

- Основные распределения в теории вероятностей
- Введение в статистику

- Роль теории вероятностей и статистики
- Распределения и их описательная статистика
- основные виды распределений
- ключевые теоремы статистики (ЦПТ и ЗБЧ)
- области применения статистики

- **Data Science** работает с неопределенностью. Мы редко знаем всё наверняка.
- **Теория Вероятностей (ТВ):** Математический аппарат для описания и измерения неопределенности.
- **Статистика:** Наука о сборе, анализе и интерпретации данных для извлечения знаний и принятия решений в условиях неопределенности.
 - Оценка параметров (например, средний чек покупателя)
 - Проверка гипотез (например, повлияла ли рекламная кампания на продажи?)
 - Построение моделей (предсказание оттока клиентов)

Сопоставление ТВ и статистики

- **Генеральная совокупность(ГС)** - все объекты, которые имеют качества, свойства, интересующие исследователя
- **Выборка** - часть генеральной совокупности элементов, которая охватывается экспериментом

Это два ключевых определения, так как теория вероятностей описывает первое, а статистика занимается вторым!

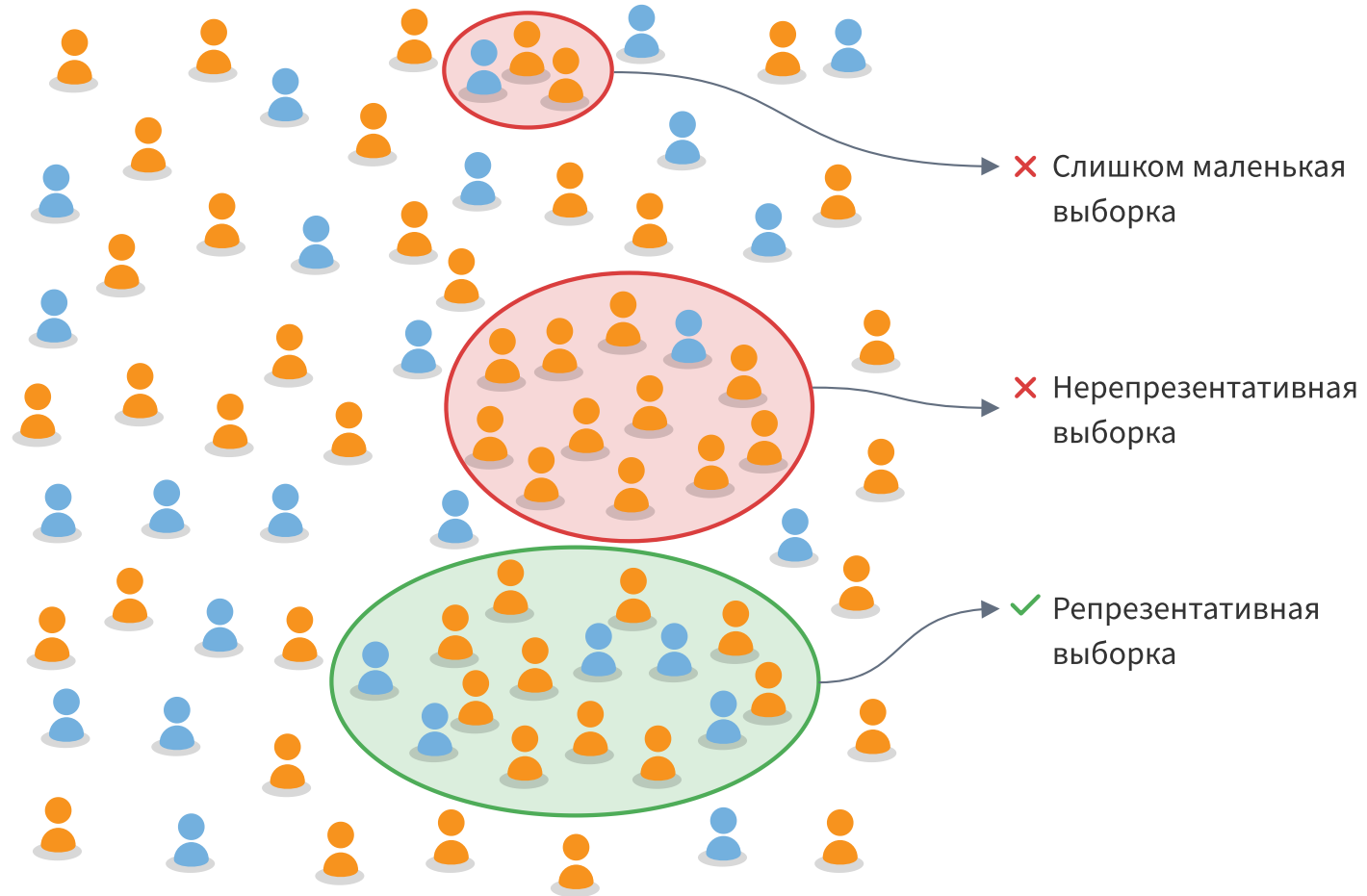
Получить в распоряжение все данные о ГС маловероятно, на практике есть лишь выборки. С помощью выдвигаемых нами гипотез(об этом завтра) мы будем пытаться оценить всю нашу ГС!

Но чтобы правильно оценить, нам нужно знать терминологию описания ГС, поэтому речь сегодня пойдет о распределениях и их характеристиках ГС.

Сопоставление ТВ и статистики

Генеральная совокупность включает

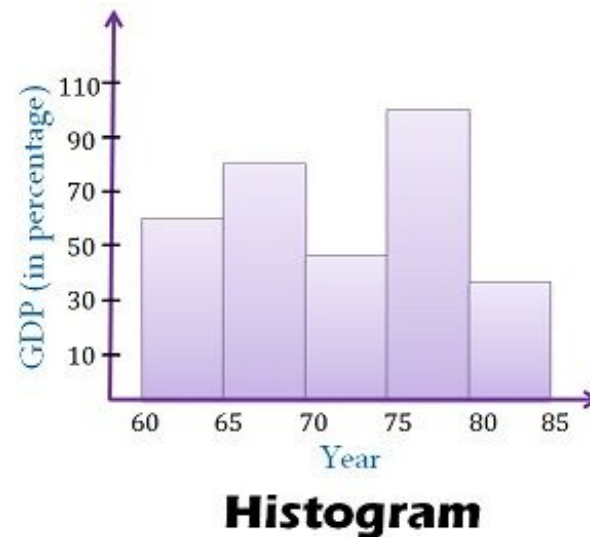
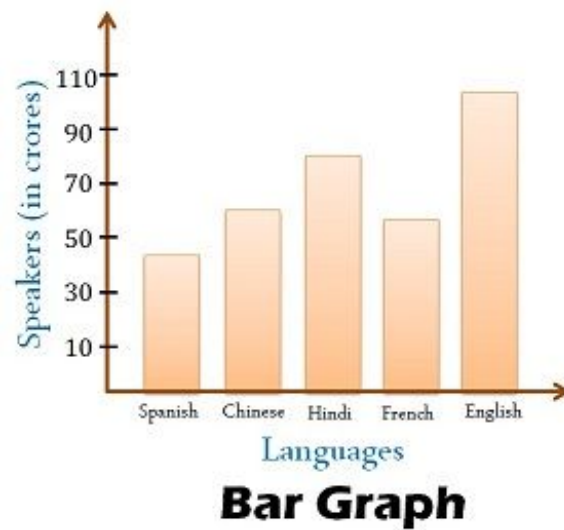
 - 1/3 и  - 2/3



- **Распределением случайной величины** или **Распределением данных** в статистике называют сопоставление значений исследуемых данных с частотой возникновения этих значений, либо с вероятностью возникновения этих значений
- В Питоне гистограмма(**histplot**), или барплот(**barplot**) позволяет взглянуть на *распределение* данных. Либо в частотном, либо в вероятностном виде.

Распределение данных

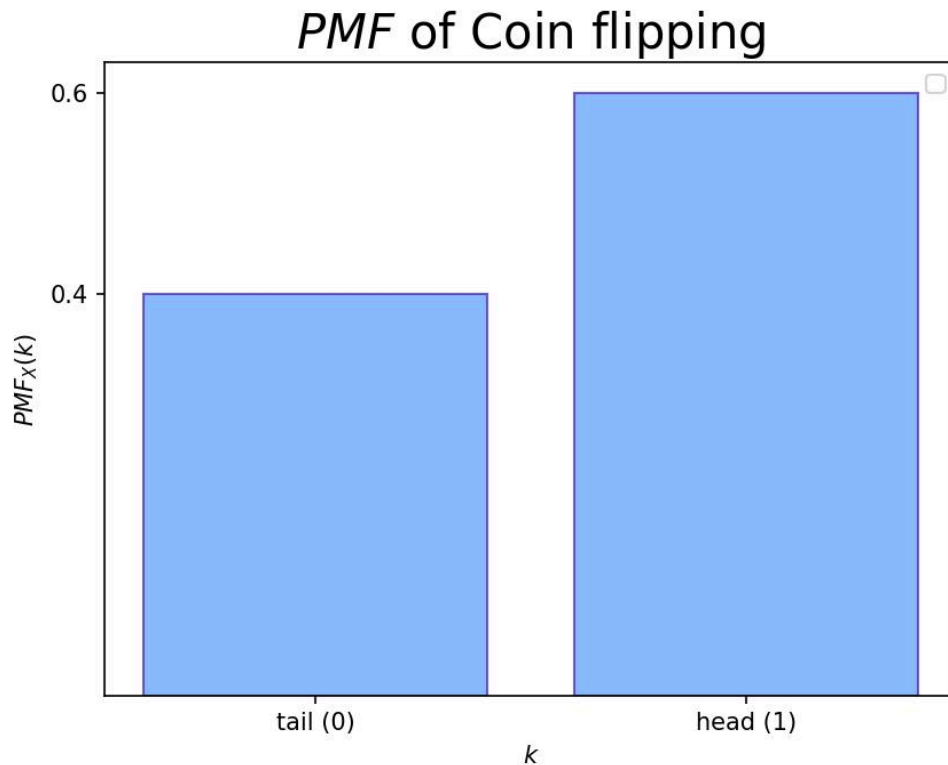
- дискретные
- непрерывные



- дискретные: отдельные значения, которые можно подсчитать
 - тип автомобиля: легковой, грузовой и пр.
 - словесное обозначение цвета: желтый, красный и пр.
 - количество людей в очереди на кассе супермаркета
- непрерывные: принимают любые значения в каком-либо интервале
 - рост
 - расстояние
 - возраст
 - пр.

Функция вероятности (pmf)

Функция, возвращающая вероятность того, что дискретная случайная величина X примет определённое значение k . Является инструментом анализа **дискретных** величин. Обозначение: $P_X(k) = P(X = k) = PMF_X(k)$, где k -дискретно.



$$X \sim Be(p)$$

$$X \sim \begin{pmatrix} tails(0) & heads(1) \\ 1-p & p \end{pmatrix}$$

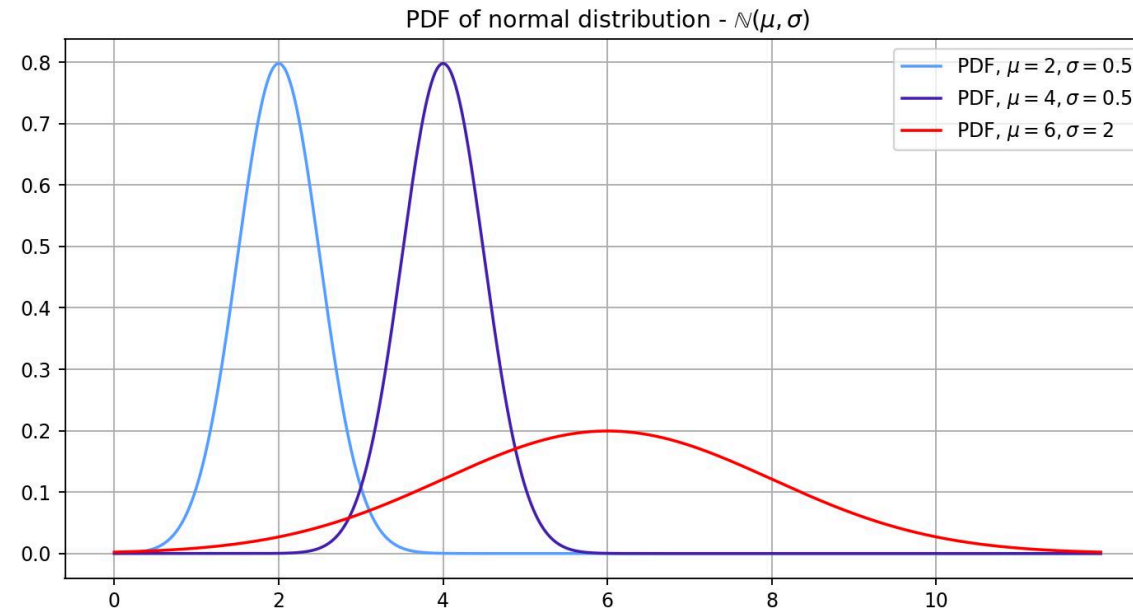
В нашем случае $p = 0.6$



`scipy.stats.bernoulli(p=0.6).pmf()` (probability mass function)

Функция плотности вероятности (pdf)

Функция, характеризующая сравнительную вероятность реализации тех или иных значений случайной переменной (переменных). Является инструментом анализа непрерывных величин. Обозначение: $f_{\xi}(x) = P(\xi = x)$, x - непрерывен.



`scipy.stats.norm(loc=2, scale=0.5).pdf()` (probability density function)

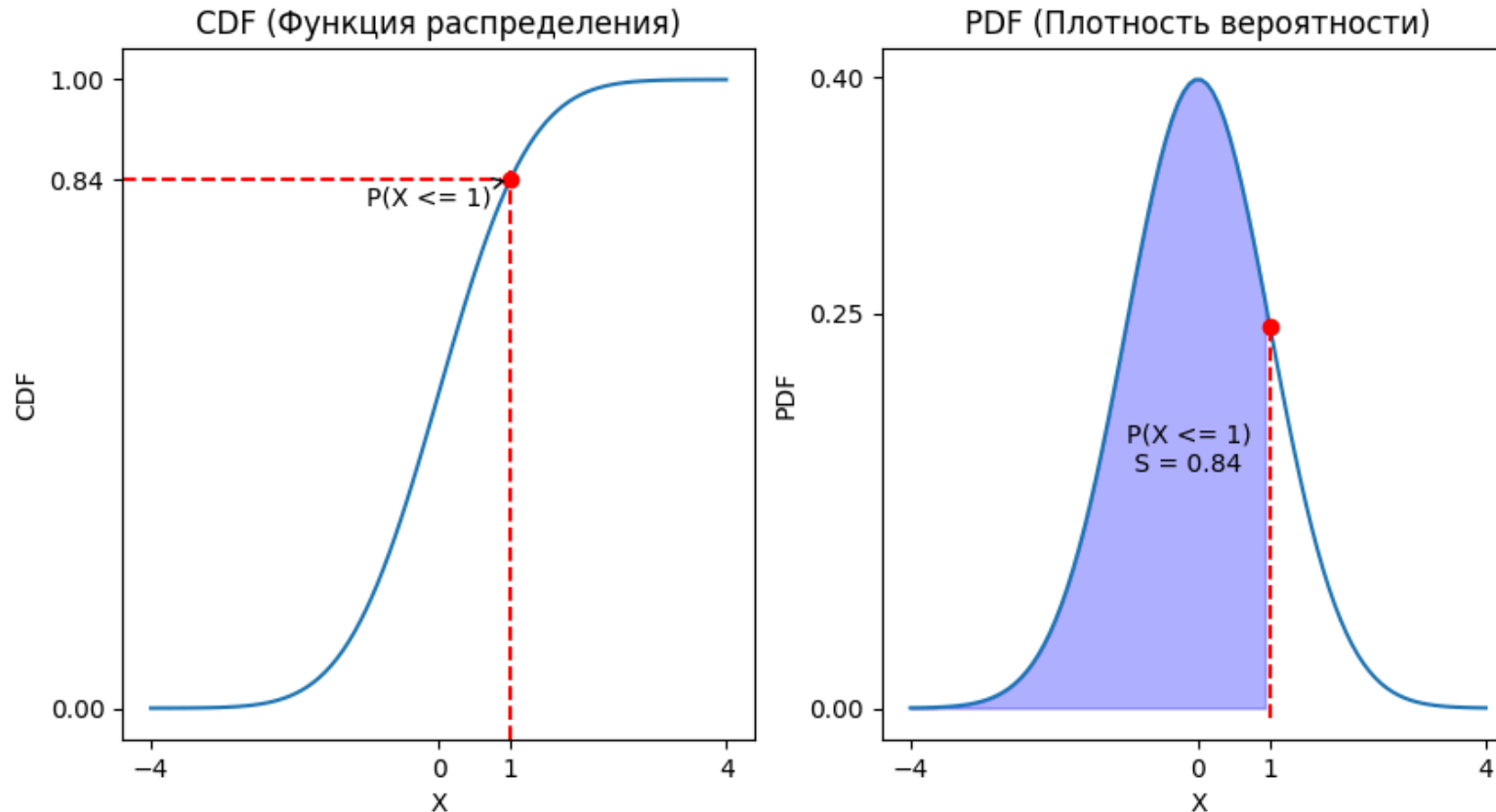
Функция распределения(cdf)

Функция, характеризующая распределение случайной величины или случайного вектора; вероятность того, что случайная величина X примет значение, меньшее либо равное x , где x — произвольное действительное число. Обозначение:

$F_{\xi}(x) = P(\xi \leq x)$. Общий для **дискретных** и **непрерывных**.

Функция распределения(cdf)

- Графическая связь с cdf и pdf



`scipy.stats.norm(loc=0, scale=1).cdf(1)` (cumulative distribution function)


Квантиль распределения(ppf)

- значение, которое заданная случайная величина не превышает с фиксированной вероятностью. Если вероятность задана в процентах, то квантиль называется *процентилем* или *перцентилем*

- Квантиль уровня γ :

Если $F(x_\gamma) = \gamma$, то значение $x_\gamma = F^{-1}(\gamma)$ - является кватилем уровня γ

- Квантиль является обратной функцией, к функции распределения

 `scipy.stats.norm(loc=0, scale=1).ppf(0.84)` (percent point function)

Основные характеристики распределений

Для ГС:

Математическое ожидание(среднее): $\mathbb{E}[X]$, $\mathbb{M}[X]$

- Сколько в среднем принимает случайная величина
- Для дискретной случайной величины $\mathbb{E}[X] = \sum_i^n x_i p_i$
- Для непрерывной случайной величины $\mathbb{E}[X] = \int_{-\infty}^{+\infty} x_i f(x_i)$

Для выборки:

Выборочное среднее является наилучшей оценкой мат.ожидания ГС, иначе говоря лучше всего приближает это значение (`np.mean()`)

Выборочное среднее: $\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$



`np.mean()`

Основные характеристики распределений

Для ГС:

Дисперсия: $\mathbb{D}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$

- Мера *разброса* случайной величины
- Часто обозначается как $\mathbb{D}[X]$, $Var[X]$ (variance)

Для выборки:

Выборочная дисперсия является наилучшей дисперсии ГС, иначе говоря лучше всего приближает это значение (`np.var()`)

Выборочная дисперсия: $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})^2$



`np.var(ddof=1)` (при `ddof=1` в знаменателе формулы $n-1$)

Стандартное отклонение: $\sigma_x = \sqrt{\mathbb{D}[X]}$

- тоже мера разброса, но измеряется в тех же величинах, что и исходная случайная величина
- вычисляется просто как корень из дисперсии

Для выборки:

- аналогично извлечь корень из выборочной дисперсии



`np.std(ddof=1)` (при `ddof=1` в знаменателе формулы $n-1$)

Справка по функциям

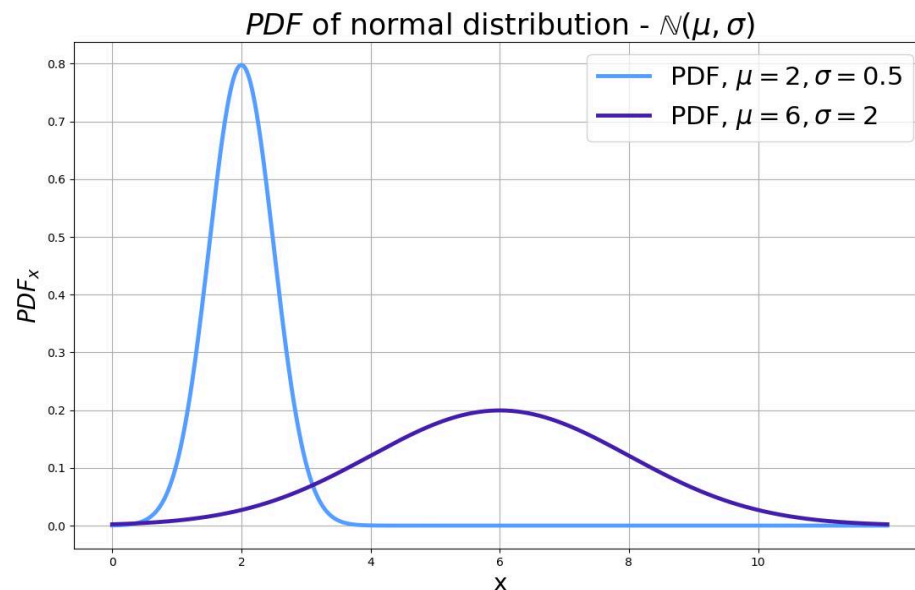
<code>scipy func</code>	Название	Вопрос	Пример
<code>pmf</code>	Функция вероятности	Какова вероятность, что дискретная величина X равна x ?	Какова вероятность того, что при броске кубика выпадет число 3?
<code>pdf</code>	Функция плотности вероятности	Какова вероятность, что непрерывная величина X находится вблизи x ?	Какова вероятность того, что рост человека будет около 170 см?
<code>cdf</code>	Функция распределения	Какова вероятность, что $X \leq x$?	Какова вероятность того, что время ожидания автобуса составит не более 10 минут?
<code>ppf</code>	Обратная функция распределения	Какое значение X соответствует вероятности p ?	Каков уровень зарплаты, что 30% людей получает меньше данной суммы?
<code>mean</code>	Среднее значение	Каково среднее значение случайной величины X ?	Каково среднее значение уровня сахара в крови в выборке пациентов?
<code>var</code>	Дисперсия	Какова дисперсия случайной величины X ?	Какова дисперсия уровня шума в офисе в течение рабочего дня?
<code>std</code>	Стандартное отклонение	Каково стандартное отклонение случайной величины X ?	Каково стандартное отклонение времени доставки посылок в течение недели?

- Ниже приведены основные типы распределений

Нормальное распределение

Возникновение в природе:

Если величина является суммой многих случайных слабо взаимозависимых величин, каждая из которых вносит малый вклад относительно общей суммы, то центрированное и нормированное распределение такой величины при достаточно большом числе слагаемых стремится к *нормальному* распределению.



- Куполообразная форма

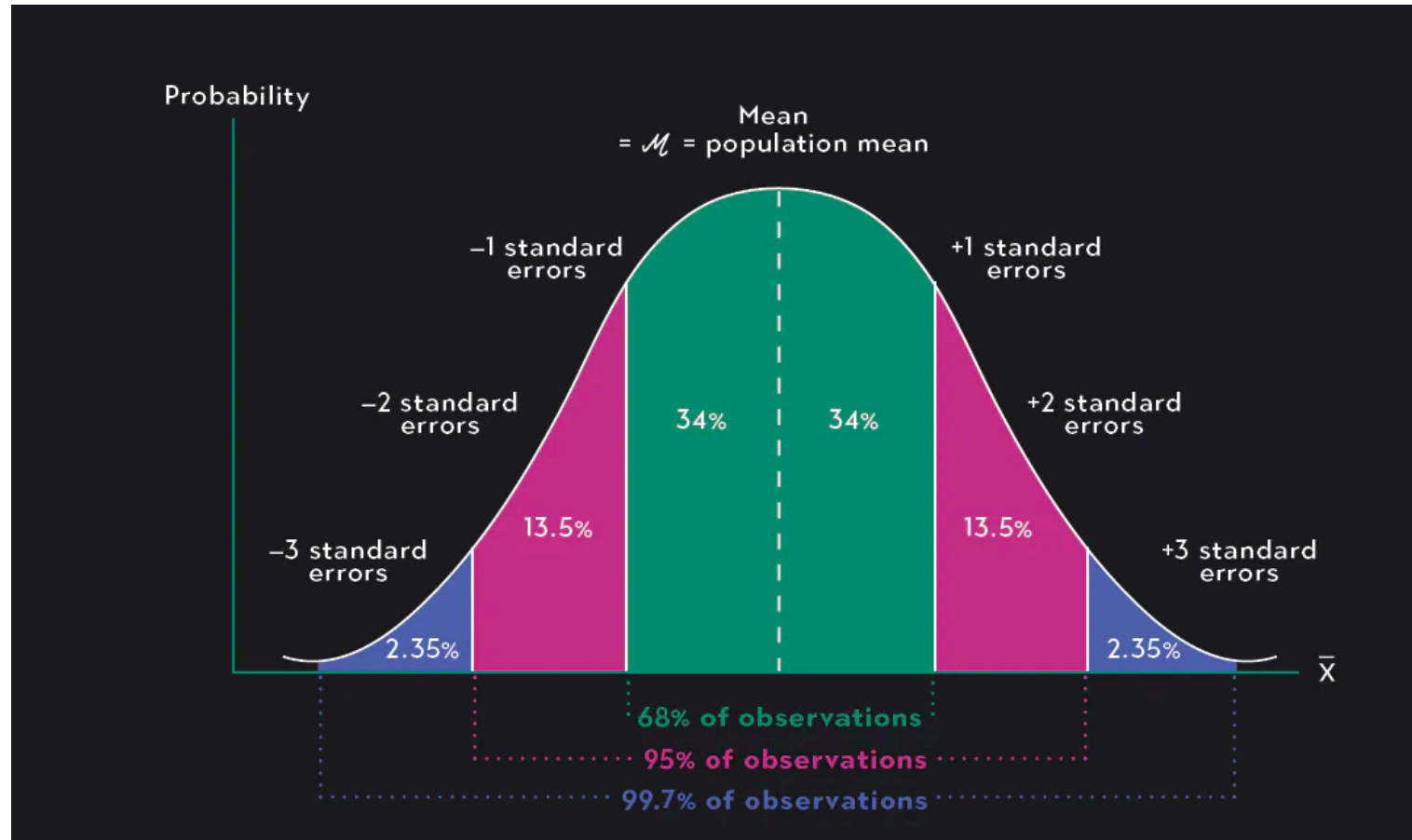
$$x \sim \mathcal{N}(\mu, \sigma)$$

$$\mathbb{E}_x = \mu$$

$$\sigma_x = \sigma$$

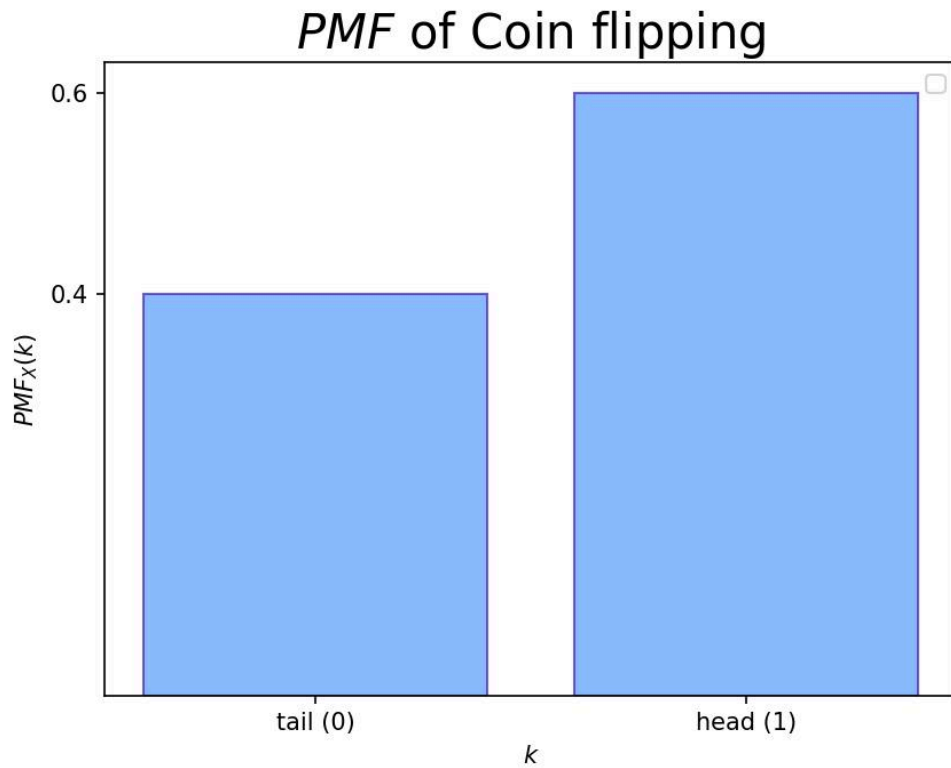
Правило трех сигм для нормального распределения

- Данное правило сформулировано именно для нормального распределения. Однако эти проценты можно узнать и для любого другого распределения



Распределение Бернулли

Дискретное распределение вероятностей, моделирующее случайный эксперимент произвольной природы, при заранее известной вероятности успеха (p) или неудачи ($q = 1 - p$). (Приведите пример из жизни такого распределения)



x - Что выпало на монете

$$x \sim Be(p)$$

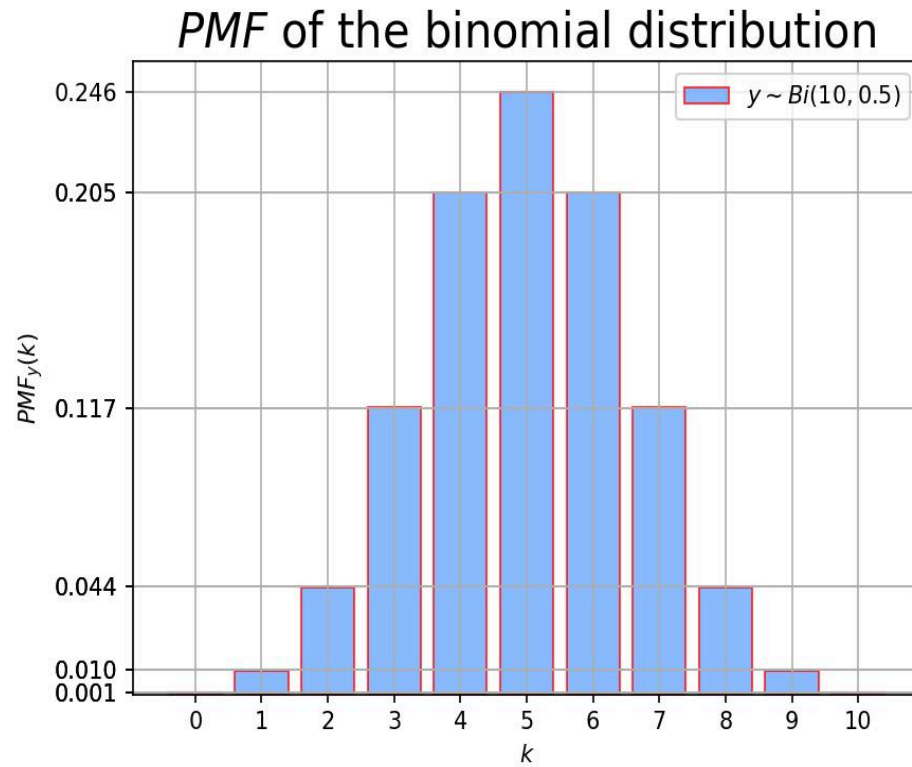
$$\mathbb{E}_x = p$$

$$\sigma_x = \sqrt{p(1-p)}$$

$$PMF_x(x = k) = p^k \cdot (1-p)^{1-k}$$

Биномиальное распределение

Распределение количества "успехов" в последовательности из n независимых случайных экспериментов, таких, что вероятность "успеха" в каждом из них постоянна и равна p . (Приведите пример из жизни такого распределения)



$$x_i \sim Be(p)$$

$$y = x_1 + x_2 + \dots + x_n$$

$$y \sim Bi(n, p)$$

$$\binom{n}{k} = \frac{n!}{(n-k)!k!} - \text{биномиальный}$$

коэффициент

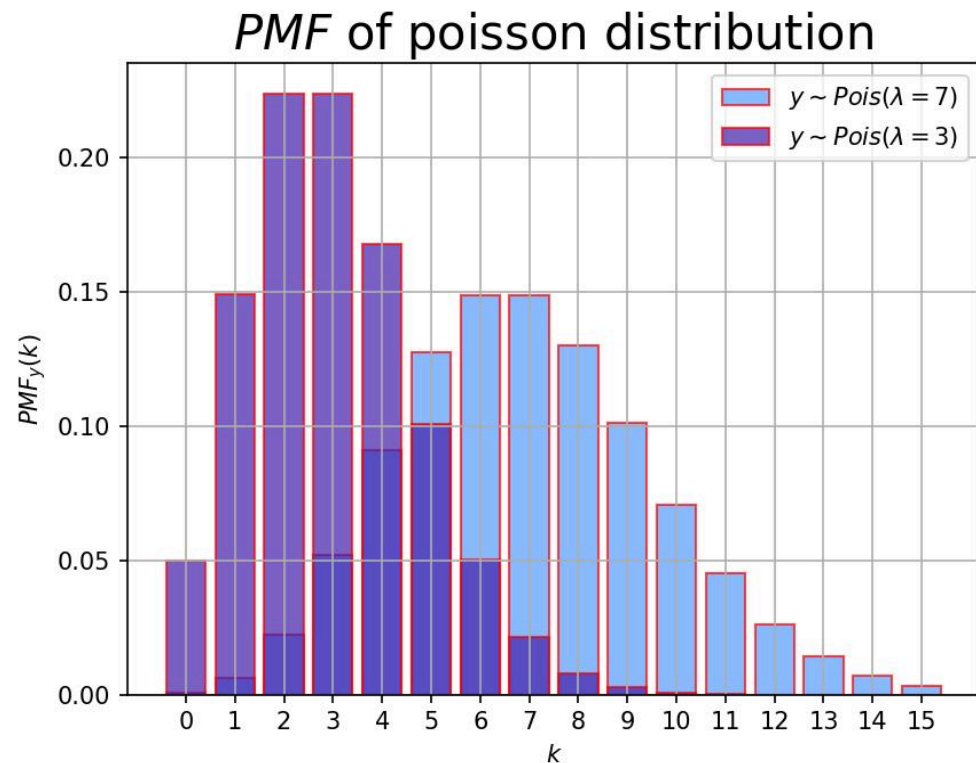
$$pmf_y(k) = \mathbb{P}(y = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\mathbb{E}_x = n * p$$

$$\sigma_x = \sqrt{n * p * (1-p)}$$

Дискретное распределение случайной величины, представляющей собой **число событий, произошедших за фиксированное время**, при условии, что данные события происходят с некоторой фиксированной средней интенсивностью и независимо друг от друга. Является приближением биномиального распределения, в случае, когда n - очень большое, а p - очень маленькое. $n \cdot p = \lambda$.

Распределение Пуассона



$$y \sim \text{Pois}(\lambda)$$

$$pmf_y(k) = \mathbb{P}(y = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

$$\mathbb{E}_x = \lambda$$

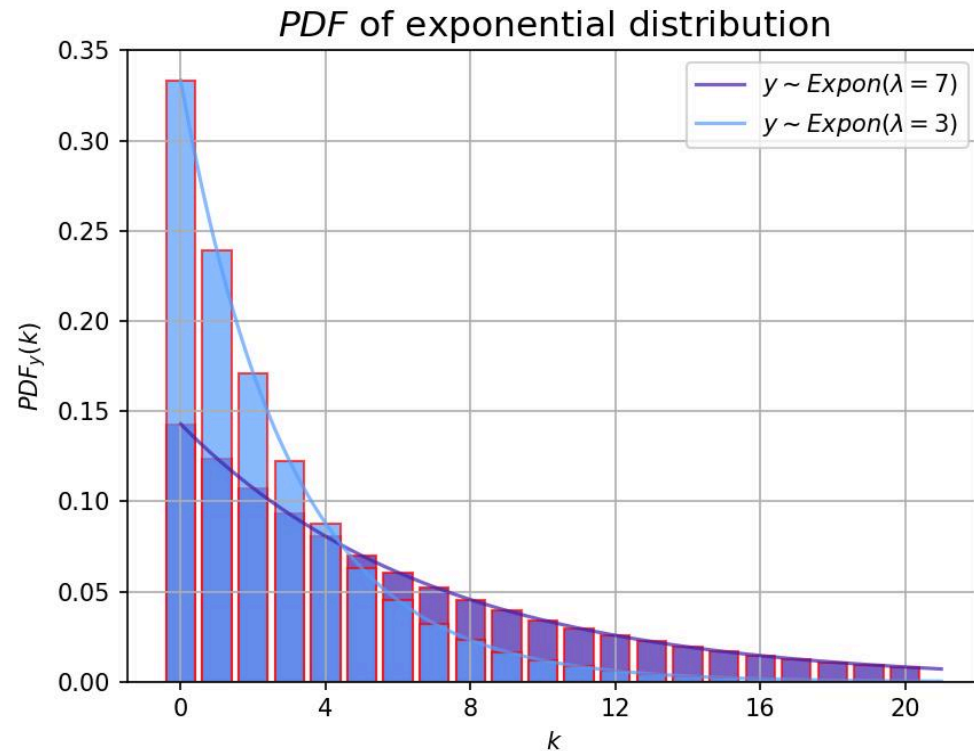
$$\sigma_x = \sqrt{\lambda}$$

- Количество опоздавших автобусов за день
- Число посетителей на веб-сайте за сутки



```
scipy.stats.poisson(lambda)
```


Экспоненциальное распределение



$$y \sim \text{exp}(\lambda)$$

$$pdf_y(k) = \mathbb{P}(y = k) = \lambda \cdot e^{-\lambda x}$$

$$\mathbb{E}_x = \frac{1}{\lambda}$$

$$\sigma_x = \frac{1}{\lambda}$$

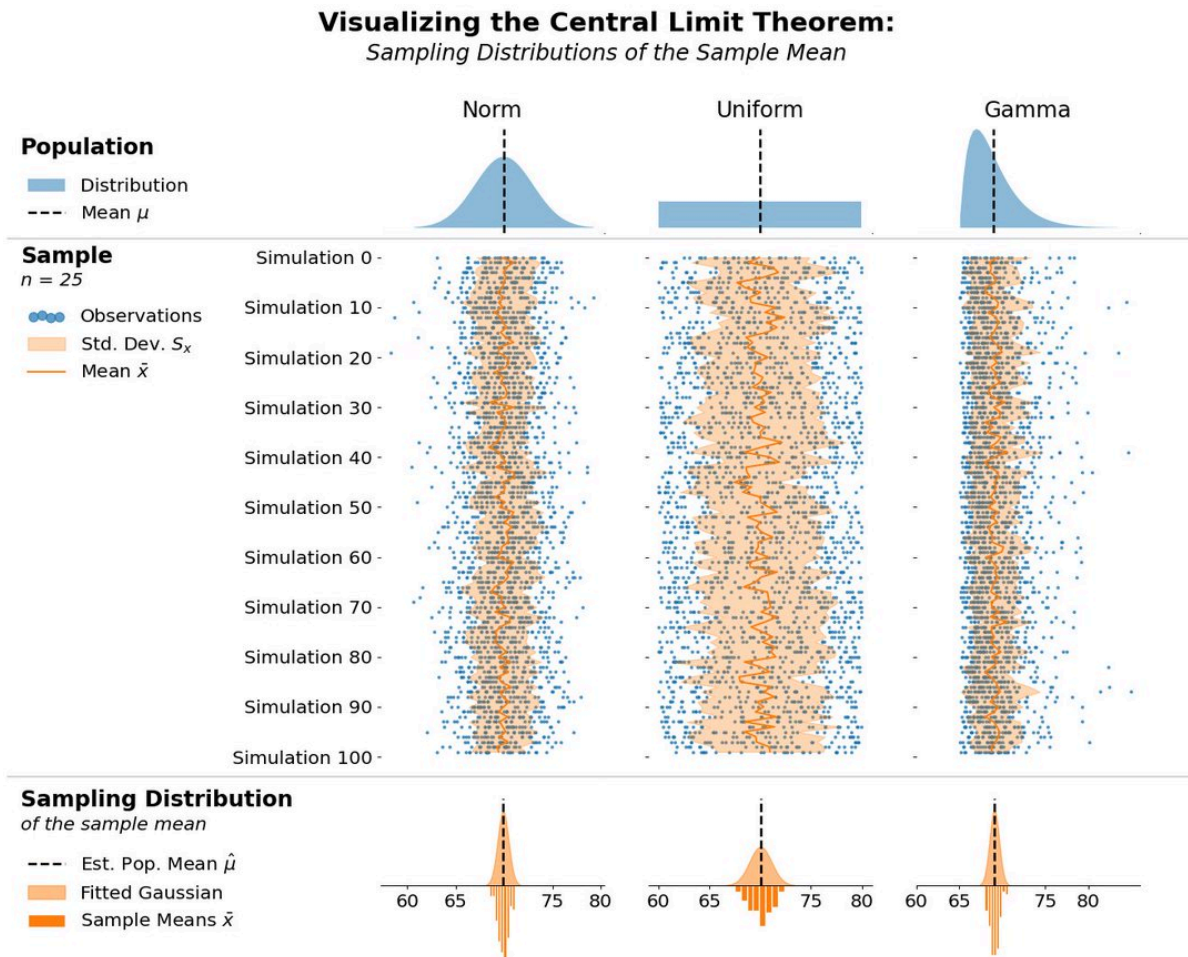
- Время между отказами оборудования
- Время потраченное на звонок с клиентом

🔥 `scipy.stats.expon(lambda)`

Центральная предельная теорема(!!!)

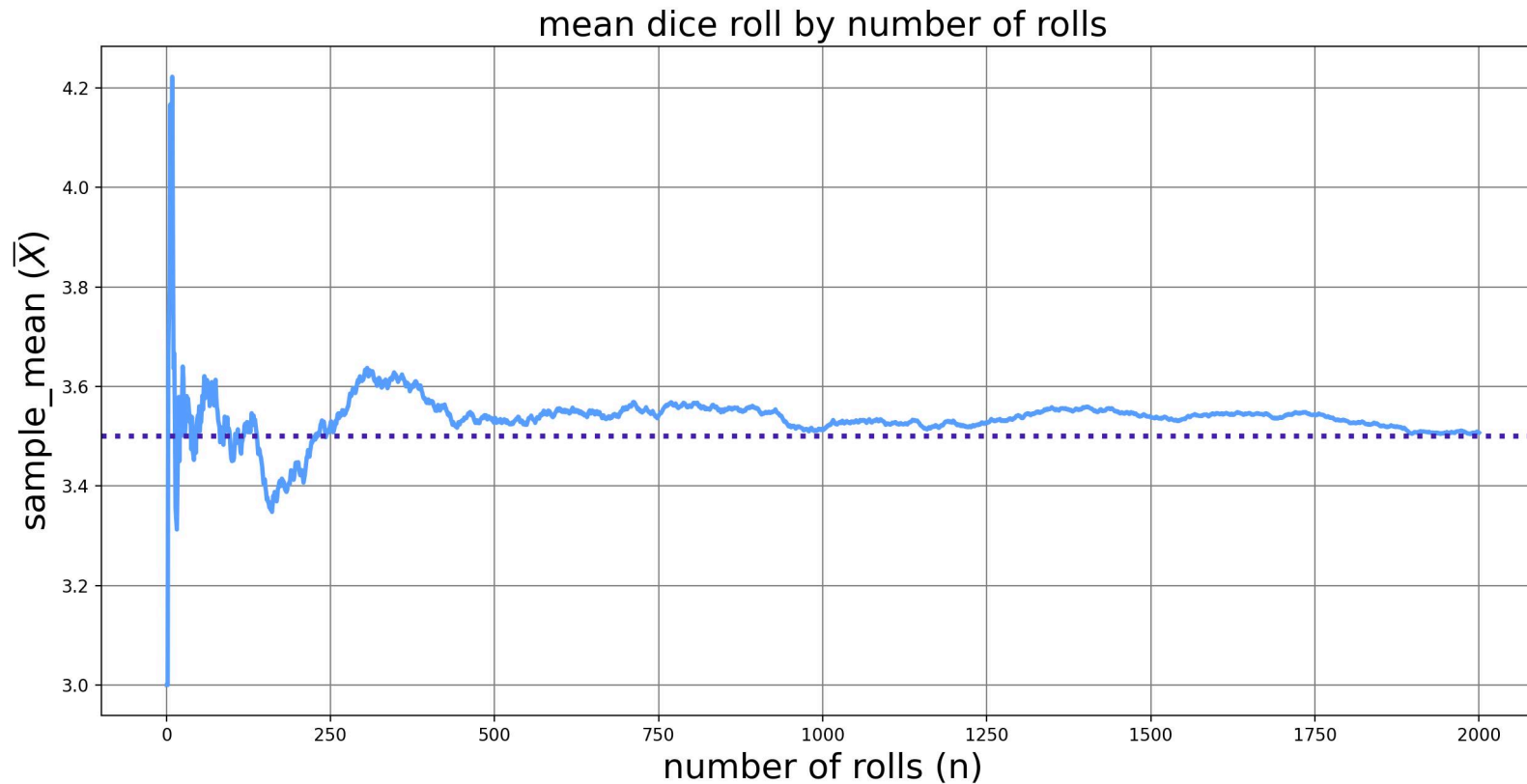
Для данной генеральной совокупности, описанной любым распределением вероятностей, имеющим среднее μ и конечную дисперсию σ^2 , распределение выборочного среднего \bar{X} , вычисленное по выборке размера n из этой совокупности будет приблизительно нормальным со средним μ (среднее значение совокупности) и дисперсией $\frac{\sigma^2}{n}$ (дисперсия совокупности деленная на n), при большом размере выборки n .

Центральная предельная теорема(!!!)



Author: Cameron Riddell; Twitter @RiddleMeCam

- Лежит в основе статистической проверки гипотез, об этом завтра



Идея простая: при увеличении числа испытаний выборочное среднее случайной величины стремится к истинному математическому ожиданию распределения.

- Случайные величины появляются повсеместно: свойства объектов, измерения и пр
- Часто они бывают близки к каким-то распределениям
- Распределение – математическая модель
- У распределений есть характеристики(параметры) – об этом надо помнить
- Всех распределений не выучить: [Univariate Distribution Relationships](#)
- Необходимо знать только базовые(в презентации), остальные, по мере необходимости