

# Курс по Математической статистике

**Даниил Потапов**

Руководитель Лаборатории Искусственного Интеллекта

РСХБ





## Что-то профессиональное:

- Стаж 11 лет, 4 компании
- GIS, Front-end, Full-stack
- Open source & open edu

## Что-то личное:

- 3,5к часов в CS 1.6
- Люблю научную фантастику
- Коллекционер книг и игр

## Что-то необычное:

- 10 лет занимался шахматами
- Арахнофоб
- Писал читы для игр

## Что-то неприличное:

- ругаюсь %#&
- КМС по литрболу
- Не умею готовить еду

# Команда курса

**ІІТМО**

**Даниил Потапов**

Рук-ль Лаборатории ИИ

PCXB



**Жукова Алина**

Head of Analytics

Novakid



**Юрий Котов**

Senior Data Engineer

Т-Банк



**Жигалов Августин**

Data Engineer

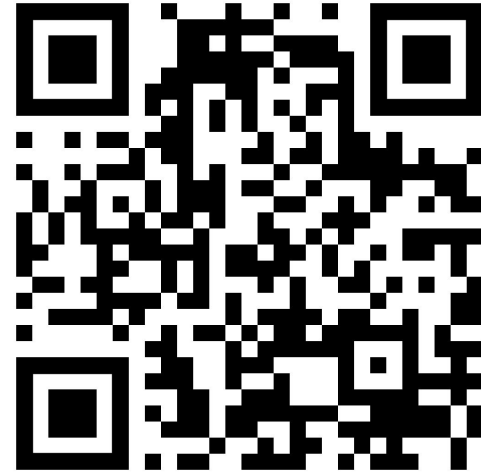
PCXB



- 6 лекцій
- 6 семінарів
- 4 домашки
- 18 часів на заняття
- 6 тижнів інтенсива

- 6 лекций
- 6 семинаров
- 4 домашки
- 18 часов на занятия
- 6 недель интенсива

- Занятия - понедельник и пятница
- Одна пара за раз, 18:40 - 20:10
- Телеграм чат



- Материалы курса будем выкладывать в Github (приватная репа)
  - [https://github.com/sharthZ23/ltmo\\_mathstat\\_2024](https://github.com/sharthZ23/ltmo_mathstat_2024)
- Надо будет позже внести свои Github ники
  - Если у вас нет аккаунта на Github, то стоит завести
- Там же мы закрепим ментора за каждым студентом
  - Тем не менее, не стесняйтесь задавать вопросы в общем чате
  - Будем начислять доп. баллы тем, кто помогает своим однокурсникам
- Об этом всем дополнительно еще в Telegram чате проинформируем
  - И заведем отдельный Google sheet





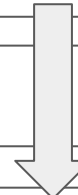
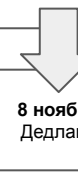
## 6 тем (лекция + семинар)

1. Введение в МатСтат
2. Статистические оценки
3. Доверительные интервалы
4. Параметрические критерии
5. Непараметрические критерии
6. МатСтат на службе у бизнеса

## Домашки

- 4 ДЗ, для тем с 2-ой по 5-ую
- 25 баллов за каждую
  - 100 баллов макс за курс
- Оценка за курс (экзамен)
  - 5 - от 91 баллов
  - 4 - от 74 до 90
  - 3 - от 60 до 73
- Зачет - оценка 3 и выше

# Расписание

	Понедельник	Пятница	ДЗ №1	ДЗ №2	ДЗ №3	ДЗ №4
Неделя 1	30 сентября Лекция №1	4 октября Семинар №1				
Неделя 2	7 октября Лекция №2	11 октября Семинар №2	7 октября Выдача			
Неделя 3	14 октября Лекция №3	18 октября Семинар №3		14 октября Выдача		
Неделя 4	21 октября Лекция №4	25 октября Семинар №4	21 октября Дедлайн		21 октября Выдача	
Неделя 5	28 октября Лекция №5	1 ноября Семинар №5		28 октября Дедлайн		28 октября Выдача
Неделя 6	Среда 6 ноября Лекция №6	8 ноября Семинар №6			4 ноября Дедлайн	 8 ноября Дедлайн

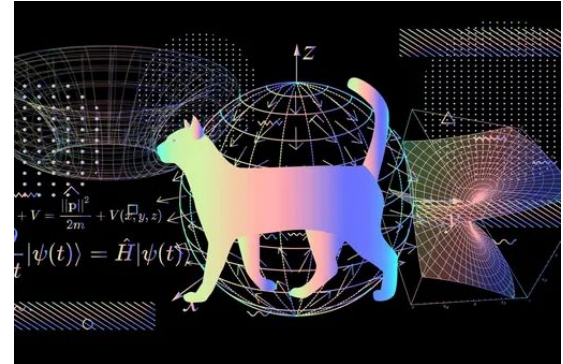
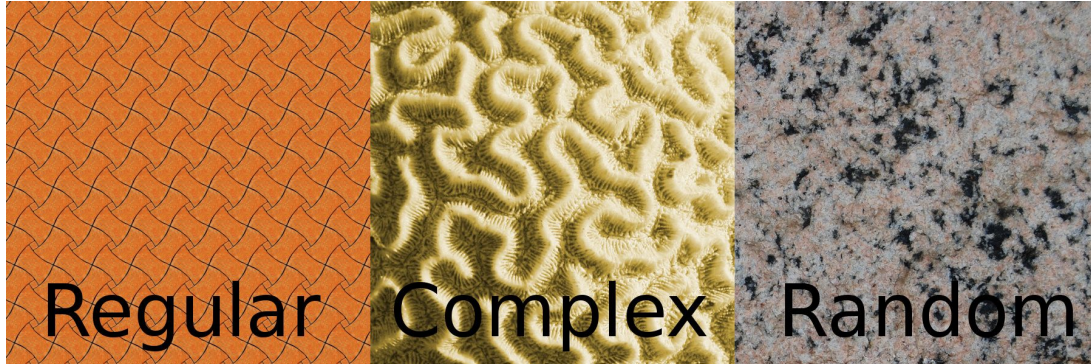


**Вопросы по орг. вопросам?**

- Про что ТерВер и МатСтат
- Основные термины ТерВера
  - Случайные величины
  - Характеристики случайных величин
- Основные термины МатСтата
  - Генеральная совокупность и выборка
  - Виды статистик
- Статистика на реальных данных
  - Независимые величины
  - Корреляция

- **Про что ТерВер и МатСтат**
- Основные термины ТерВера
  - Случайные величины
  - Характеристики случайных величин
- Основные термины МатСтата
  - Генеральная совокупность и выборка
  - Виды статистик
- Статистика на реальных данных
  - Независимые величины
  - Корреляция

# Случайность вокруг нас



# Демон Лапласа



Пьер-Симон де Лаплас (1775)

«Мы можем рассматривать настоящее состояние Вселенной как следствие его прошлого и причину его будущего. Разум, которому в каждый определённый момент времени были бы известны все силы, приводящие природу в движение, и положение всех тел, из которых она состоит, будь он также достаточно обширен, чтобы подвергнуть эти данные анализу, смог бы объять единым законом движение величайших тел Вселенной и мельчайшего атома; для такого разума ничего не было бы неясного и будущее существовало бы в его глазах точно так же, как прошлое»



# Два подхода к вероятности



Тóмас Бáйес  
1702, Лондон — 17 апреля 1761



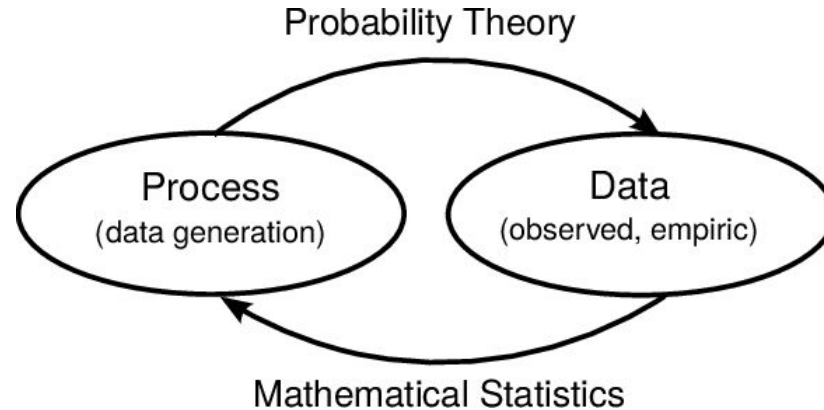
Рóналд Ёйлмер Фíшер  
17 февраля 1890 — 29 июля 1962)

- **Лаплас** развил *байесовские* идеи исходя из детерминизма
- Демон Лапласа - Точное предсказание вселенной в случае возможности измерения положения каждого атома, но издержки огромны (парадокс разрешим)
- Возникающая неопределенность – результат огромного разрыва между совершенством природы и несовершенством человеческого познания
- Таким образом, случайность – следствие нашей ограниченности
- **Вероятность – способ измерения случайности, причем субъективно**

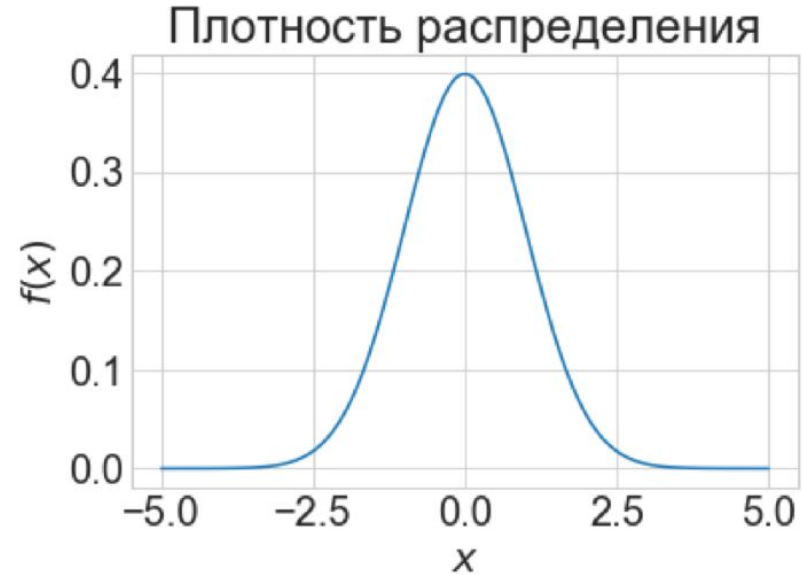
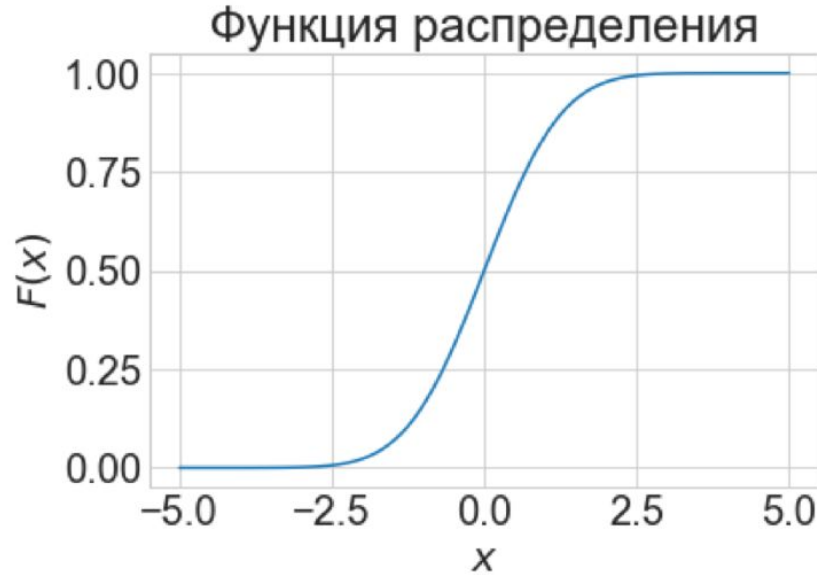
- **Фишер** считал, что наука не может рассматривать вероятность как нечто субъективное
- Можно оценивать вероятность только тех событий, которые происходят более одного раза
- Вопрос “Какова вероятность, что кандидат N победит на выборах?” не имеет ответа, так как событие уникально и не обладает частотой
- **Вероятность должна быть объективной**



- Мир вокруг нас порождает данные мириадами различных процессов. Механизмы порождения изучаются **теорией вероятностей**



- Наблюдаемые данные – объект изучения **математической статистики**. По выборкам из этих данных мы пытаемся понять, каким процессом они порождены



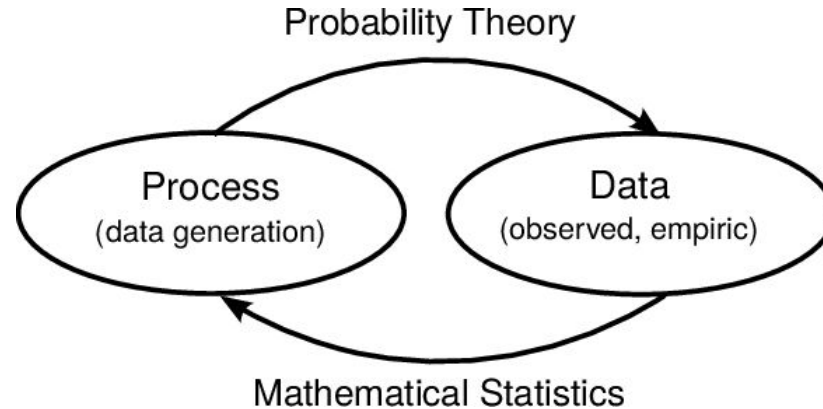
**Модель** – наше предположение о том, как устроен “мир”, то есть какие есть процессы и какие данные они порождают. Все это базируется на наших знаниях и предположениях

- “Мир” порождает данные неизвестным нам механизмом
- Мы изучаем данные и их свойства
- На основе данных пытаемся восстановить структуру механизма
- Фиксация своих предположений и гипотез в виде моделей
- Восстановление структуры механизма на основе выбранной модели
- Проверка корректности нашей модели на имеющихся данных

- Про что ТерВер и МатСтат
- **Основные термины ТерВера**
  - Случайные величины
  - Характеристики случайных величин
- Основные термины МатСтата
  - Генеральная совокупность и выборка
  - Виды статистик
- Статистика на реальных данных
  - Независимые величины
  - Корреляция

**Случайная величина  $X$**  – произвольная измеримая функция, заданная на пространстве элементарных событий  $\Omega$  и принимающая значения в  $\mathbb{R}$

Это означает, что каждому элементарному событию  $w$  мы будем ставить в соответствие некоторое число  $X(w)$



# Случайная величина

Дискретная - множество значений конечно или счётно

- Значение игральной кости
- Число звонков в КЦ
- Число кликов

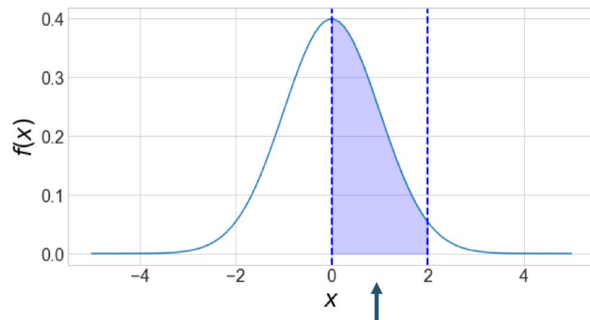
Непрерывная - бесконечное число значений

- Вес
- Рост
- Зарплата
- Время

Распределение задается таблицей

$x_i$	1	2	3	4
$p_i$	$\frac{1}{16}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{16}$

Распределение задается функцией плотности

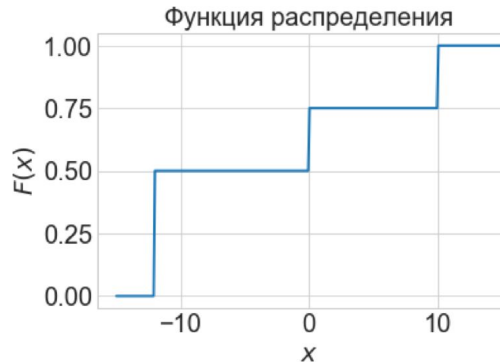


Площадь равна вероятности попасть на отрезок от 0 до 2

**Функция распределения** – функция, которая определяет вероятность события  $X \leq x$ , то есть

$$F(x) = \mathbb{P}(X \leq x) = \sum \mathbb{P}(X = k) \cdot [X \leq x],$$

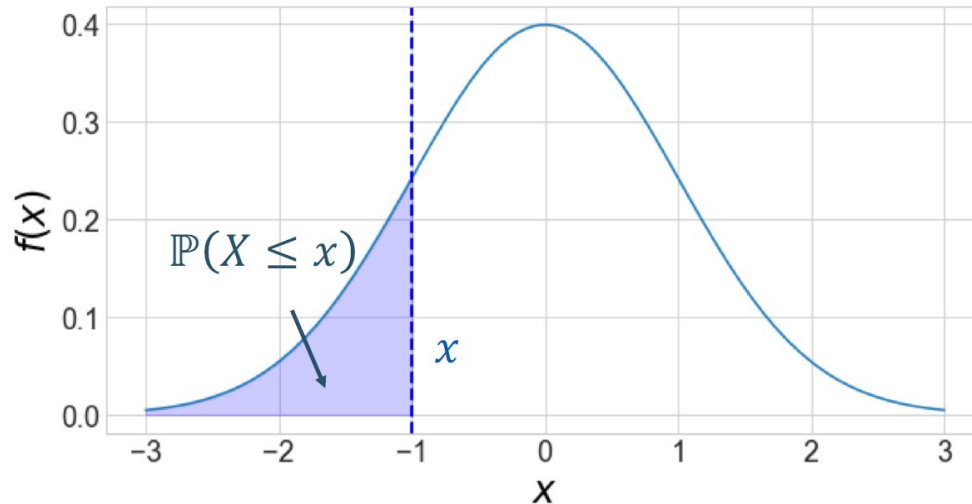
$$[X \leq x] = \begin{cases} 1, & X \leq x \\ 0, & \text{иначе} \end{cases}$$



$X$	$-12$	$0$	$10$
$\mathbb{P}(X = k)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$

**Функция распределения** – функция, которая определяет вероятность события  $X \leq x$ , то есть

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^x f(t) dt, f(t) - \text{плотность}$$





**Математическое ожидание** – среднее значение случайной величины

$$\mathbb{E}(X) = \sum_{k=1}^n k \cdot \mathbb{P}(X = k) \qquad \mathbb{E}(X) = \int_{-\infty}^{+\infty} t \cdot f(t) dt$$

**Дисперсия** – мера разброса случайной величины вокруг её среднего

$$Var(X) = \mathbb{E}(X - \mathbb{E}(X))^2 = \sum_{k=1}^n (k - \mathbb{E}(X))^2 \cdot \mathbb{P}(X = k)$$

$$Var(X) = \mathbb{E}(X - \mathbb{E}(X))^2 = \int_{-\infty}^{+\infty} (t - \mathbb{E}(X))^2 \cdot f(t) dt$$

# Дисперсия и среднее квадратическое отклонение

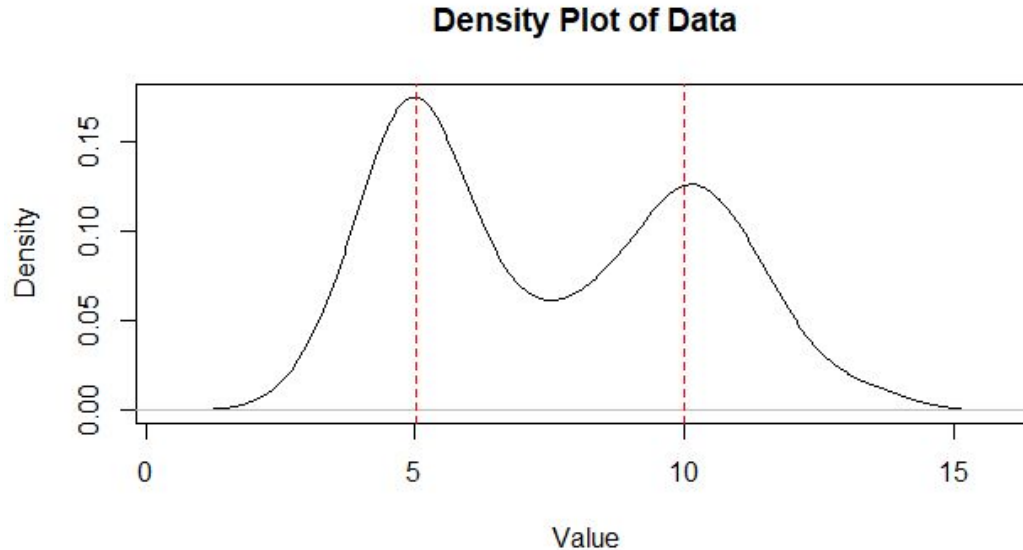
Более простая формула для дисперсии

$$\begin{aligned} \text{Var}(X) &= \mathbb{E}(X - \mathbb{E}(X))^2 \\ &= \mathbb{E}(X^2 - 2 \cdot X \cdot \mathbb{E}(X) + \mathbb{E}^2(X)) \\ &= \mathbb{E}(X^2) - 2 \cdot \mathbb{E}(X) \cdot \mathbb{E}(\mathbb{E}(X)) + \mathbb{E}^2(X) \\ &= \mathbb{E}(X^2) - 2 \cdot \mathbb{E}(X) \cdot \mathbb{E}(X) + \mathbb{E}^2(X) \\ &= \mathbb{E}(X^2) - \mathbb{E}^2(X) \end{aligned}$$

**Среднее квадратическое отклонение** - корень от дисперсии (чтобы убрать “квадрат”)

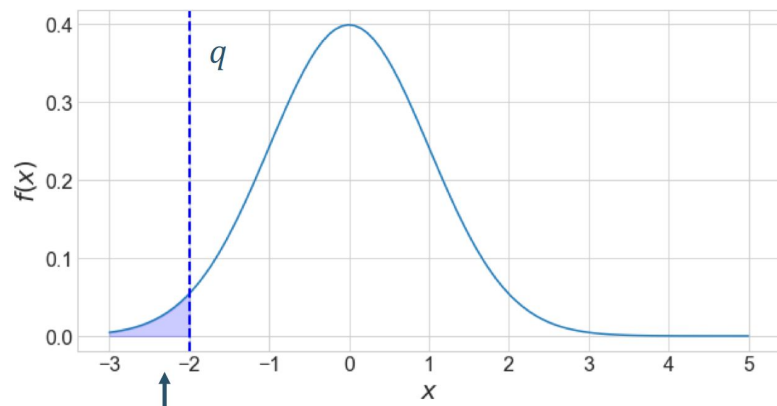
$$\sigma(X) = \sqrt{\text{Var}(X)}$$

**Мода** – значение, которому соответствует наибольшая вероятность (для дискретной случайной величины) или локальный максимум плотности распределения (для непрерывной)



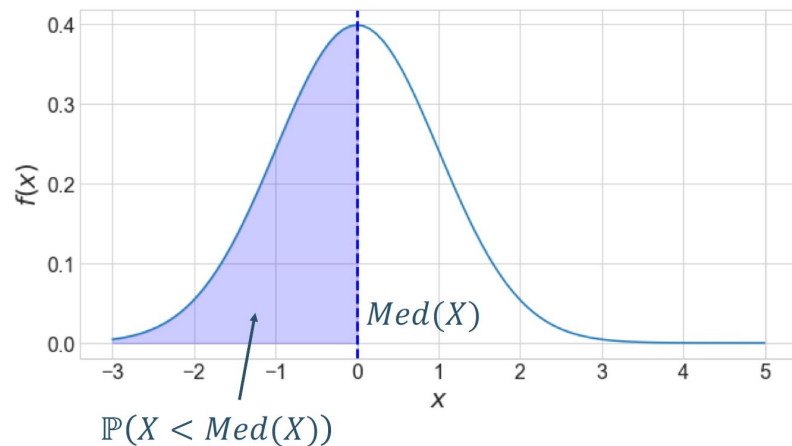
**Квантиль уровня  $\gamma$**  – это такое число  $q$ , что

$$\mathbb{P}(X \leq q) = \gamma$$



**Медиана** – это квантиль-0.5

$$\mathbb{P}(X < Med(X)) = \mathbb{P}(X > Med(X)) = 0.5$$

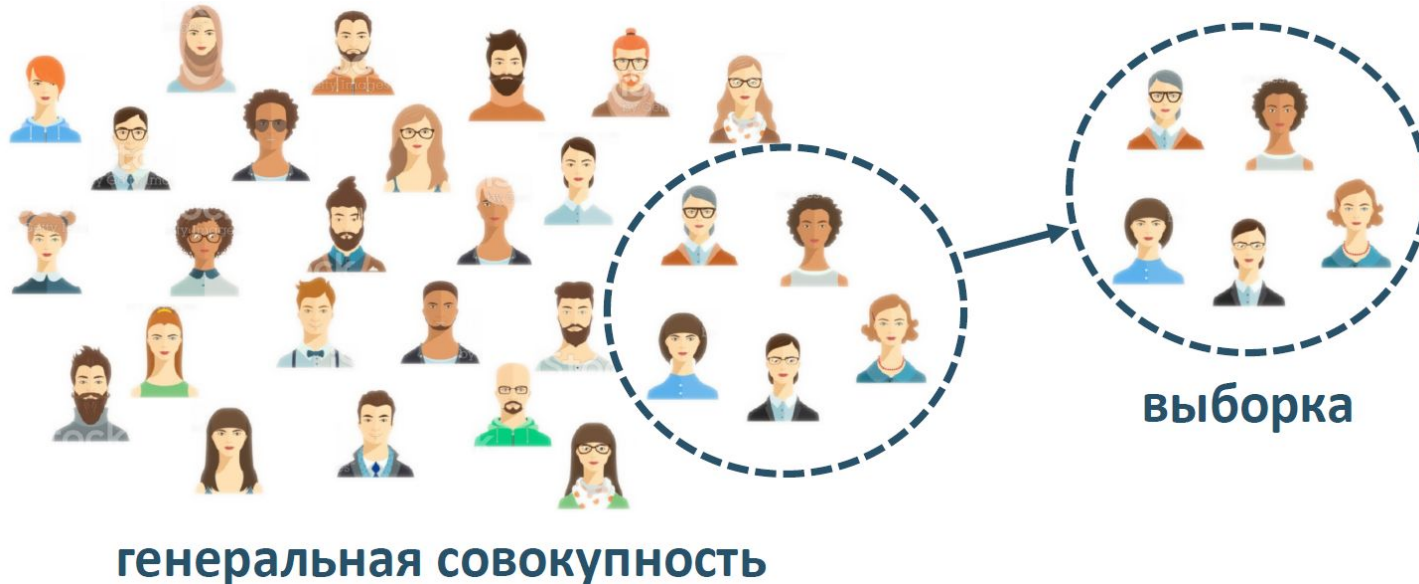


- Про что ТерВер и МатСтат
- Основные термины ТерВера
  - Случайные величины
  - Характеристики случайных величин
- **Основные термины МатСтата**
  - Генеральная совокупность и выборка
  - Виды статистик
- Статистика на реальных данных
  - Независимые величины
  - Корреляция

# Генеральная совокупность и выборка

**Генеральная совокупность** – это все объекты, которые нас интересуют при исследовании

**Выборка** – это та часть генеральной совокупности, по которой мы собрали данные для исследования



- Выборки позволяют сделать выводы о всей генеральной совокупности
- Чтобы выводы были корректными, выборка должны быть репрезентативной
- Репрезентативная выборка – отражает свойства генеральной совокупности

Вы хотите исследовать средний рост в своем городе. Как будете формировать свою выборку?

- Опросить своих друзей
- Опросить людей на автобусной остановке
- Опросить своих знакомых из спортивного кружка

- Выборки позволяют сделать выводы о всей генеральной совокупности
- Чтобы выводы были корректными, выборка должны быть репрезентативной
- Репрезентативная выборка – отражает свойства генеральной совокупности

Вы хотите исследовать средний рост в своем городе. Как будете формировать свою выборку?

☐ Нет ☐ Опросить своих друзей

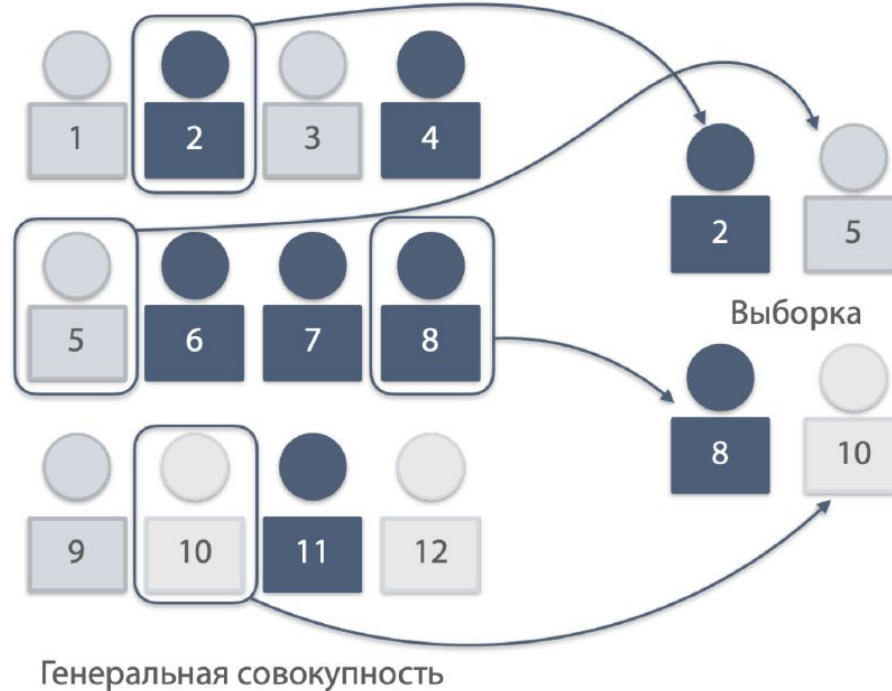
☒ Да ☐ Опросить людей на автобусной остановке

☐ Нет ☐ Опросить своих знакомых из спортивного кружка



# Получение репрезентативности

Один из способов достижения репрезентативности случайный отбор наблюдений



Выборка размера  $n$  –  $X_1, X_2, \dots, X_n \sim iid$

Каждое наблюдение можно рассматривать как случайную величину, которая имеет такое же распределение как и генеральная совокупность

Базовые предположения:

- Наблюдения независимы друг от друга
- Наблюдения имеют одинаковое распределение (как у генеральной совокупности)

*iid* расшифровывается как *identically independently distributed* (независимы и одинаково распределены)

Выборка:  $X_1, X_2, \dots, X_n \sim iid$

**Статистика** – функция от наблюдений (среднее, медиана, максмин и тд)

Каждая статистика – случайная величина, так как она вычисляется на основе случайной выборки, т.е. на основе других случайных величин

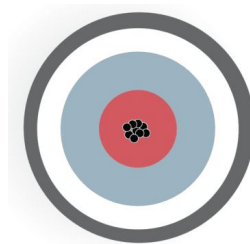


# Виды статистик

**Меры центральной тенденции** - это числа, которые могут описать множество значений в наборе данных одним числом.

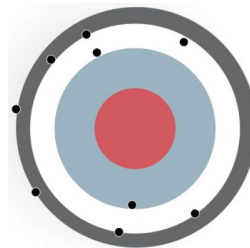
Отвечают на вопрос “На что похожие типичные данные для выборки”

Примеры: *среднее, медиана, мода*



**Меры разброса** отвечают на вопрос “Как сильно данные могут отличаться от типичных для этой выборки”

Примеры: *дисперсия, размах, отклонение*



# Среднее и медиана

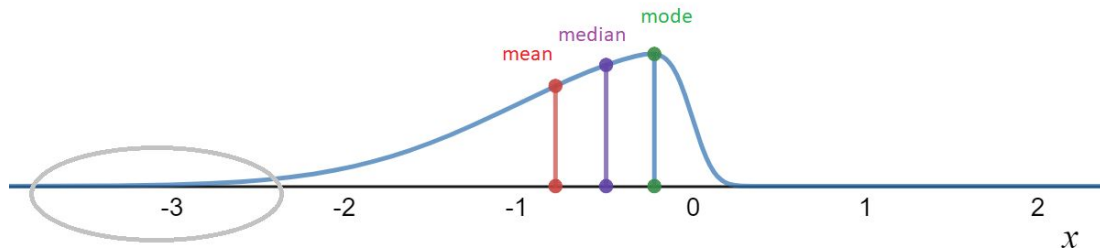
Все статистики - аналоги характеристик случайных величин, подсчитанных на выборке (иногда называют выборочными)

**Среднее**  $\bar{x} = \frac{X_1 + \dots + X_n}{n} = \frac{1}{n} \sum_{i=1}^n X_i$

Для подсчета **медианы** надо отсортировать выборку и взять середину (или среднее двух чисел в середине, если выборка четного размера)

## Свойства

- Среднее и медиана отражают типичное или ожидаемое значение
- Среднее чувствительно к выбросам в данных, медиана нет
- Соответственно, если в выборке нет выбросов, они примерно совпадают



# Выборочная дисперсия и отклонение

## Выборочная дисперсия

$$\hat{\sigma}^2 = \frac{(X_1 - \bar{x})^2 + \dots (X_n - \bar{x})^2}{n} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{x})^2 \quad \hat{\sigma}^2 = \overline{x^2} - \bar{x}^2$$

## Стандартное отклонение

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} \quad \text{лет} = \sqrt{\text{лет в квадрате}}$$

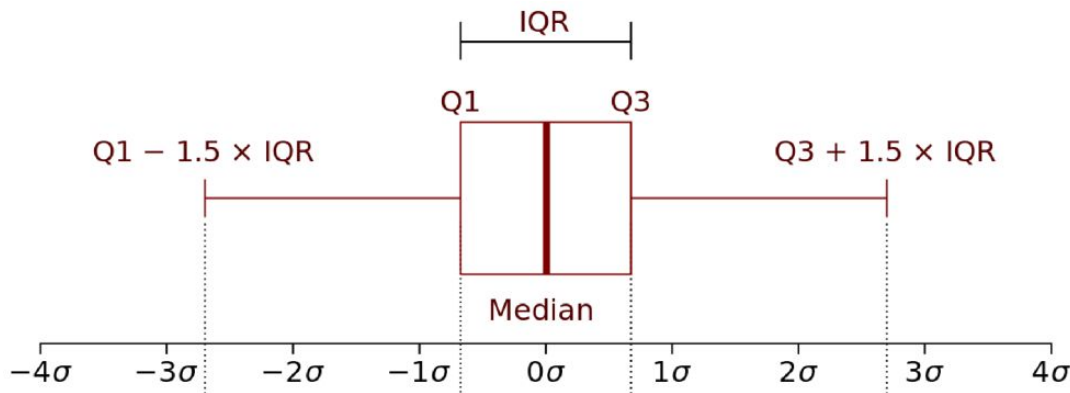
## Несмещенная выборочная дисперсия

$$s^2 = \frac{(X_1 - \bar{x})^2 + \dots (X_n - \bar{x})^2}{n - 1} = \frac{1}{n - 1} \sum_{i=1}^n (X_i - \bar{x})^2$$

# Перцентиль

Перцентиль порядка  $k$  – это такое число, что  $k\%$  выборки меньше этого числа

- Проще всего вычислять его по упорядоченной выборке  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$
- Квартили – перцентили с шагом в 0.25  $x_{(0.25 \cdot [n+1])}$   $x_{(0.5 \cdot [n+1])}$   $x_{(0.75 \cdot [n+1])}$
- Интерквартильный размах  $IQR = x_{(0.75 \cdot [n+1])} - x_{(0.25 \cdot [n+1])}$



- Про что ТерВер и МатСтат
- Основные термины ТерВера
  - Случайные величины
  - Характеристики случайных величин
- Основные термины МатСтата
  - Генеральная совокупность и выборка
  - Виды статистик
- **Статистика на реальных данных**
  - Независимые величины
  - Корреляция

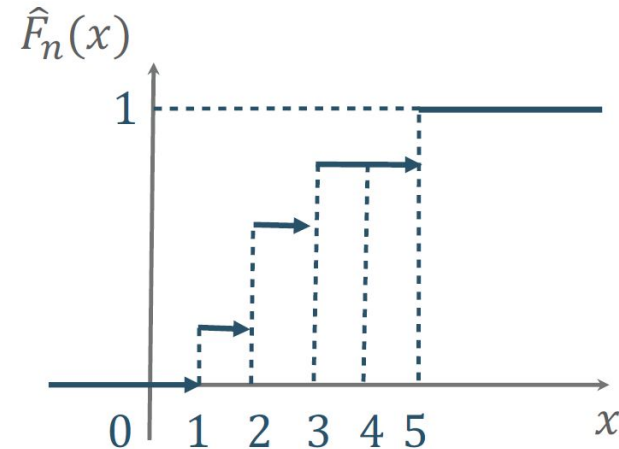


**Функция распределения** – функция, которая определяет вероятность события  $X \leq x$

**Эмпирическая функция распределения** – функция, которая определяет для каждого  $x$  частоту события  $X \leq x$

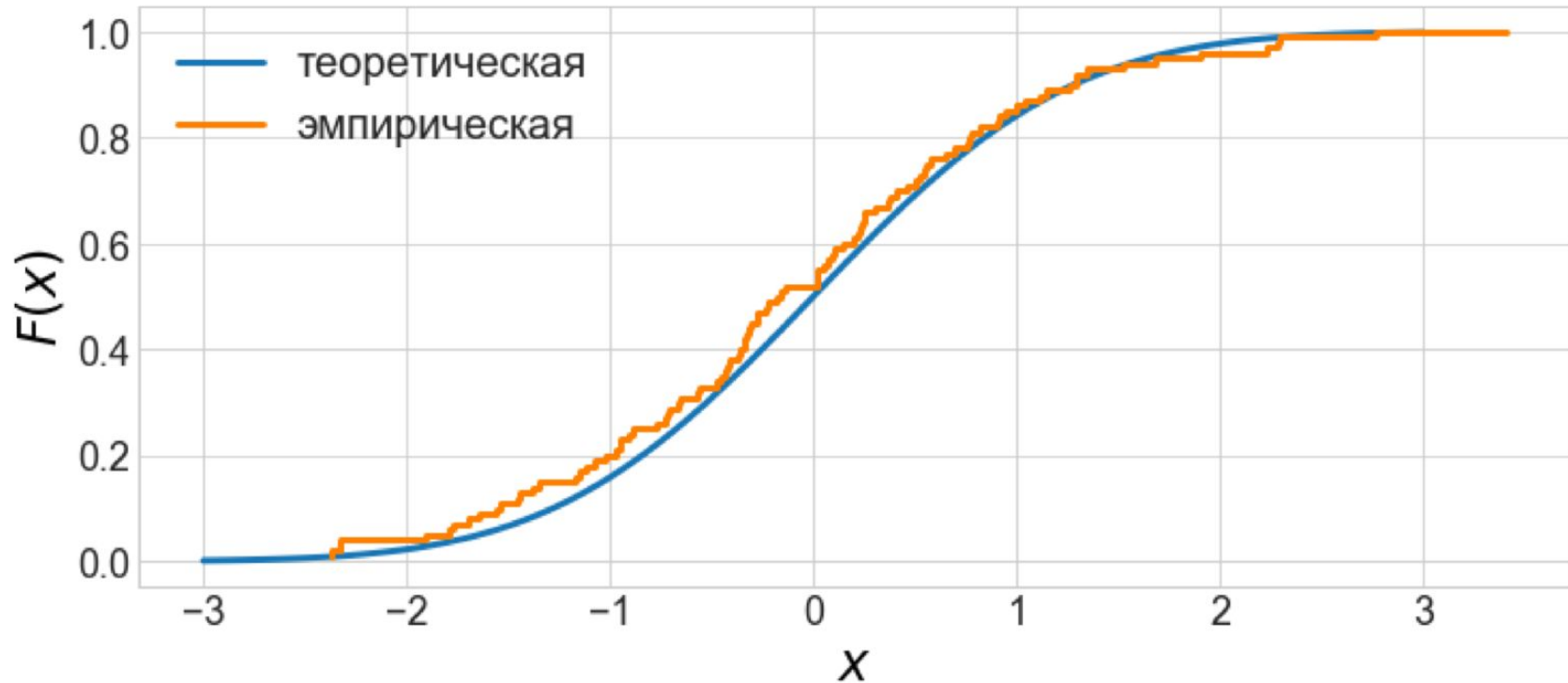
$$\hat{F}_n(x) = \hat{\mathbb{P}}(X \leq x) = \frac{1}{n} \sum_{i=1}^n [X_i \leq x]$$

$[ ]$  - индикаторная функция  $[X_i \leq x] = \begin{cases} 1, & X_i \leq x \\ 0, & \text{иначе} \end{cases}$

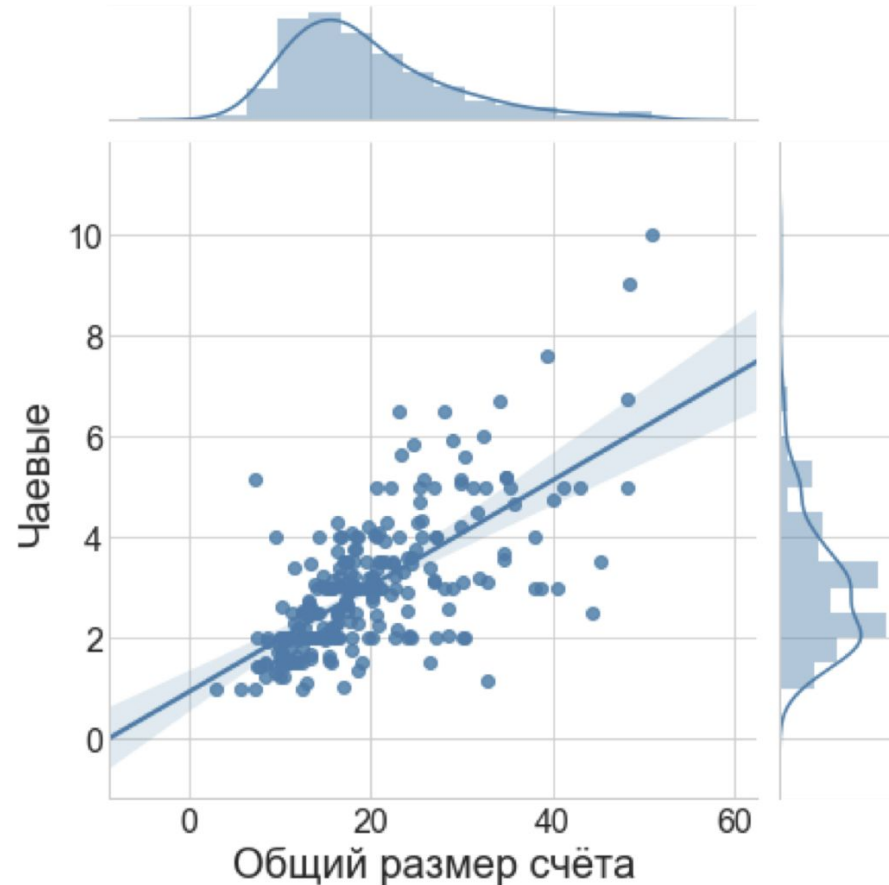


# Эмпирическая функция распределения

Чем больше выборка, тем чаще ступеньки и тем больше эмпирическая функция распределения похожа на теоретическую (чем больше данных - тем лучше)



- Случайные величины часто взаимосвязаны между собой
- Нужен какой-то способ измерять взаимосвязь между ними



**Независимость** заключается в том, что события не связаны, а значит их вероятности не влияют друг на друга

Говорят, что события  $A$  и  $B$  независимы, если

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$$

Говорят, что случайные величины  $X$  и  $Y$  независимы, если

$$F(x, y) = \mathbb{P}(X \leq x, Y \leq y) =$$

$$\mathbb{P}(X \leq x) \cdot \mathbb{P}(Y \leq y) = F_X(x) \cdot F_Y(y)$$

Тоже самое для плотностей

$$f(x, y) = f_X(x) \cdot f_Y(y)$$

**Ковариация** – это мера, показывающая степень совместной изменчивости двух случайных величин.

$$Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}(X)) \cdot (Y - \mathbb{E}(Y))]$$

## Свойства

- Положительная ковариация указывает на то, что величины имеют тенденцию изменяться в одном направлении (одна увеличивается, и другая тоже увеличивается), тогда как отрицательная ковариация указывает на противоположное изменение
- Если случайные величины независимы, то их ковариация равна 0
- Важно! Обратное неверно, если ковариация равна 0, то величины могут быть зависимыми
- Если величины  $X$  и  $Y$  зависимы, то

$$\mathbb{E}(X \cdot Y) = \mathbb{E}(X) \cdot \mathbb{E}(Y) + Cov(X, Y)$$

$$Var(X + Y) = Var(X) + Var(Y) + 2 \cdot Cov(X, Y)$$

Ковариация имеет размерность равную произведению размерностей случайных величин.  
Если  $X$  – деньги,  $Y$  – вес, ковариация измеряется в  
деньги · вес

Это неудобно  $\Rightarrow$  вводится безразмерный коэффициент корреляции:

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sigma(X) \cdot \sigma(Y)}$$

Коэффициент корреляции характеризует тесноту и направленность линейной связи между случайными величинами и принимает значение от  $-1$  до  $1$

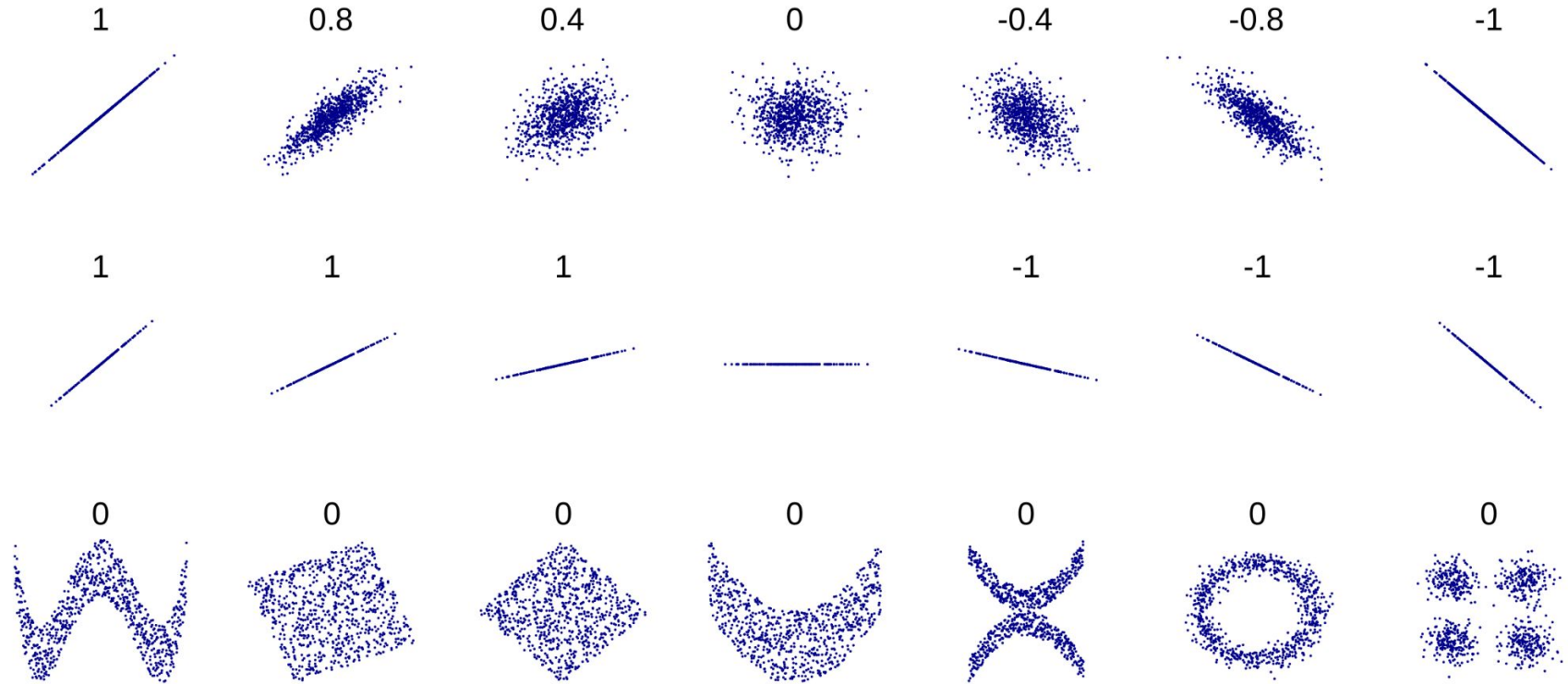
Выборочные аналоги:

(корреляция Пирсона)

$$\widehat{Cov}(X, Y) = \overline{xy} - \bar{x} \cdot \bar{y} = \frac{1}{n} \sum_{i=1}^n x_i \cdot y_i - \left( \frac{1}{n} \sum_{i=1}^n x_i \right) \cdot \left( \frac{1}{n} \sum_{i=1}^n y_i \right)$$

$$\hat{\rho}(X, Y) = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\hat{\sigma}_x \cdot \hat{\sigma}_y}$$

# Корреляция Пирсона

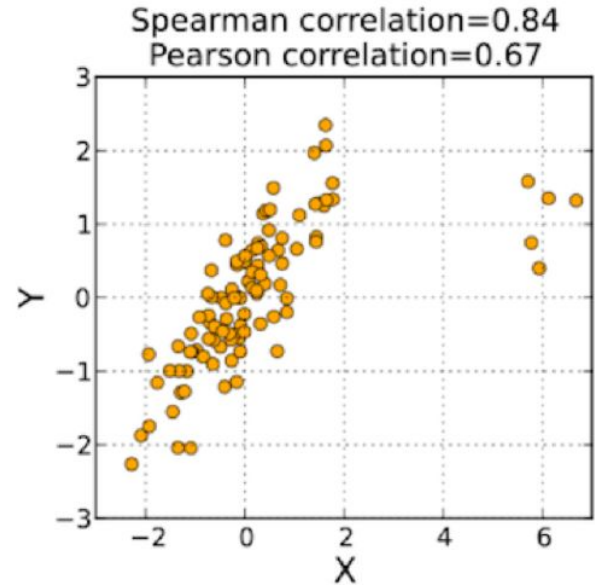
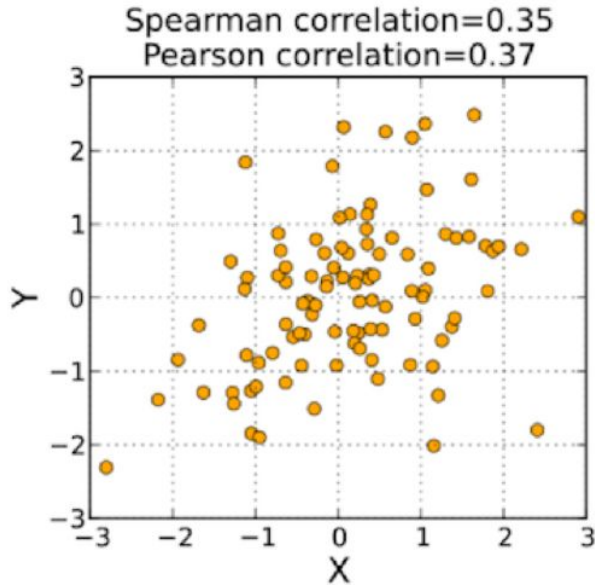
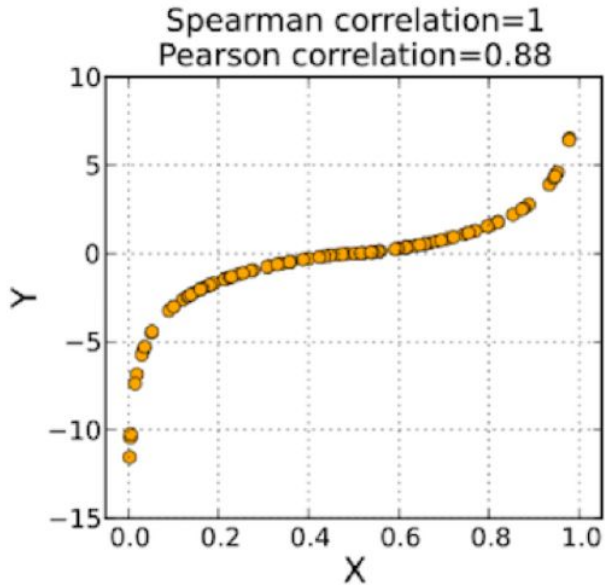


Игра “Угадай корреляцию” -  
<https://www.guessthecorrelation.com/>

# Корреляция Спирмена

Корреляция Пирсона улавливает только линейные зависимости.

Корреляция Спирмена, напротив, улавливает “монотонность”





# Корреляция Спирмена

**Корреляция Спирмена** – мера силы монотонной взаимосвязи. Вычисляется как корреляция Пирсона между рангами наблюдений.

## Правила выставления ранга

1. Порядковый номер наблюдения – ранг
2. Если встречаются несколько одинаковых значений, им присваивается одинаковое значение ранга, равное среднему арифметическому их порядковых номеров

Пример:

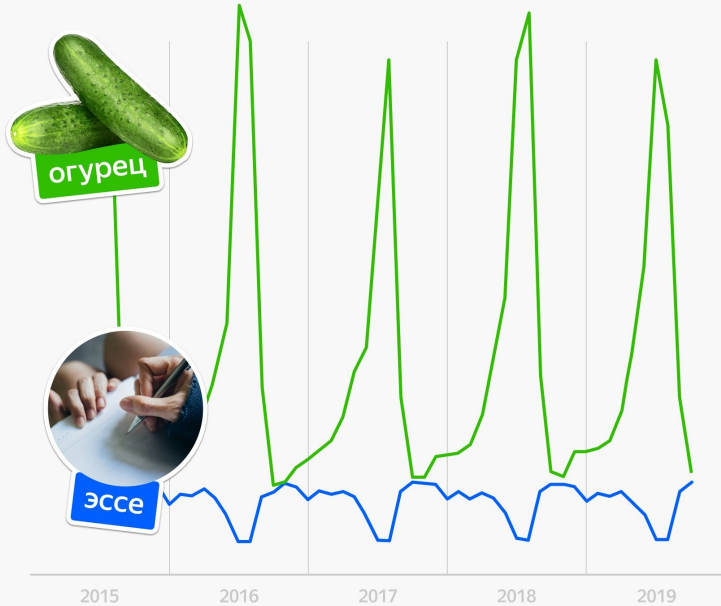
	$X$	$Y$
<b>Выборка:</b>	10, 8, 6, 7, 4, 10, 9, 5	9, 9, 4, 5, 6, 8, 10, 7
<b>Порядок:</b>	7, 5, 3, 4, 1, 8, 6, 2	6, 7, 1, 2, 3, 5, 8, 4
<b>Ранг:</b>	7.5, 5, 3, 4, 1, 7.5, 6, 2	6.5, 6.5, 1, 2, 3, 5, 8, 4
	$r_x$	$r_y$

$$\hat{\rho}_s(X, Y) = \hat{\rho}_p(r_x, r_y) \approx 0.645$$

# Корреляция != Причинность

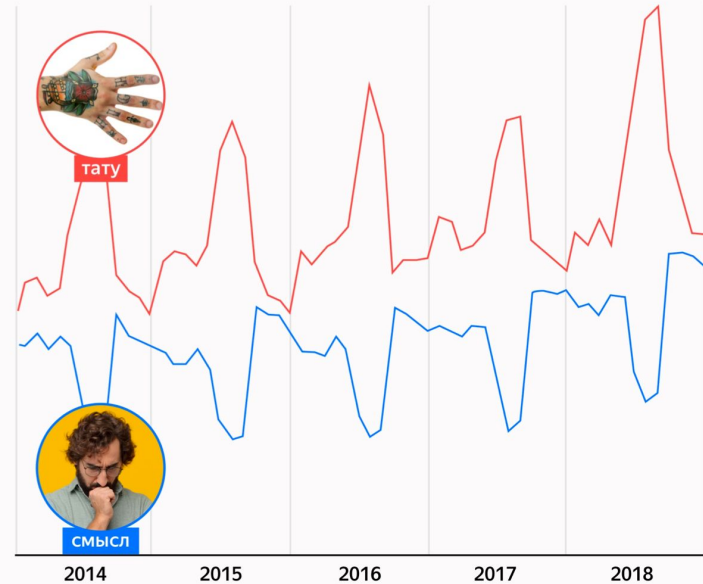
Парадоксы в Поиске — Яндекс

Когда в Поиске взлетает доля запросов со словом **огурец**, становится меньше запросов со словом **эссе**

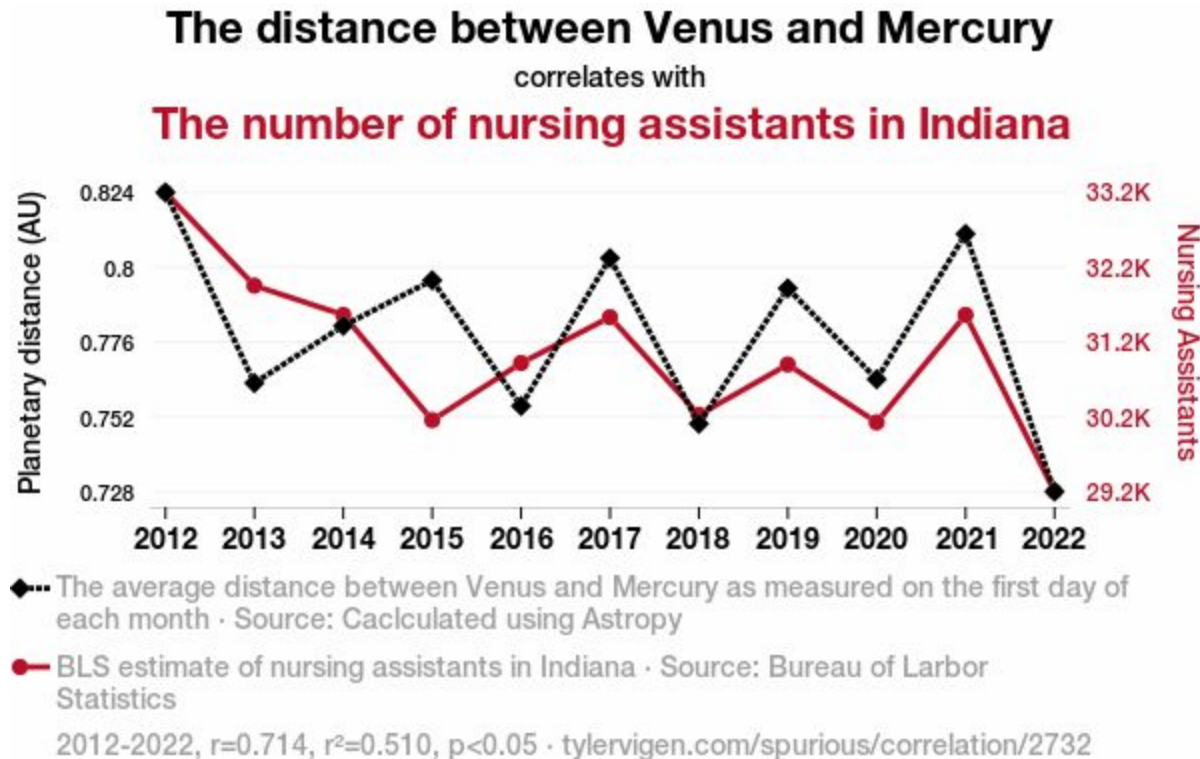


Парадоксы в Поиске — Яндекс

Когда в Поиске растёт интерес к **тату**, снижается доля запросов со словом **смысл**



# Ложная корреляция



Корреляция между величинами может быть вызвана общей причиной:

- Общий тренд в данных
- Спрос на мороженое и число грабежей коррелируют из-за погоды
- Цены на различные продукты могут коррелировать из-за инфляции

Мир сложный, но этим он и прекрасен :)

Что почитать в параллель прохождению курса

- Занимательная статистика. Манга - Син Такахаси
  - Там целая серия книг - [Лабиринт](#)
- Математическая статистика - Н. И. Чернова. [сайт НГУ](#)
- Теория вероятностей - Н. И. Чернова. [сайт НГУ](#)
- The Probability and Statistics Cookbook - <http://statistics.zone/>
- Курс по ТВиМС от Бориса Демешева - [github](#)

**Спасибо за внимание!**