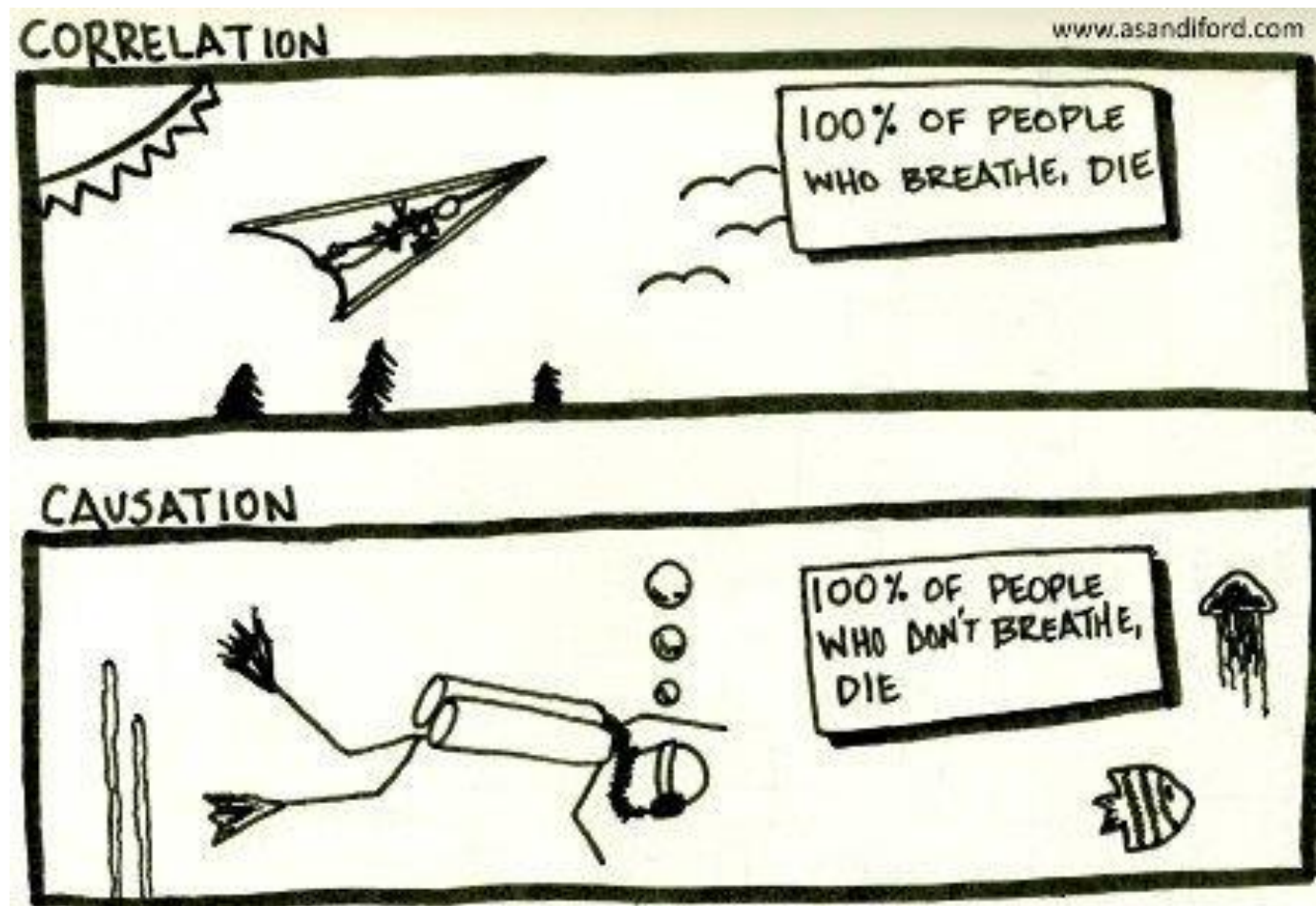
tylervigen.com

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