

# Продвинутая статистика



Корреляционный анализ данных (продолжение). Линейная регрессия.  
Функция потерь. Матрицы ковариаций и корреляций. Правило трех сигм.  
Центральная предельная теорема. Виды распределений: дискретные и непрерывные распределения. Дискретное равномерное распределение.  
Логнормальное распределение. Экспоненциальное распределение.  
Распределение Бернулли. Биноминальное распределение.  
t-критерий Стьюдента.

**Даниил Корбут**

Специалист по Анализу Данных



**Даниил Корбут**  
DL Researcher  
Insilico Medicine, Inc

Окончил бакалавриат ФИВТ  
МФТИ (Анализ данных) в 2018г  
Учусь на 2-м курсе  
магистратуры ФИВТ МФТИ  
Работал в Statsbot и Яндекс.  
Алиса.  
Сейчас в Insilico Medicine, Inc,  
занимаюсь генерацией  
активных молекул и  
исследованиями старения с  
помощью DL.

# Нахождение зависимости случайных величин

**Дисперсия** — квадрат среднеквадратичного отклонения от среднего значения (насколько данные разбросаны)

$$\sigma^2(x) = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

**Ковариация** — наличие зависимости между величинами

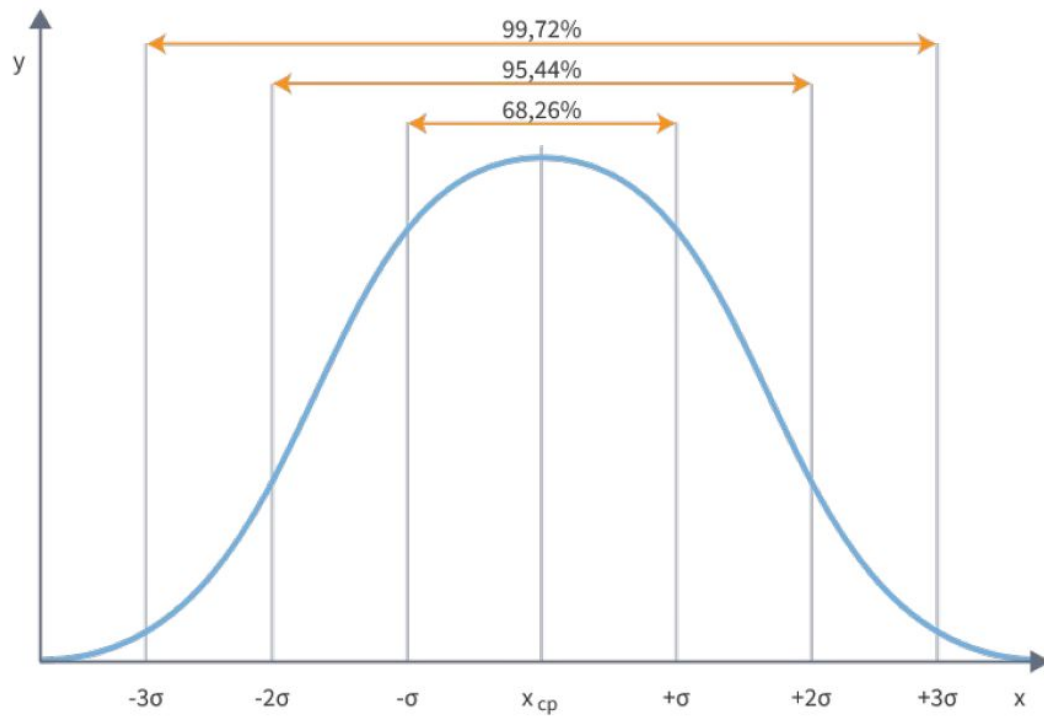
$$\sigma(x, y) = \frac{1}{n} \sum_{i=1}^n (x - \mu_x)(y - \mu_y)$$

Ковариация не равна нулю — можно предположить зависимость.

Ковариация показывает разброс величин относительно друг друга.  
Проблема ковариации: данные могут иметь разный масштаб.

**Корреляция** — нормированная ковариация.

# Правило трёх сигм



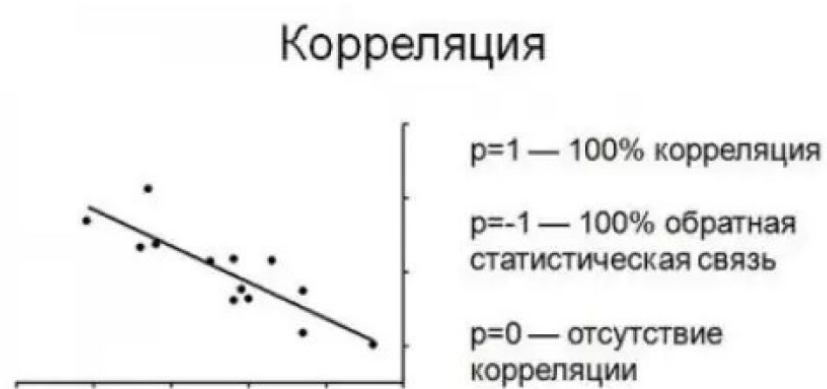
<https://wiki.loginom.ru/articles/3-sigma-rule.html>

# Корреляция Пирсона - нормированная ковариация

**Корреляция Пирсона** — нормированная ковариация, определяет силу зависимости

$$\sigma(x, y) = \frac{Cov(x, y)}{\sqrt{Var(x)}\sqrt{Var(y)}} = \frac{\frac{1}{n} \sum_{i=1}^n (x - \mu_x)(y - \mu_y)}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu_x)^2} \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \mu_y)^2}}$$

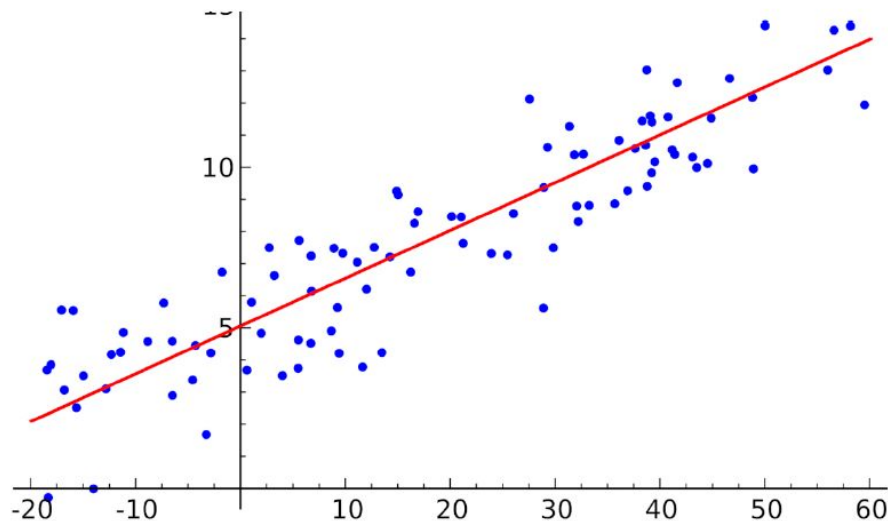
# Корреляция Пирсона



[http://economic-definition.com/Exchange\\_Terminology/Koefficient\\_korrelyacii\\_Correlation\\_coefficient\\_\\_eto.html](http://economic-definition.com/Exchange_Terminology/Koefficient_korrelyacii_Correlation_coefficient__eto.html)

# Линейная регрессия

**Линейная регрессия** — модель зависимости переменной  $x$  от одной или нескольких других переменных (факторов, регрессоров, независимых переменных) с линейной функцией зависимости



Модель:

$$y = f(x, b) + \varepsilon,$$

где  $\varepsilon$  - случайная ошибка модели

Функция регрессии имеет вид

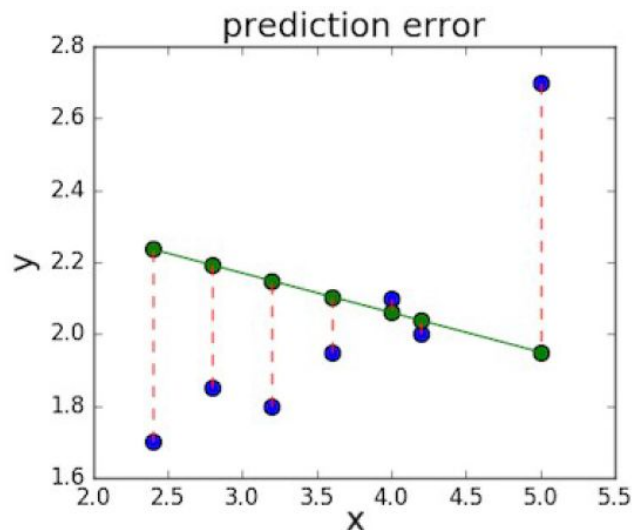
$$f(x, b) = b_0 + b_1x_1 + b_2x_2 + \dots + b_kx_k$$

$b_j$  - параметры (коэффициенты) регрессии  
 $x_j$  - атрибуты

<https://neurohive.io/ru/osnovy-data-science/linejnaja-regressija/>

# Функция потерь

**Функция потерь** — мера количества ошибок, которые линейная регрессия делает на наборе данных



Метод наименьших квадратов:

$$\sum_i e_i^2 = \sum_i (y_i - f_i(x))^2 \rightarrow \min_x.$$

<https://neurohive.io/ru/osnovy-data-science/linejnaja-regressija/>

[https://ru.wikipedia.org/wiki/Метод\\_наименьших\\_квадратов](https://ru.wikipedia.org/wiki/Метод_наименьших_квадратов)



# Алгоритм построения модели линейной регрессии

Для того, чтобы построить модель линейной регрессии в python, необходимо:

- 1) выбрать предсказываемую величину ( $y$ ) и независимую величину ( $x$ )  
( $x$  величина может быть многомерной,  $y$  – только одномерная)
- 2) разделить данные на тренировочные (80%) и тестовые (20%)
- 3) создать модель линейной регрессии (с помощью библиотеки `sklearn`)
- 4) обучаем модель на тренировочных данных
- 5) посчитать ошибку на тестовых данных (с помощью функции потерь)
- 6) оценить качество модели
- 7) сделать график

# Центральная предельная теорема (ЦПТ)

Давайте рассматривать выборки из случайных величин.

Выборка из  $X \sim F(x)$  :  
 $X^n = (X_1, X_2, \dots, X_n)$

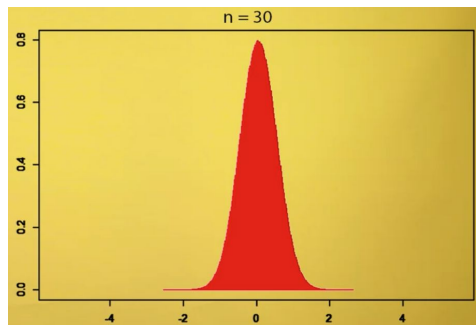
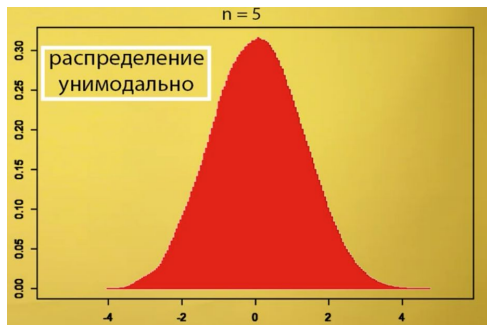
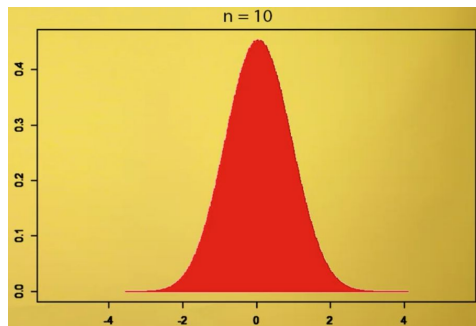
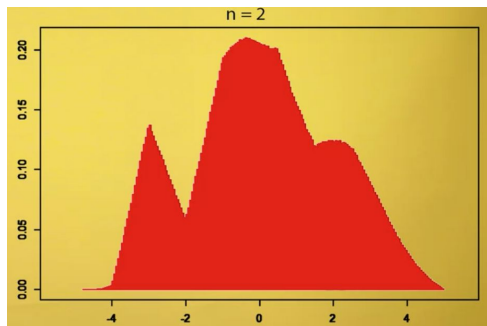
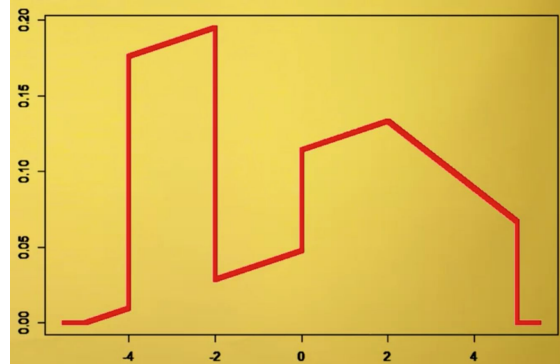
Выборочное среднее:  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$

У выборочного среднего пишем нижний индекс  $n$ , просто чтобы понимать с выборкой какого размера мы работаем. Давайте подумаем, как связано выборочное среднее с исходным распределением?

$$\bar{X}_n \sim ?$$

# Центральная предельная теорема (ЦПТ)

Будем работать с таким “странным” распределением. Давайте будем семплировать выборки объема  $n$ , считать по ним выборочные средние и повторять так много-много раз. И давайте построим гистограмму этих выборочных средних.



На плотность какого распределения похожи полученные графики?

# Центральная предельная теорема

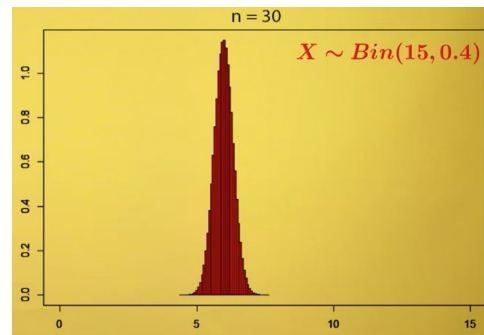
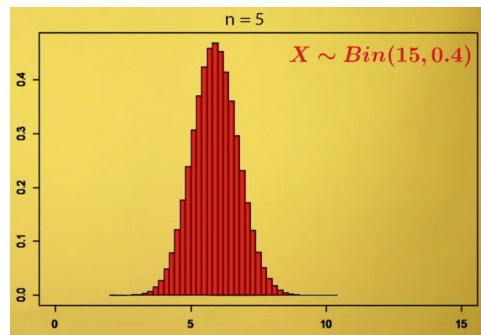
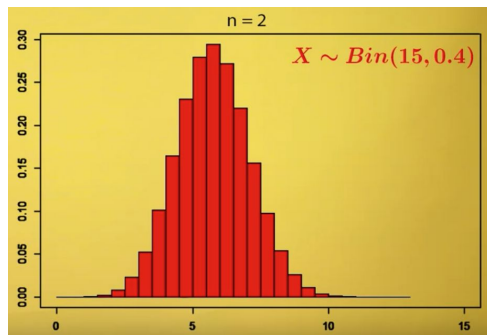
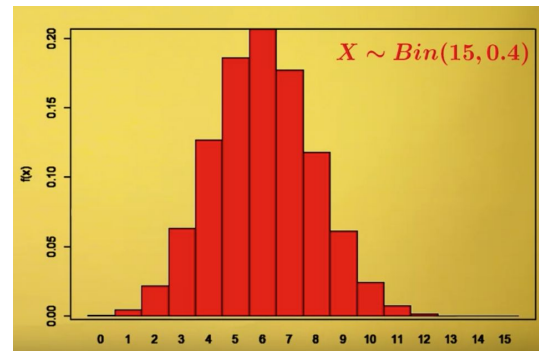
$$X \sim F(x),$$

$$X^n = (X_1, X_2, \dots, X_n) \Rightarrow$$

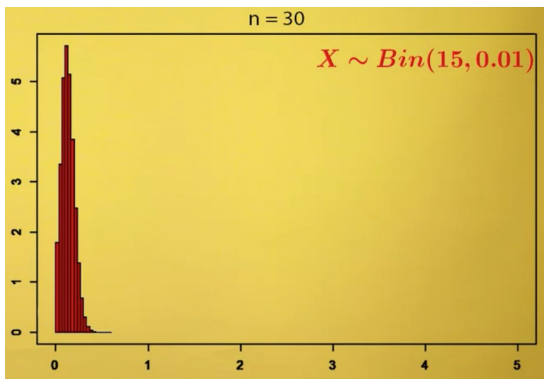
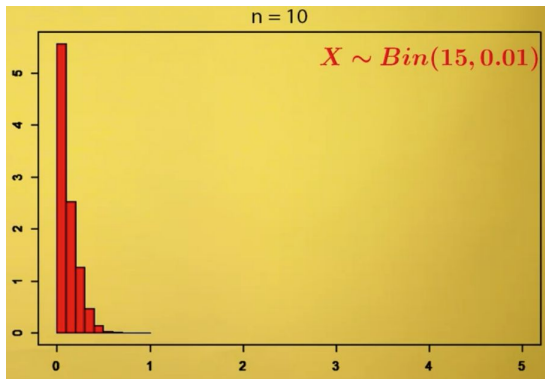
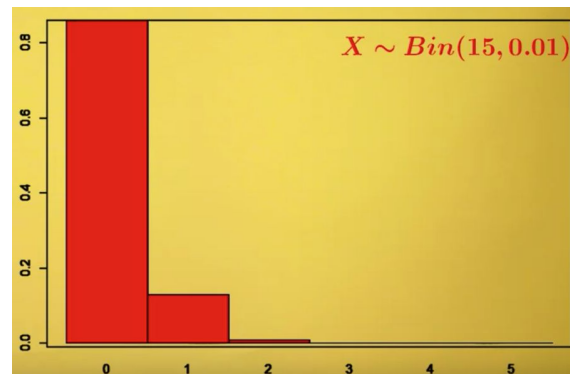
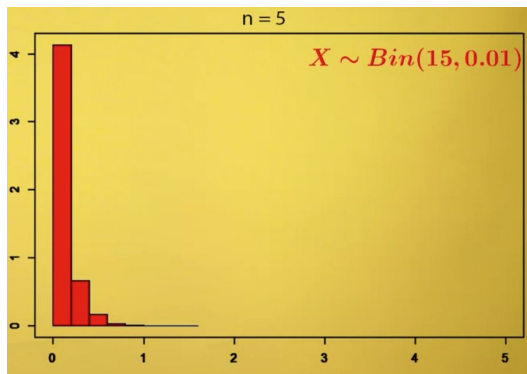
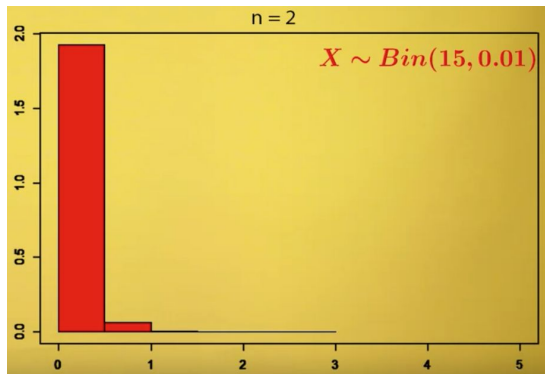
$$\bar{X}_n \approx \sim N(\mathbb{E}X, \frac{\mathbb{D}X}{n})$$

С ростом  $n$  точность аппроксимации увеличивается

Интересно, что это справедливо не только для абсолютно непрерывных распределений, но и для дискретных.

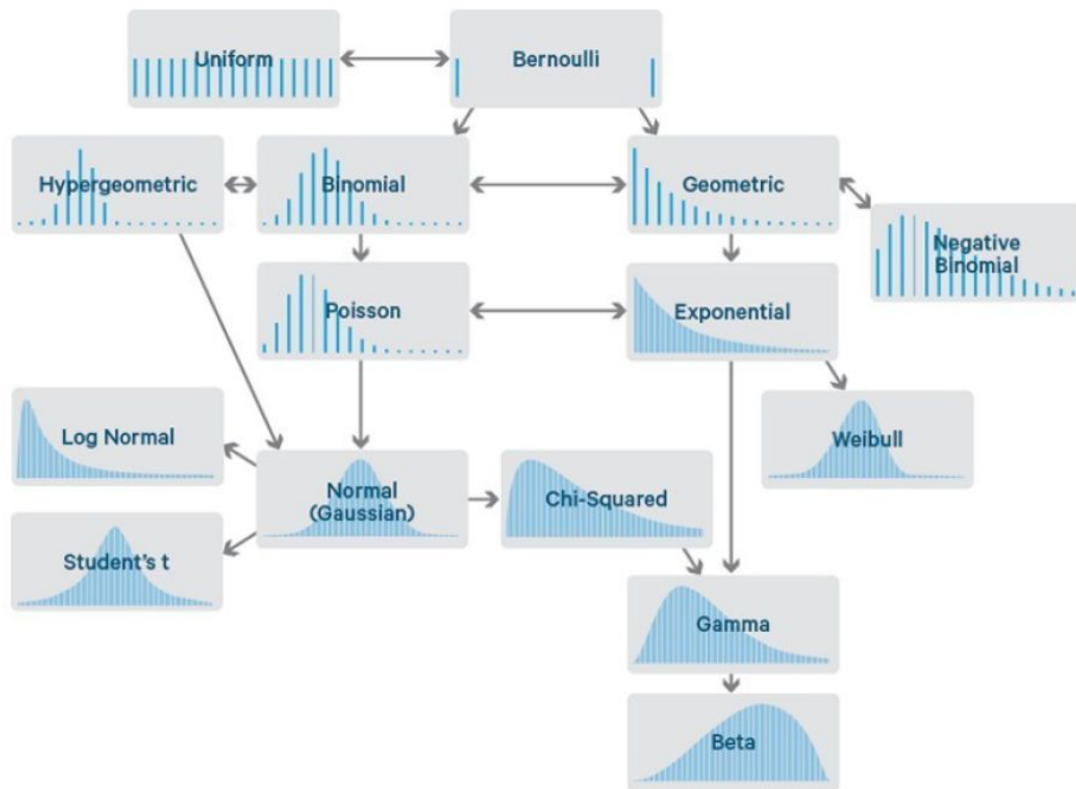


# Центральная предельная теорема



Когда распределение  $X$  не слишком  
скошено, распределение  $\bar{X}_n$  хорошо  
описывается нормальным при  $n \geq 30$ .

# Виды распределений



<https://habr.com/ru/post/331060/>



# Дискретные и непрерывные распределения

**Дискретной случайной** величиной называется случайная величина, которая в результате испытания принимает отдельные значения с определёнными вероятностями. Число возможных значений дискретной случайной величины может быть конечным и бесконечным. Примеры дискретной случайной величины: запись показаний спидометра или измеренной температуры в конкретные моменты времени.

**Непрерывной случайной** величиной называют случайную величину, которая в результате испытания принимает все значения из некоторого числового промежутка. Число возможных значений непрерывной случайной величины бесконечно. Пример непрерывной случайной величины: измерение скорости перемещения любого вида транспорта или температуры в течение конкретного интервала времени.

# Распределение Бернулли

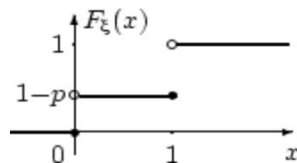
**Случайная величина** — переменная, значения которой представляют собой исходы какого-нибудь случайного феномена или эксперимента.

Простыми словами: это численное выражение результата случайного события.

$$y = X(\omega)$$

Случайная величина  **$X$**  имеет **распределение Бернулли**, если она принимает всего два значения: 1 и 0 с вероятностями  $p$  и  $q=1-p$  соответственно.

$$F_{\xi}(x) = P(\xi < x) = \begin{cases} 0, & x \leq 0; \\ 1-p, & 0 < x \leq 1 \\ 1, & x > 1. \end{cases}$$



$$\mathbb{P}(X = 1) = p,$$

$$\mathbb{P}(X = 0) = q.$$

Принято говорить, что событие  $\{X = 1\}$  соответствует «успеху», а  $\{X = 0\}$  «неудаче». Эти названия условные, и в зависимости от конкретной задачи могут быть заменены на противоположные.



# Биномиальное распределение

Случайная величина  $\xi$  имеет **биномиальное распределение** (англ. *binomial distribution*) с параметрами  $n \in \mathbb{N}$  и  $p \in (0, 1)$  и пишут:  $\xi \in \mathbb{B}_{n,p}$  если  $\xi$  принимает значения  $k = 0, 1, \dots, n$  с вероятностями  $P(\xi = k) = \binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k}$ .

Случайная величина с таким распределением имеет смысл числа успехов в  $n$  испытаниях схемы Бернулли с вероятностью успеха  $p$ .

Таблица распределения  $\xi$  имеет вид

$\xi$	0	1	...	$k$	...	$n$
$P$	$(1 - p)^n$	$n \cdot p \cdot (1 - p)^{n-1}$	...	$\binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k}$	...	$p^n$

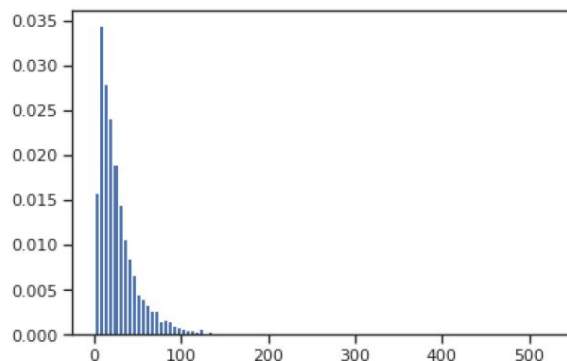
# Логнормальное распределение

Логнормальное распределение задается плотностью вероятности:

$$p(x) = \frac{1}{\sigma x \sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}$$

Где  $\mu$  - это среднее значение,  $\sigma$  - стандартное отклонение

Пример:



$$\mu = 3$$

$$\sigma = 0.9$$

# Экспоненциальное распределение

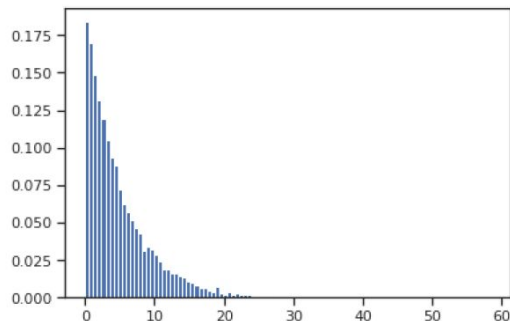
Экспоненциальное распределение задается плотностью вероятности:

$$f(x; \frac{1}{\beta}) = \frac{1}{\beta} \exp(-\frac{x}{\beta})$$

если  $x \geq 0$ , иначе  $f(x; \frac{1}{\beta}) = 0$

где  $\beta$  - это параметр

Пример:



$$\beta = 5$$

## Распределение Стьюдента

Мы хотим сгенерировать нормальное распределение, но по некоторым причинам не можем вычислить среднеквадратичное отклонение (например, выборка маленькая). Мы можем найти выборочное среднее и выборочную дисперсию по выборке.

Пусть  $x_1, \dots, x_n$  — выборка размером  $n$

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

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

## t-критерий Стьюдента

Случайная величина  $t$  имеет распределение Стьюдента с  $n-1$  степенями свободы, где  $n$  — размер выборки.

$$t = \frac{\bar{X} - m}{s_X / \sqrt{n}}$$

Данный критерий был разработан Уильямом Госсетом для оценки качества пива в компании Гиннесс. В связи с обязательствами перед компанией по неразглашению коммерческой тайны (руководство Гиннеса считало таковой использование статистического аппарата в своей работе), статья Госсета вышла в 1908 году в журнале «Биометрика» под псевдонимом «Student» (Студент).

## t-критерий Стьюдента

Применяется для проверки нулевой гипотезы  $H_0 : E(X) = m$

о равенстве математического ожидания  $E(X)$  некоторому неизвестному значению  $m$ .

$$t = \frac{\bar{X} - m}{s_X / \sqrt{n}}$$

При нулевой гипотезе распределение этой статистики  $t(n-1)$ .

Следовательно, при превышении значения статистики по абсолютной величине критического значения данного распределения (при заданном уровне значимости) нулевая гипотеза отвергается.

**Спасибо за внимание!**