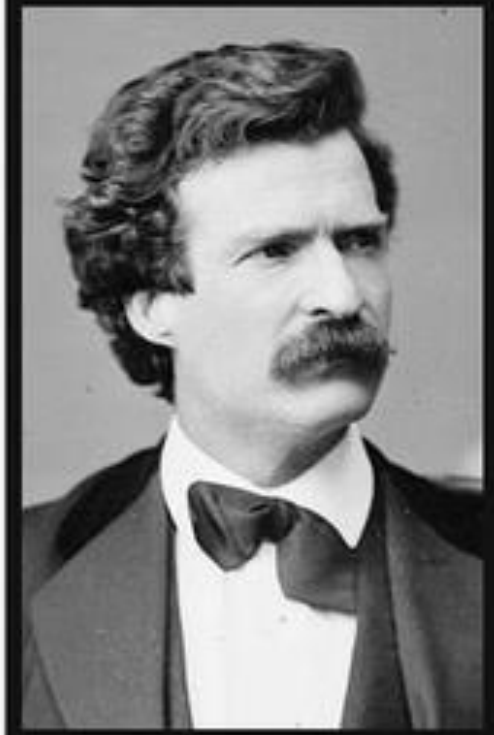


Существуют три вида лжи: ложь, наглая ложь и статистика.

(Марк Твен)

План

- Зачем нужна математическая статистика?
- Литература для изучения
- Генеральная совокупность и выборка
- Описательные статистики
- Корреляция
- A/B тестирование и статистические гипотезы

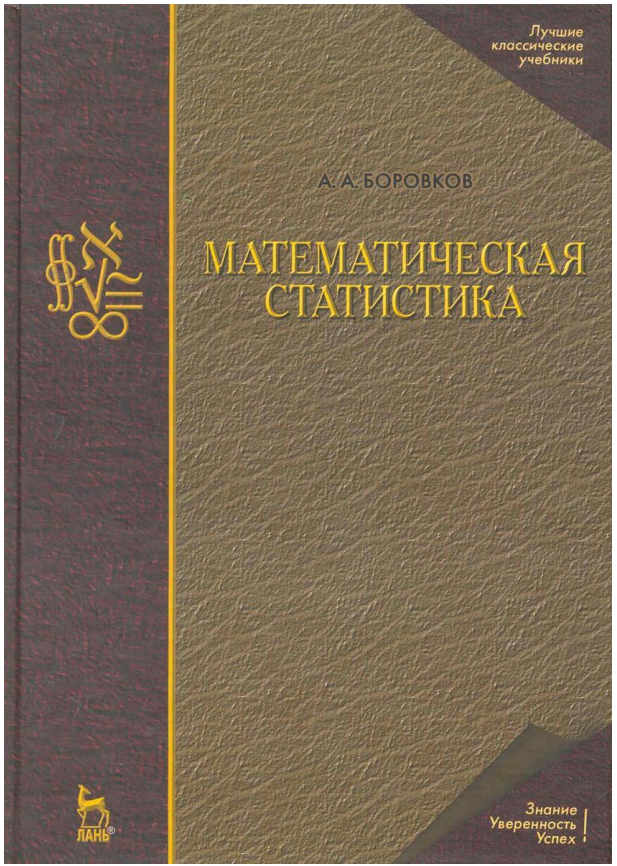
Что такое data science?

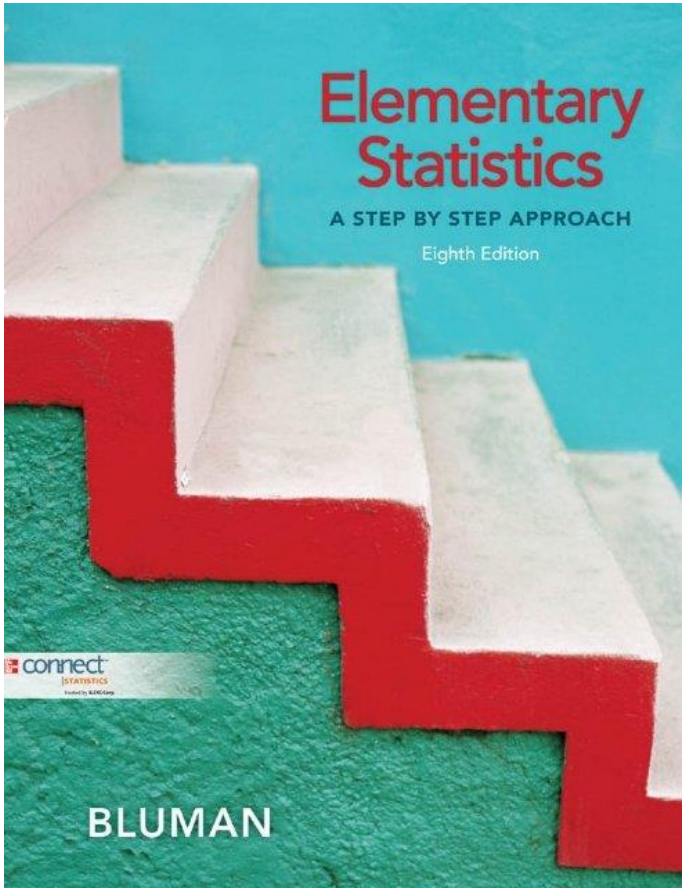
data science = (statistics + informatics + computing + communication + sociology + management)

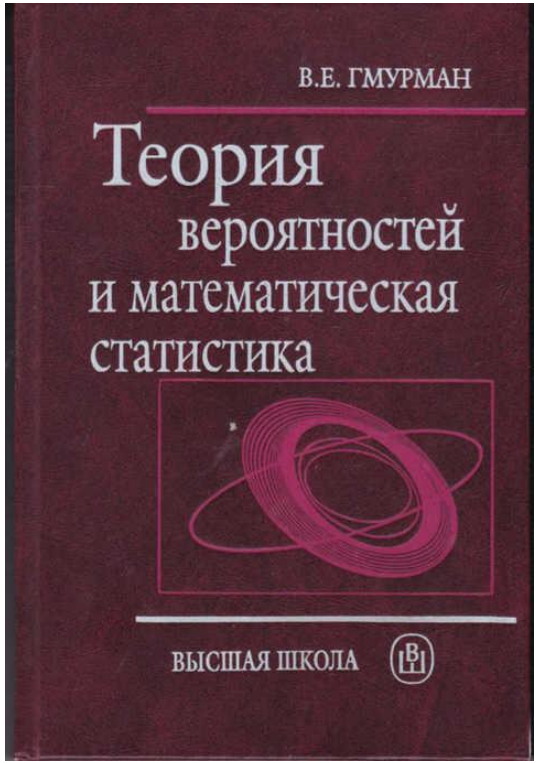

Владимир Савельев
Статистика
и котики
Самая дружелюбная книга
об анализе данных
При поддержке ЦИиР Юрия Корженевского


Чарльз Уилан
Голая
статистика
Самая
интересная
книга
о самой
скучной
науке


МАНГА ЗАНИМАТЕЛЬНАЯ
СТАТИСТИКА
Син Токехаси


Лучшие классические учебники
А. А. БОРОВКОВ
МАТЕМАТИЧЕСКАЯ СТАТИСТИКА
ЛАНЬ
Знание
Уверенность
Успех


Elementary Statistics
A STEP BY STEP APPROACH
Eighth Edition
connect
BLUMAN


В.Е. ГМУРМАН
Теория вероятностей и математическая статистика
ВЫСШАЯ ШКОЛА

Генеральная совокупность

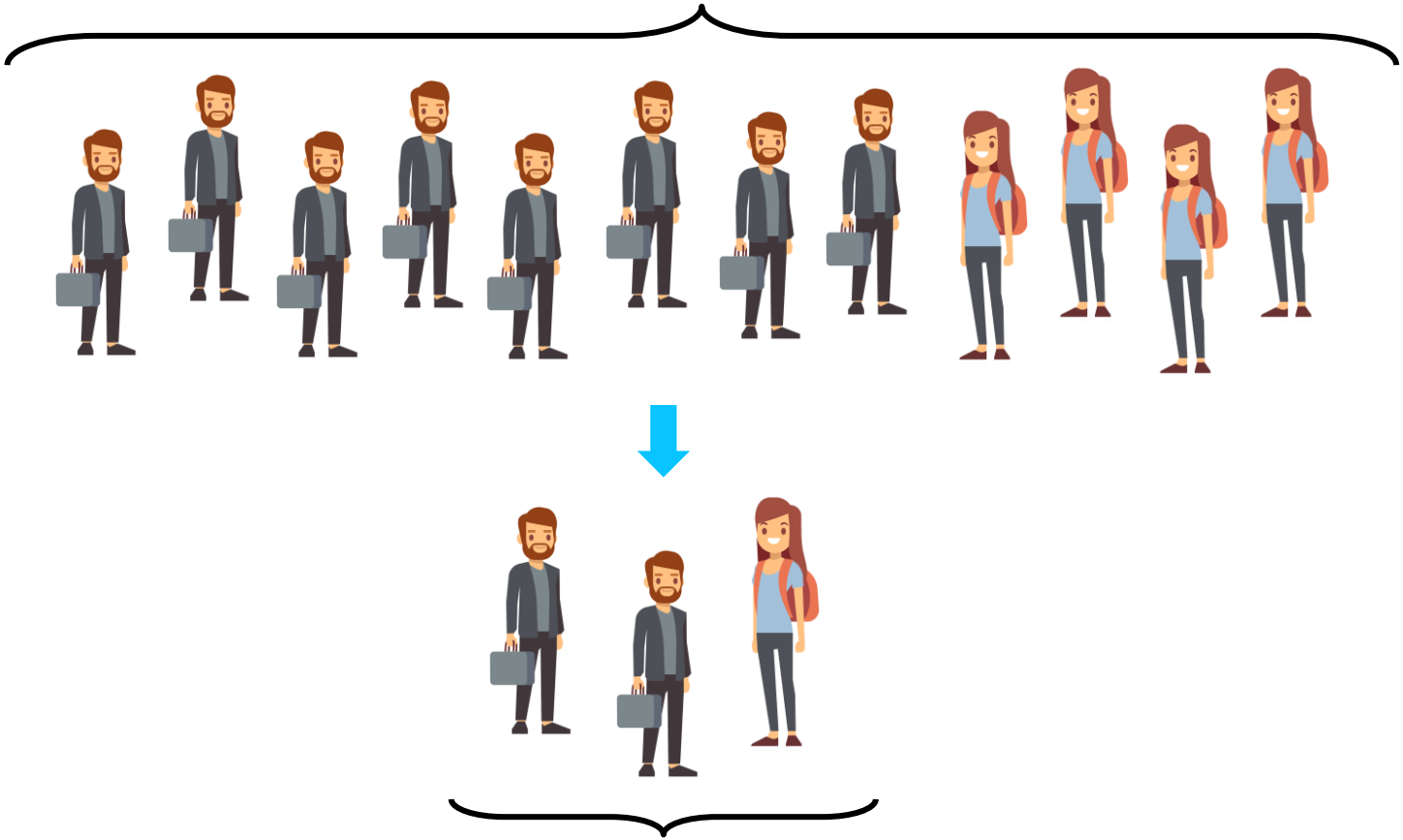
Генеральная совокупность — совокупность всех объектов, относительно которых предполагается делать выводы при изучении конкретной задачи

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

Генеральная совокупность и выборка

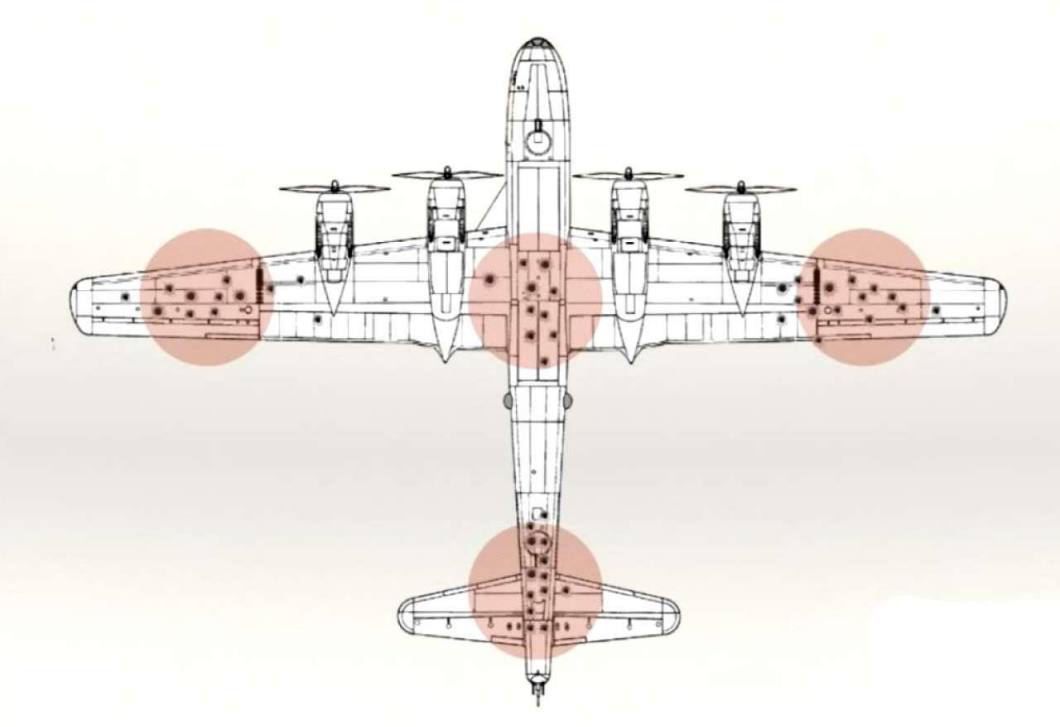


Репрезентативность выборки

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

Репрезентативность выборки

Ошибка выжившего



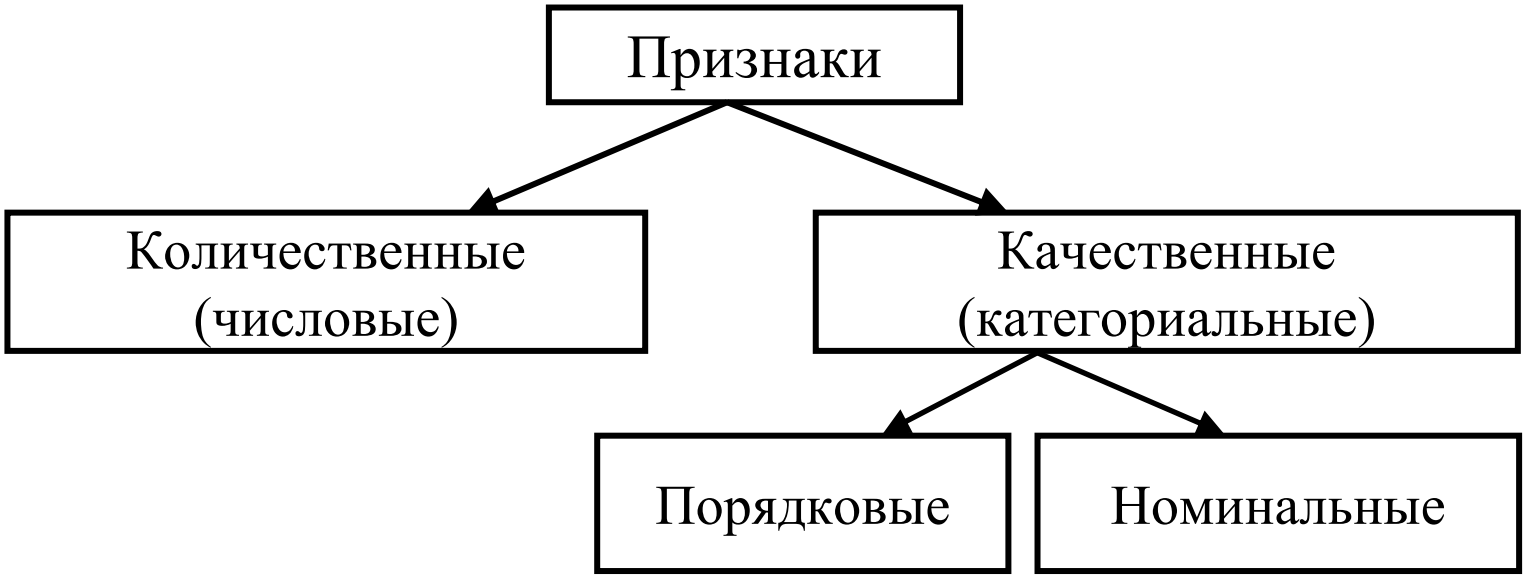
Ошибка выжившего

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

Меры центральной тенденции

- Среднее арифметическое
- Медиана
- Мода

Среднее арифметическое

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

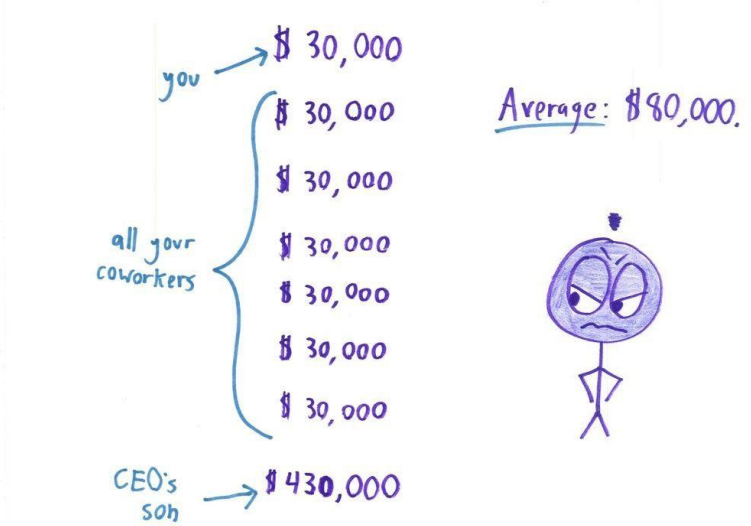
Пример: 1,2,6,6,7

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

*Иллюстрации взяты из книги «Статистика и котики»

Среднее арифметическое

Минус данной МЦТ: чувствительность к выбросам



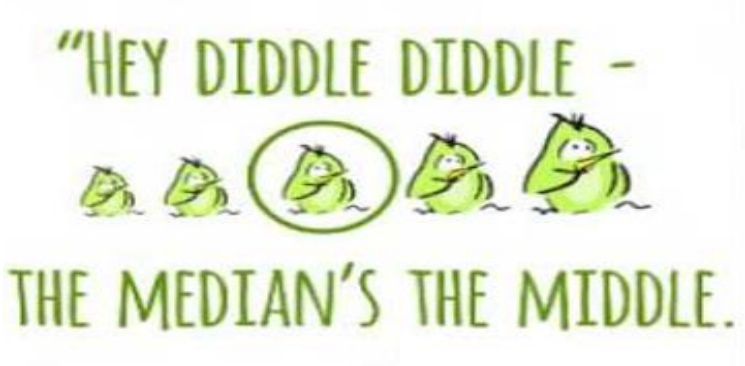
*Иллюстрации взяты из книги «Статистика и котики»

Усеченное среднее

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

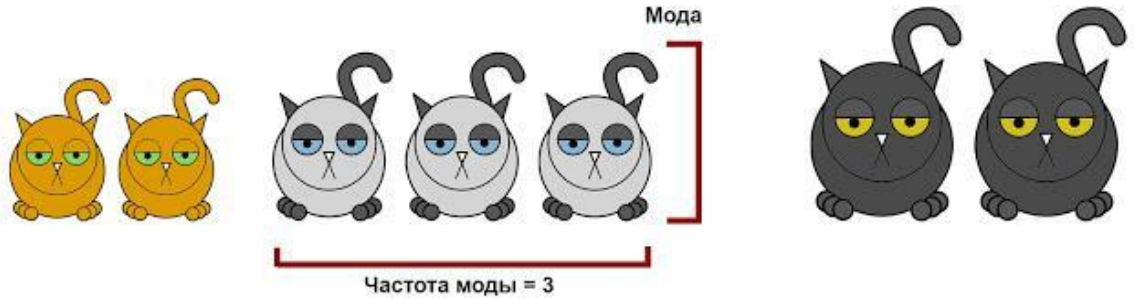
1. Дан числовой ряд: 1,5,3,9,11, 2, 14, 6
2. Расположим числа в порядке возрастания:

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



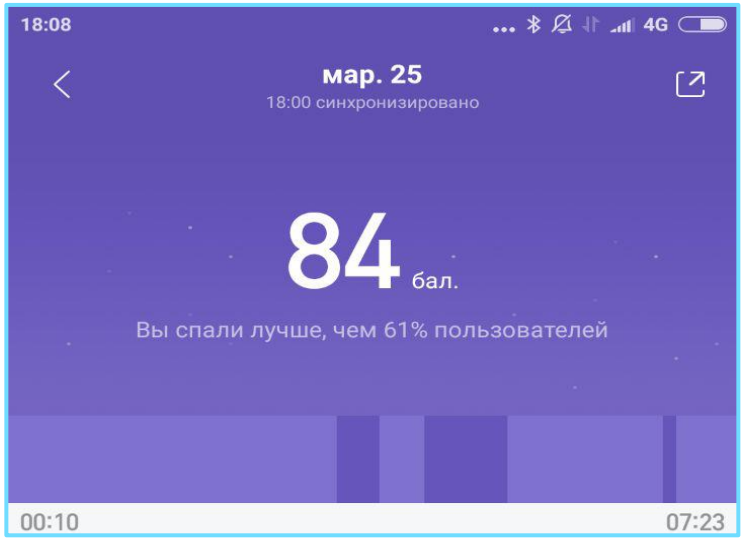
*Иллюстрации взяты из книги «Статистика и котики»

Меры разброса

- Размах
- Межквартильный размах
- Стандартное отклонение
- Дисперсия

Квантили и процентили

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



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

Стандартное отклонение

Дисперсия

Меры и типы признаков

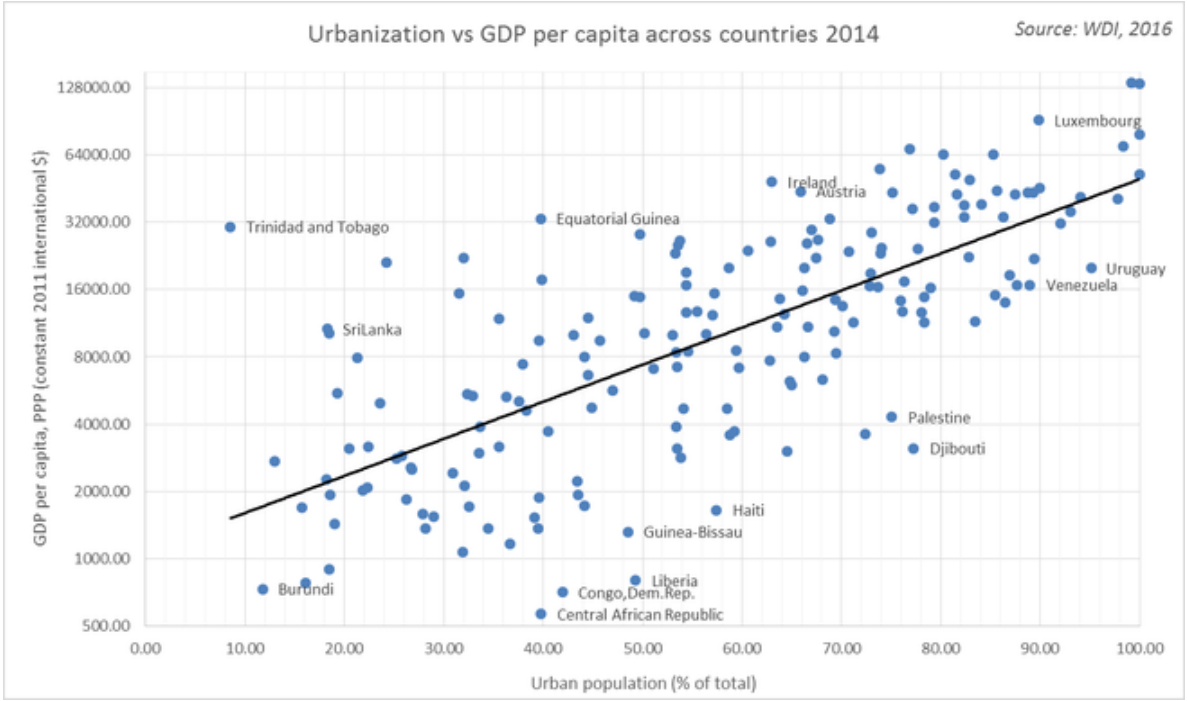
Типы данных	Меры центра			Меры разброса		
	Мода	Медиана	Среднее	Размах	Q-Q	Ст.Откл.
Номинальные	✓	✗	✗	✗	✗	✗
Порядковые	✓	✓	✗	✓	✓	✗
Количественные	✓	✓	✓	✓	✓	✓

Корреляция

Корреляция – мера взаимосвязи двух величин

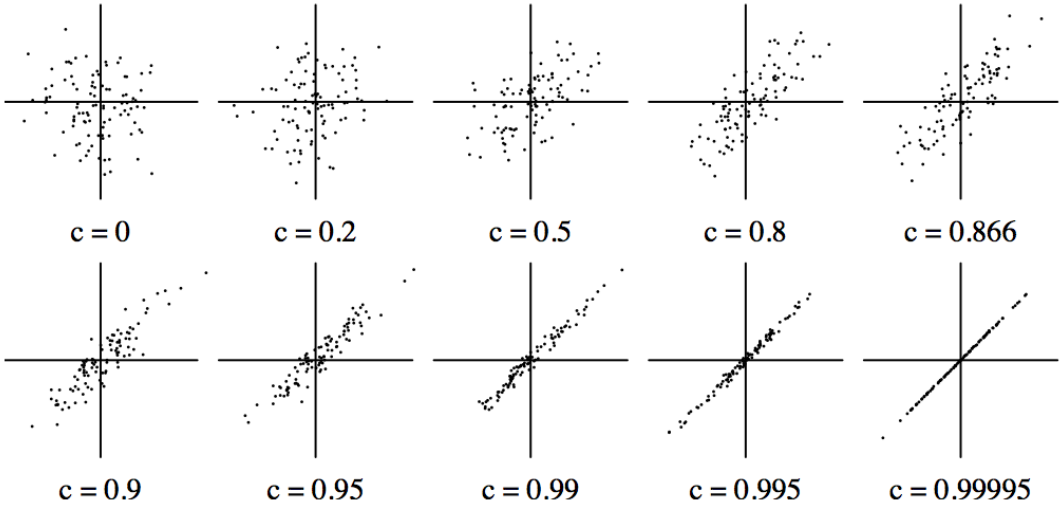
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

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

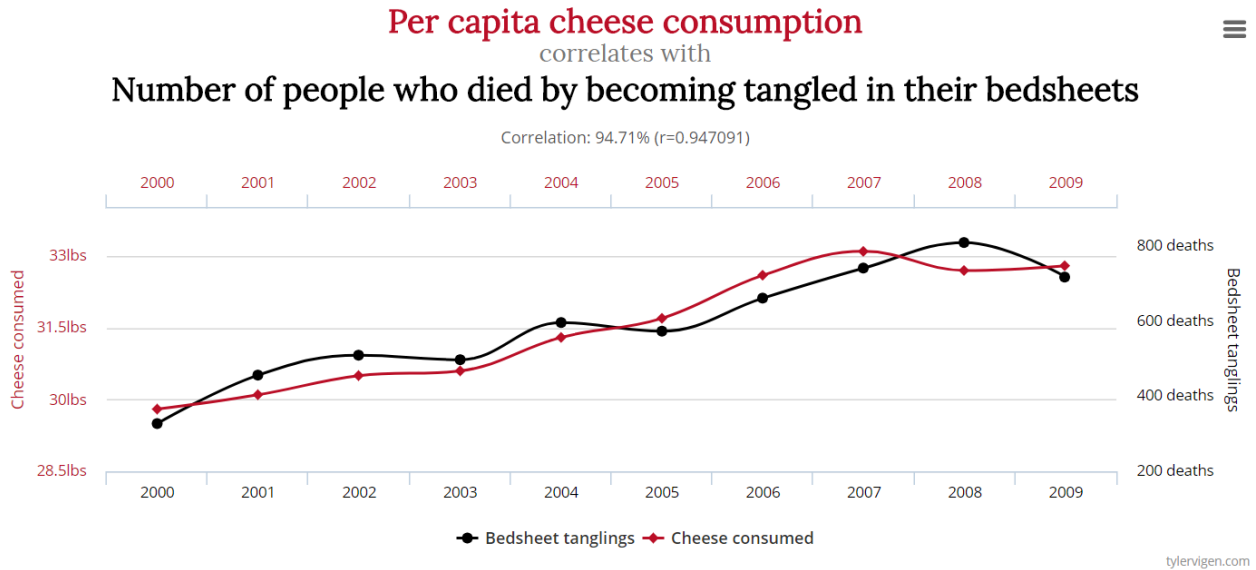


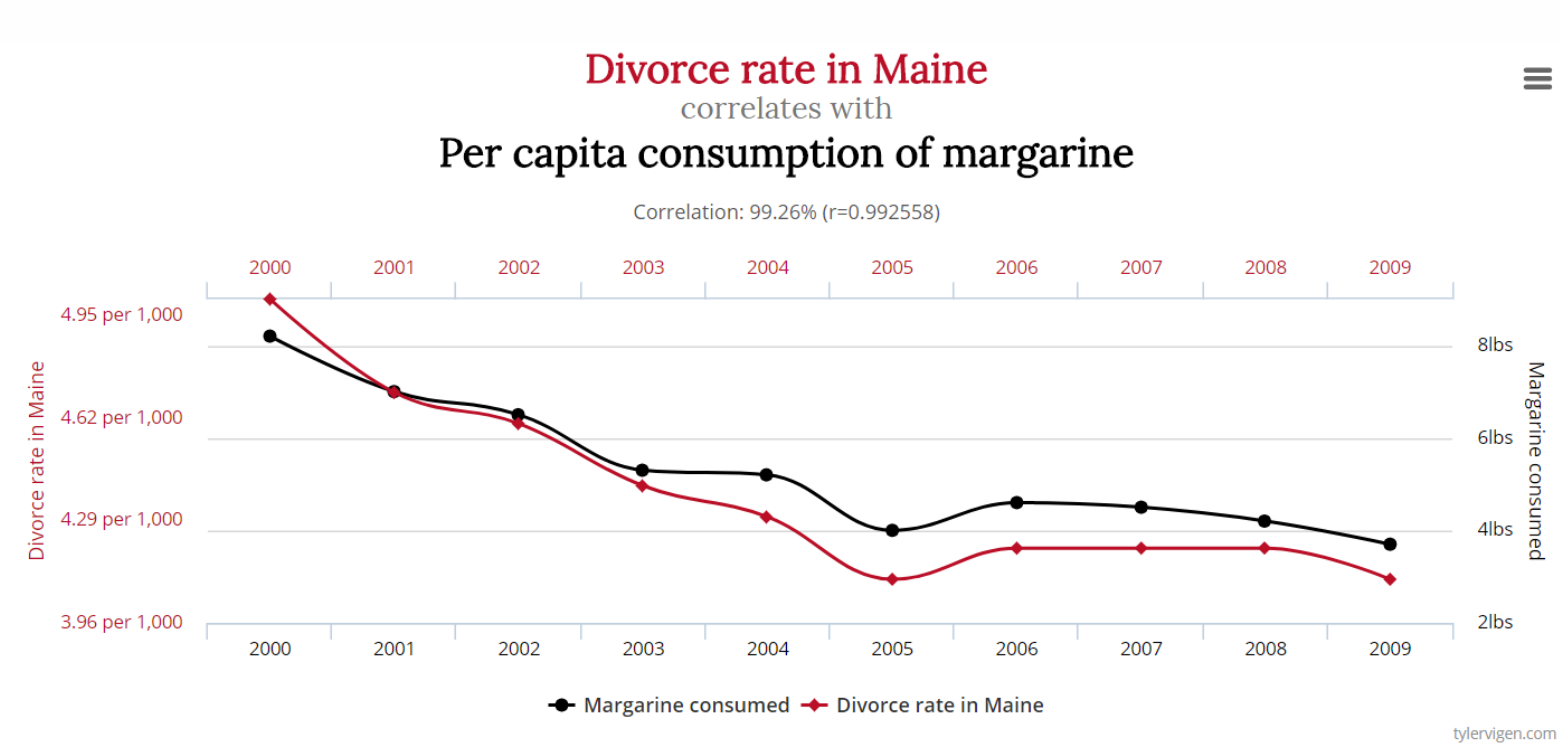
Корреляция

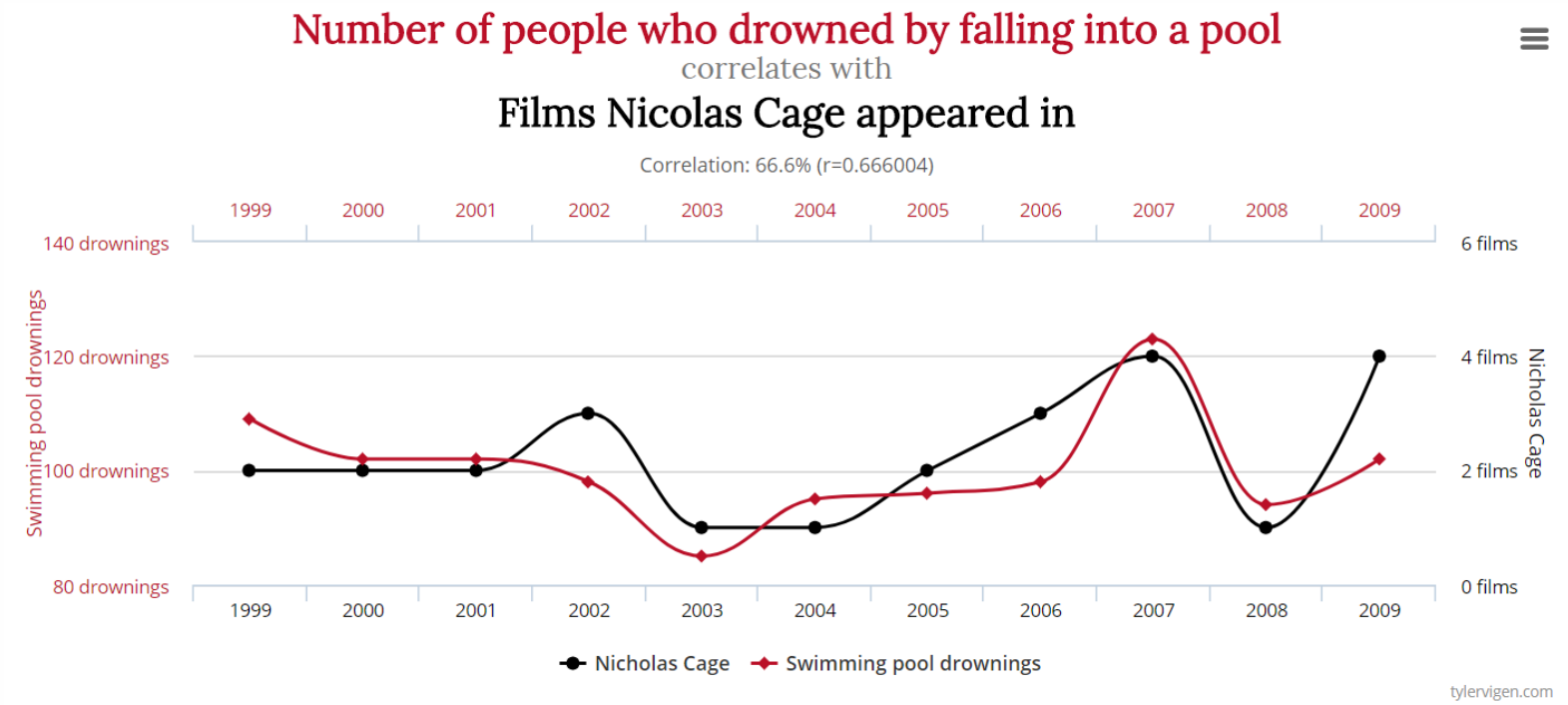
Свойства корреляции:

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

Корреляция

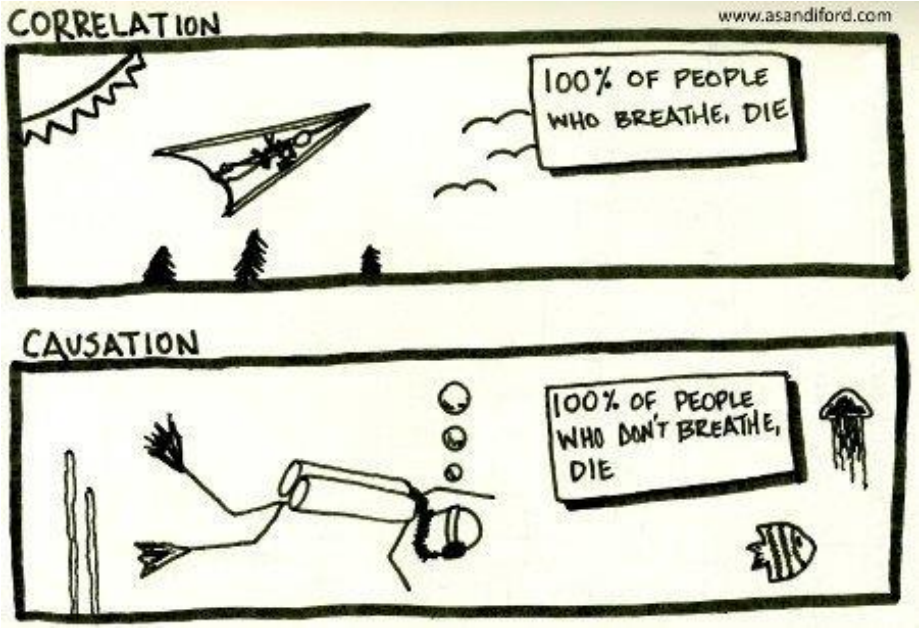

Per capita cheese consumption
correlates with
Number of people who died by becoming tangled in their bedsheets
Correlation: 94.71% (r=0.947091)
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
33lbs
31.5lbs
30lbs
28.5lbs
Cheese consumed
800 deaths
600 deaths
400 deaths
200 deaths
Bedsheet tanglings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


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



Data sources: Centers for Disease Control & Prevention and Internet Movie Database

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



Качество оценок

- Несмещённость означает, что в среднем мы оценим параметр верно.

- Состоятельность означает, что при увеличении размера выборки ошибка оценки уменьшается.

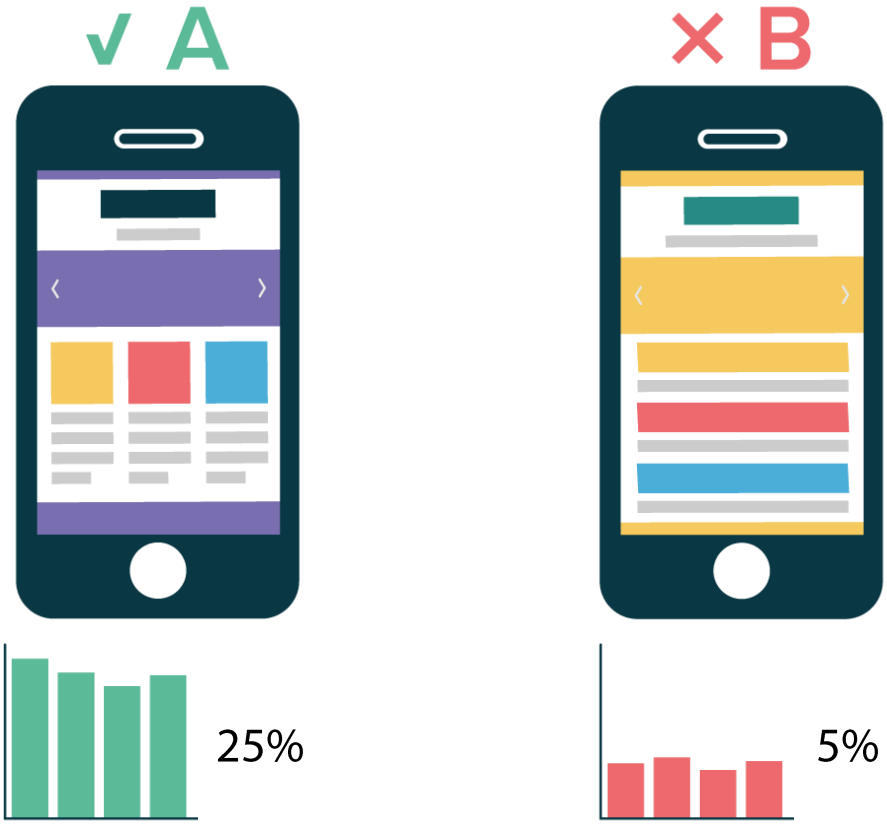
A/B тестирование

A/B-тестирование (англ. A/B testing, Split testing) — метод исследования, суть которого заключается в том, что контрольная группа элементов сравнивается с набором тестовых групп, в которых один или несколько показателей были изменены, для того, чтобы выяснить, какие из изменений улучшают целевой показатель

A/B тестирование

- Все пользователи сайта случайным образом разделяются на две группы: сегмент A и сегмент B
- Сегмент A является контрольным, пользователи из него видят ресурс в первоначальном виде
- Пользователям из сегмента B показывают измененный вариант
- Оценивается, есть ли статистически значимые различия в интересующей нас метрике между двумя сегментами

A/B тестирование



Статистическая гипотеза

Статистическая гипотеза — это любое утверждение о виде или свойствах распределения наблюдаемых в эксперименте случайных величин

Нулевая и альтернативная гипотеза

Существует два вида гипотез:

H_0 – нулевая гипотеза

H_A – альтернативная гипотеза

Нулевая и альтернативная гипотеза: пример

H_0 : Между средними оценками в двух группах студентов нет отличий

H_A : Средние оценки студентов в двух группах различны

Нулевая и альтернативная гипотеза: пример

H_0 : Коэффициент корреляции между временем, потраченным на наш курс, и уровнем счастья равен нулю

H_A : Коэффициент корреляции между временем, потраченным на наш курс, и уровнем счастья значимо отличается от нуля

Несимметричность гипотез

Нулевая и альтернативная гипотеза не равнозначны

Мы не можем доказать, что нулевая гипотеза является верной

Несимметричность гипотез

У нас есть 2 варианта:

1. Отвергаем нулевую гипотезу в пользу альтернативной
2. Не отвергаем нулевую гипотезу в пользу альтернативной

Ошибка I и II рода

	H0 верна	**H0 неверна**
H0 принимается	H0 принята верно	Ошибка II рода
H0 отвергается	Ошибка I рода	H0 отвергнута верно

Ошибка I и II рода

Ошибка I рода

Ошибка II рода

habr.com

P-значение

P-значение — величина, используемая при тестировании статистических гипотез. Фактически это вероятность ошибки при отклонении нулевой гипотезы (ошибки первого рода)

P-значение: пример

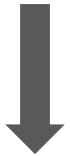
Пусть есть следующие гипотезы:

H_0 : Между средними оценками в двух группах студентов нет отличий

H_A :Средние оценки студентов в двух группах различны

P-значение: пример

Получили p-значение, равное 0.13



Если мы примем гипотезу о различных оценках, то ошибемся с вероятностью 0.13

Практическая и статистическая значимость

Статистическая значимость не подразумевает важность или практическую ценность

Нужно не только проверять справедливость статистической гипотезы, но и оценивать, насколько велик размер эффекта

Практическая и статистическая значимость

- Было проведено исследование, которое изучало влияние физических упражнений на набор веса

Практическая и статистическая значимость

- Было проведено исследование, которое изучало влияние физических упражнений на набор веса
- За три года женщины, упражнявшиеся не меньше часа в день, набрали значимо меньше веса, чем женщины, упражнявшиеся меньше 20 минут в день ($p < 0.001$)

Практическая и статистическая значимость

- Было проведено исследование, которое изучало влияние физических упражнений на набор веса
- За три года женщины, упражнявшиеся не меньше часа в день, набрали значимо меньше веса, чем женщины, упражнявшиеся меньше 20 минут в день ($p < 0.001$)
- Разница в набранном весе составила 150 г., что является совершенно незначительным эффектом с практической точки зрения!

Практическая и статистическая значимость

- В 2002 году клинические испытания гормонального препарата Премарин были досрочно прерваны, хотя статистически значимых побочных эффектов не было: было обнаружено, что его приём ведёт к значимому увеличению риска развития рака груди на 0.08%, риска инсульта на 0.08% и инфаркта на 0.07%

Практическая и статистическая значимость

- В 2002 году клинические испытания гормонального препарата Премарин были досрочно прерваны, хотя статистически значимых побочных эффектов не было: было обнаружено, что его приём ведёт к значимому увеличению риска развития рака груди на 0.08%, риска инсульта на 0.08% и инфаркта на 0.07%
- Несмотря на то что побочный эффект очень маленький, на практике он дает тысячи смертей

Понятие статистического критерия

Статистический критерий — это строгое математическое правило, по которому принимается или отвергается та или иная статистическая гипотеза с известным уровнем значимости

ВОПРОСЫ?