

# Юстина Иванова

Data scientist

Корреляция. Линейная регрессия.  
Виды распределений.  
Центральная предельная теорема.  
Статистический анализ данных в python.



Спикер



## Юстина Иванова,

- PhD в университет Больцано (Италия)
- Data scientist по компьютерному зрению в компании ОЦРВ, Сочи
- Выпускница МГТУ им. Баумана
- Магистр по Artificial Intelligence в University of Southampton (Англия)



# 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.

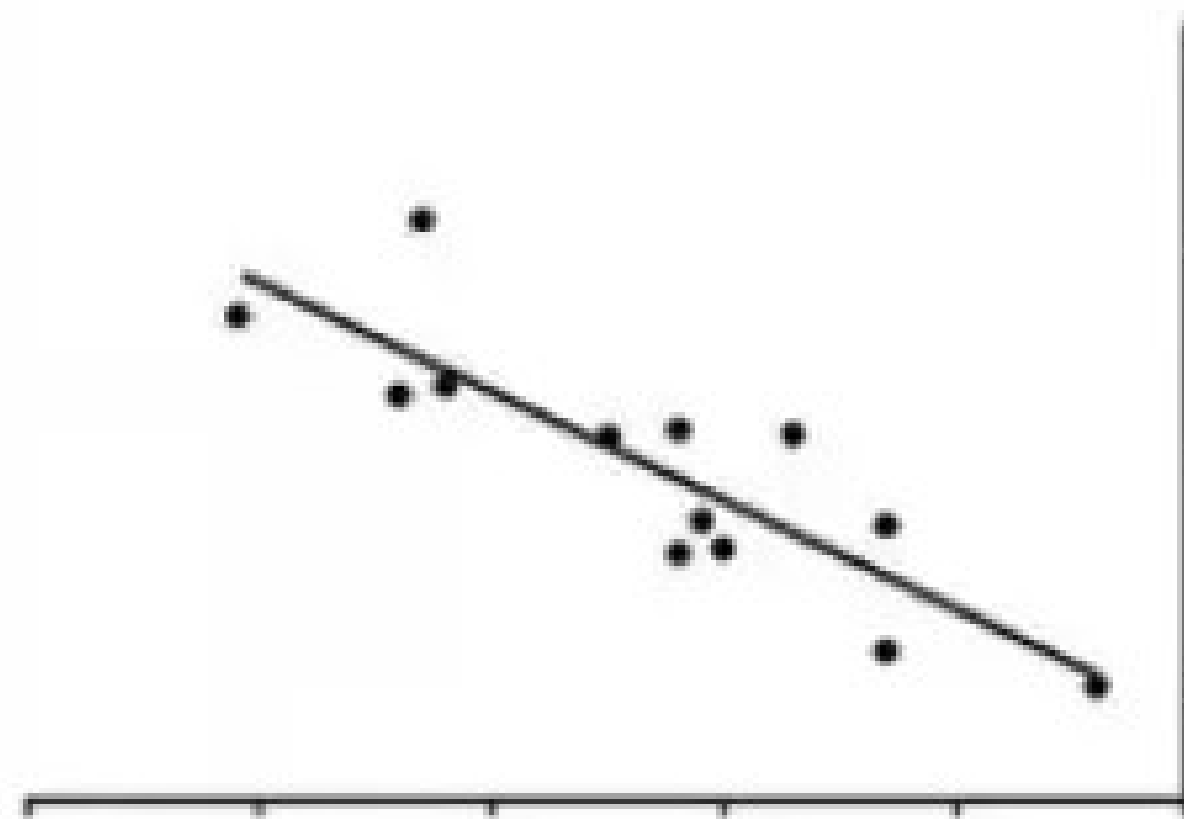
# Корреляция Пирсона — нормированная ковариация

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

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

# Корреляция Пирсона

## Корреляция



$r=1$  — 100% корреляция

$r=-1$  — 100% обратная  
статистическая связь

$r=0$  — отсутствие  
корреляции

# Линейная регрессия

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

Модель:

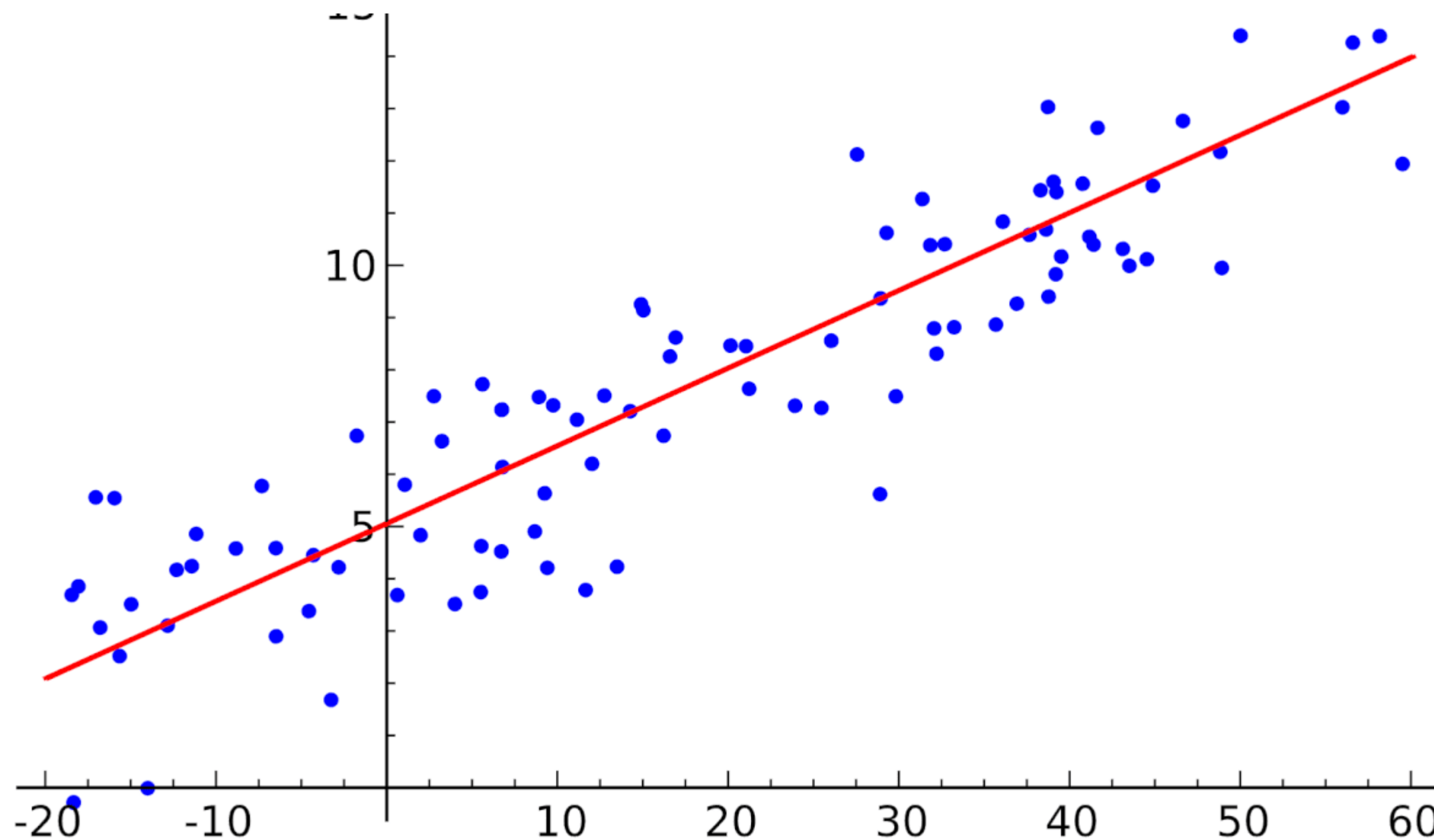
$$y = f(x, b) + \varepsilon,$$

где  $\varepsilon$  - случайная ошибка модели

Функция регрессии имеет вид

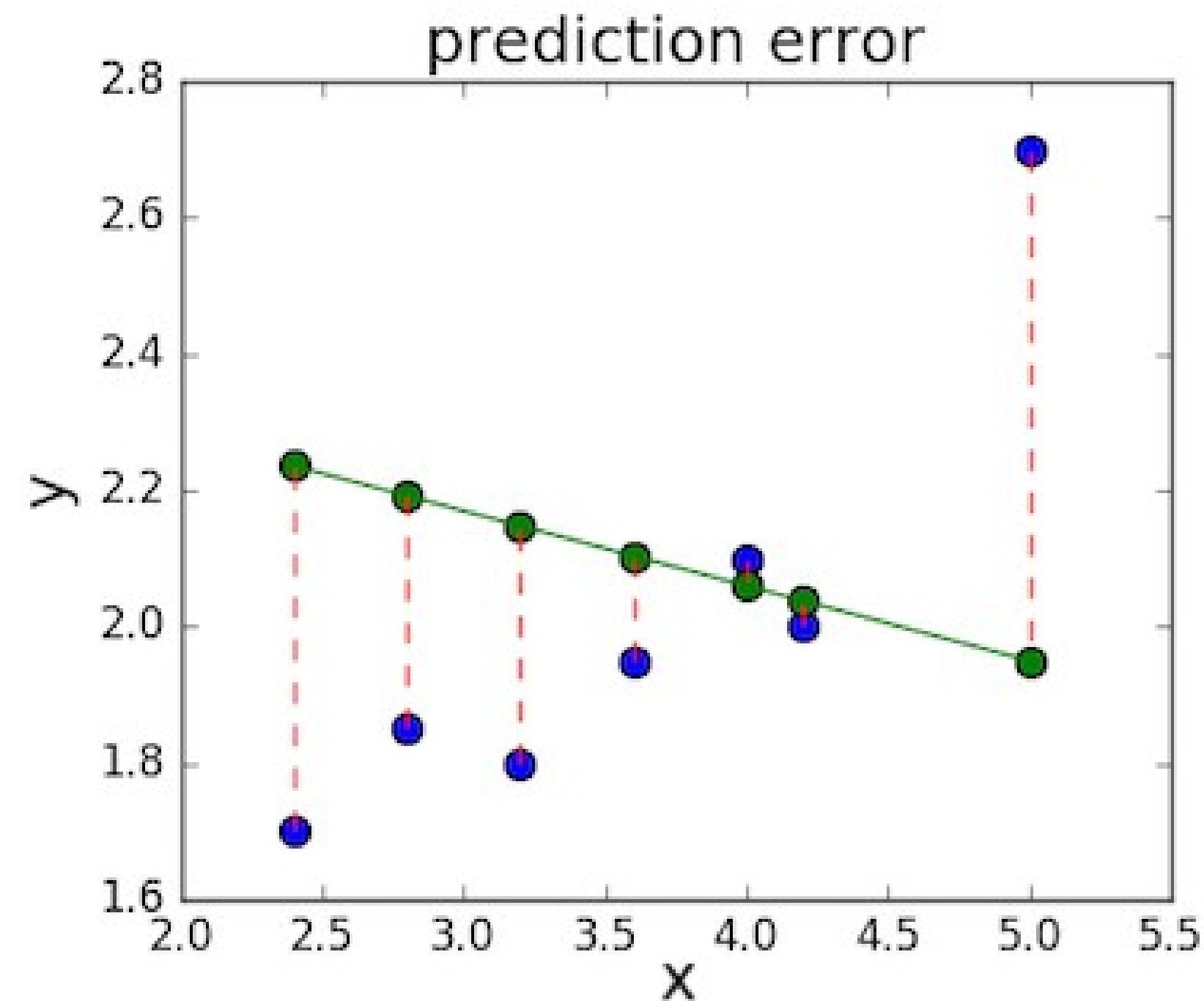
$$f(x, b) = b_0 + b_1x_1 + b_2x_2 + \dots + b_kx_k$$

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



# 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



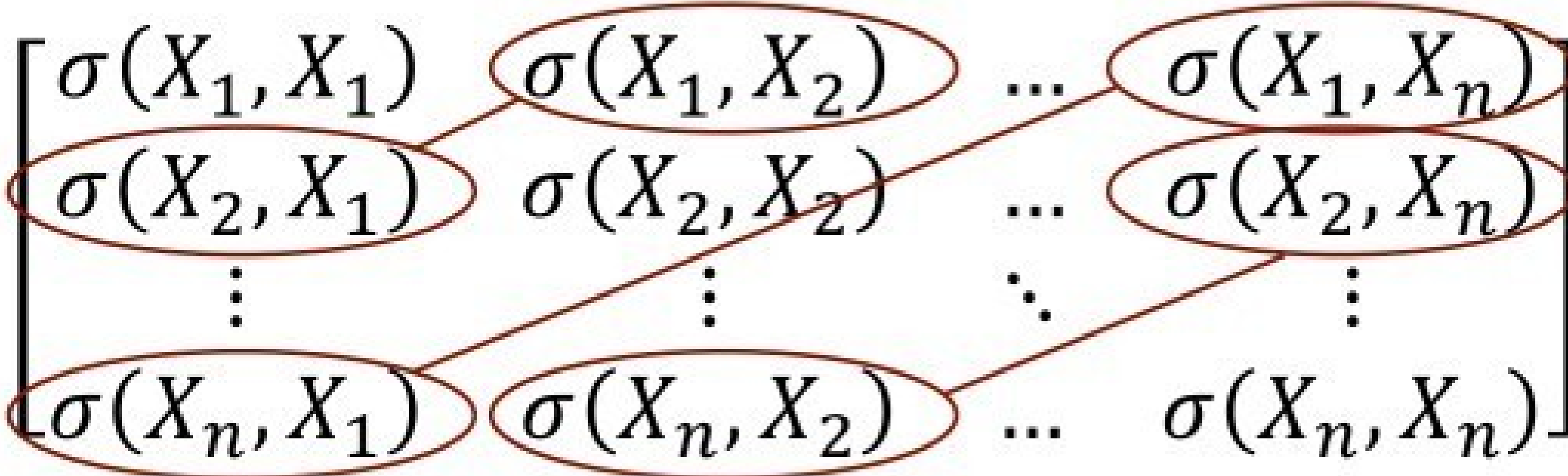
# Матрица корреляций

Матрица корреляций подсчитывается с помощью формул, которые показывают как данные зависят друг от друга в пространстве  $n$  значений (каждый элемент матрицы равен коэффициенту Пирсона).

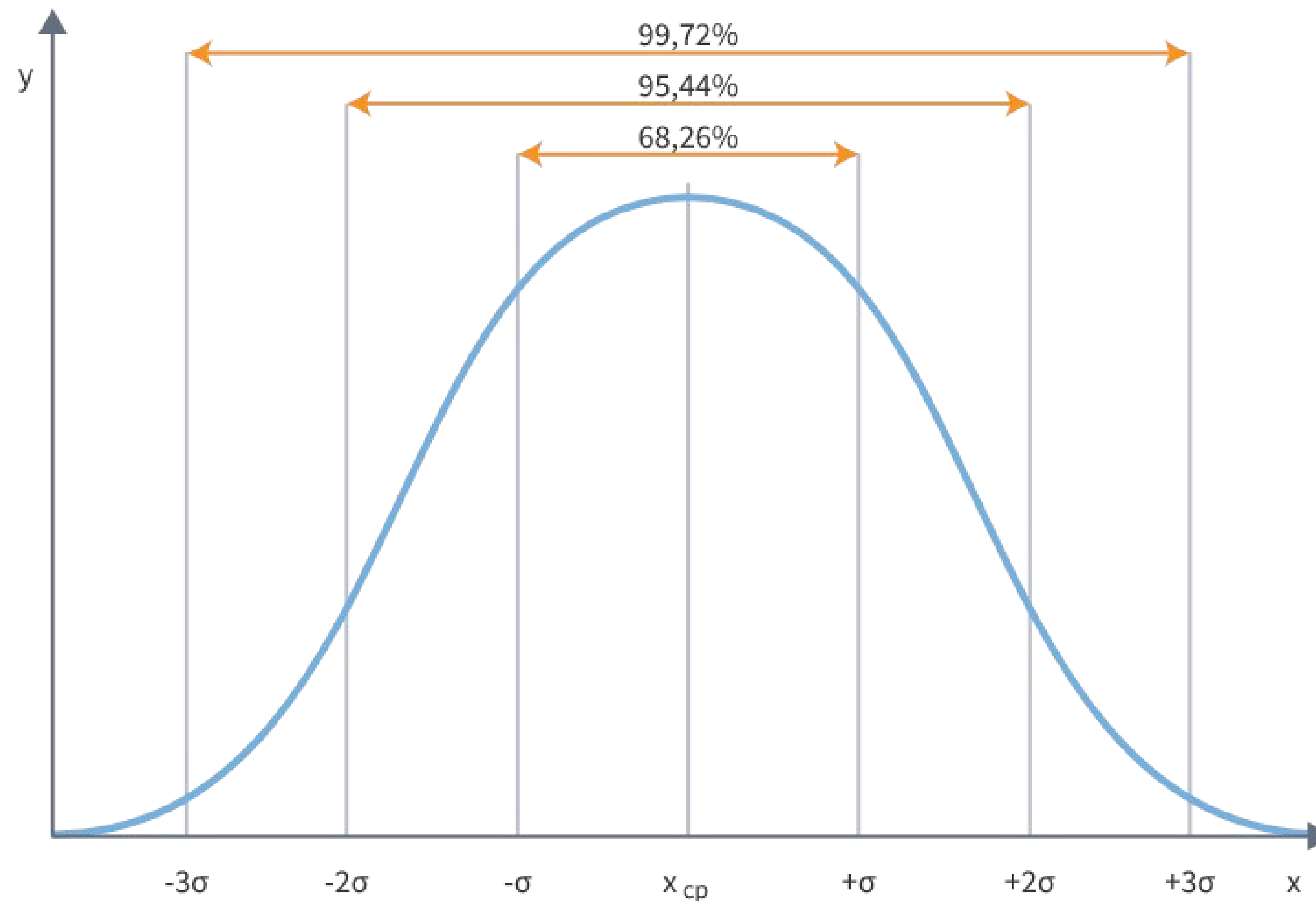
$$\Sigma = \begin{bmatrix} \sigma(X_1, X_1) & \sigma(X_1, X_2) & \dots & \sigma(X_1, X_n) \\ \sigma(X_2, X_1) & \sigma(X_2, X_2) & \dots & \sigma(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \sigma(X_n, X_1) & \sigma(X_n, X_2) & \dots & \sigma(X_n, X_n) \end{bmatrix}$$

# Свойства матрицы корреляций

Матрица корреляций симметрична.

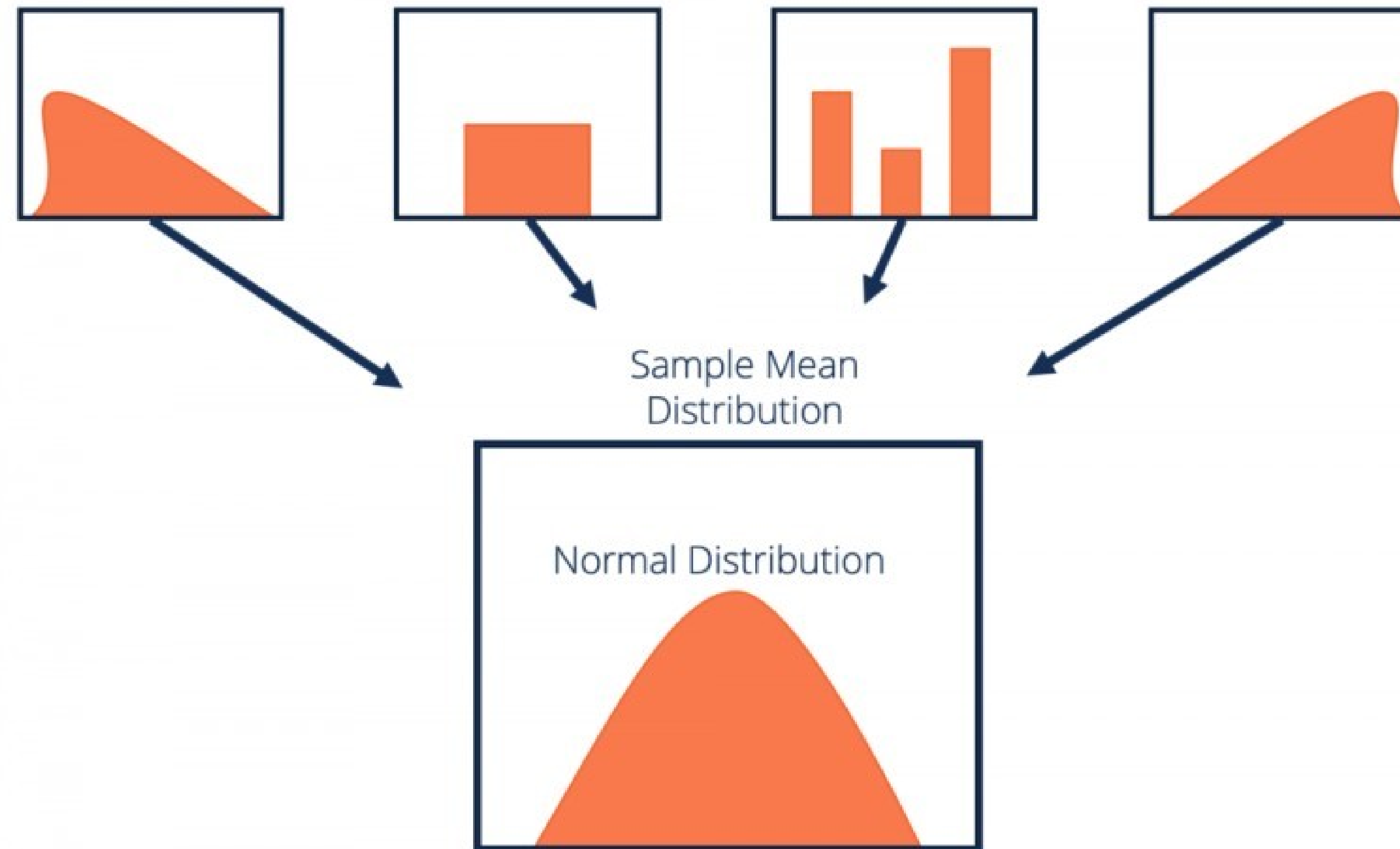
$$\Sigma = \begin{bmatrix} \sigma(X_1, X_1) & \sigma(X_1, X_2) & \dots & \sigma(X_1, X_n) \\ \sigma(X_2, X_1) & \sigma(X_2, X_2) & \dots & \sigma(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \sigma(X_n, X_1) & \sigma(X_n, X_2) & \dots & \sigma(X_n, X_n) \end{bmatrix}$$


# Правило трех сигм (нормальное распределения)



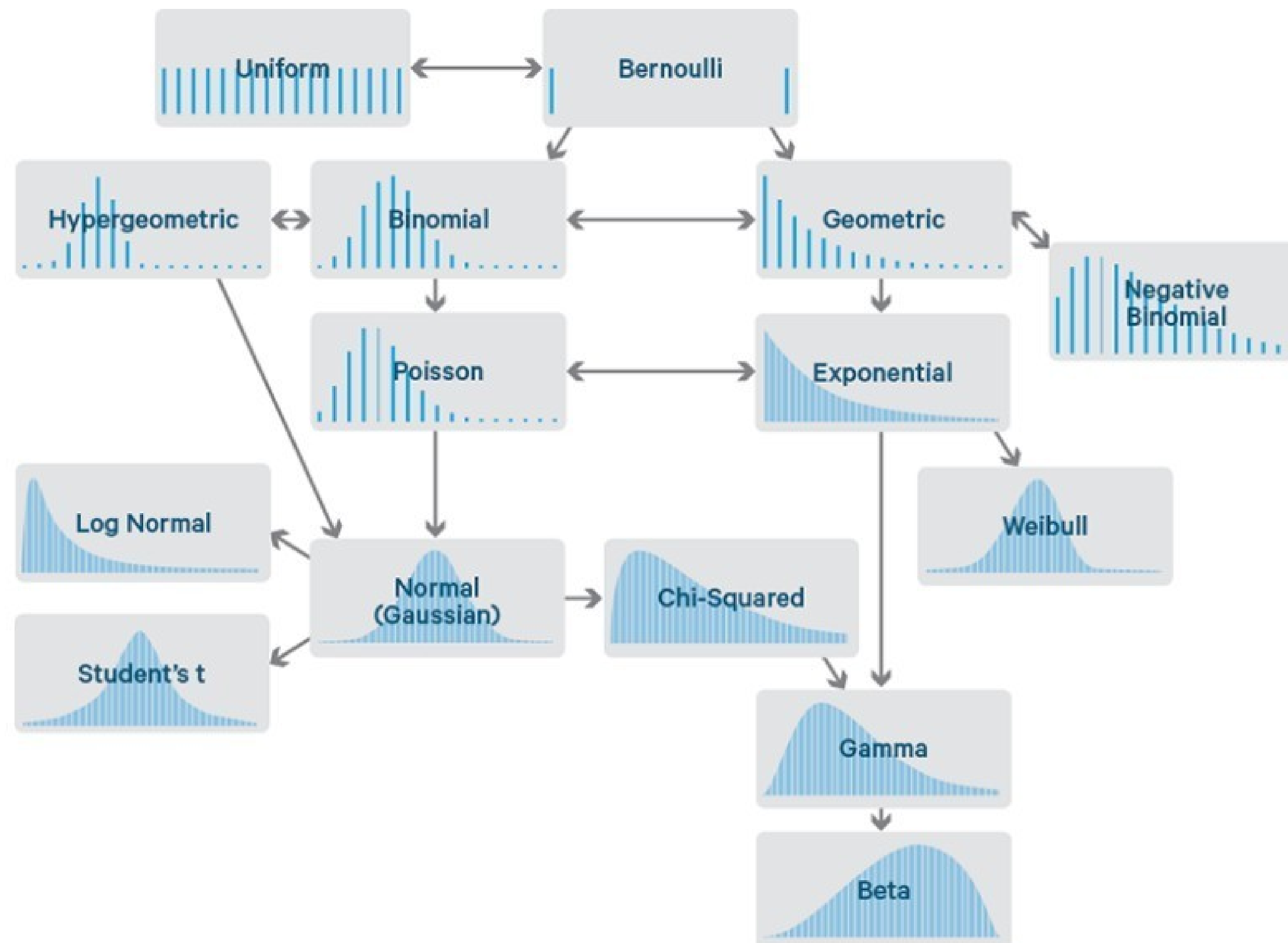


# Центральная предельная теорема



Центральные предельные теоремы (Ц. П. Т.) — класс теорем в теории вероятностей, утверждающих, что сумма достаточно большого количества слабо зависимых случайных величин, имеющих примерно одинаковые масштабы (ни одно из слагаемых не доминирует, не вносит в сумму определяющего вклада), имеет распределение, близкое к нормальному

# Виды распределений



# 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.



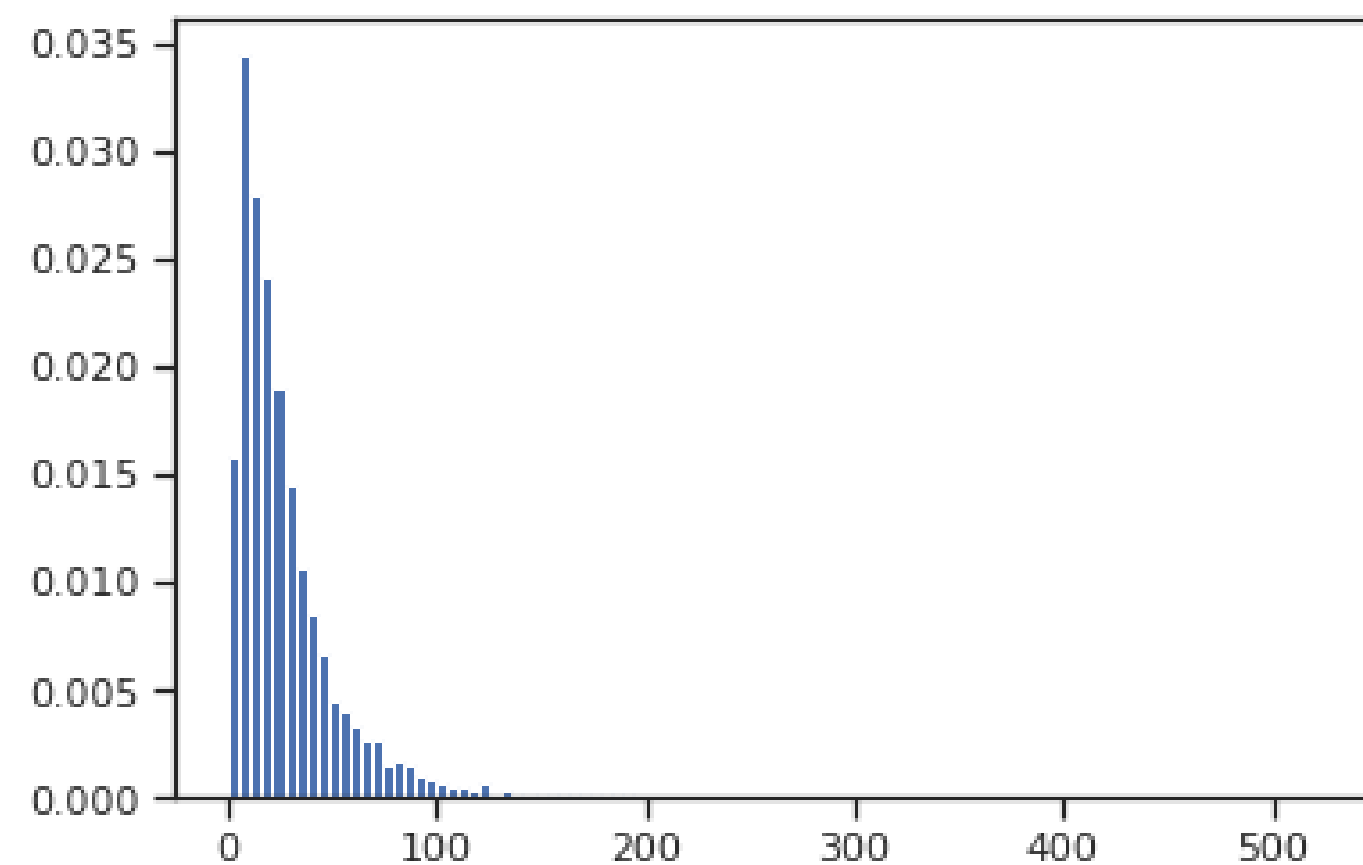
# Логнормальное распределение

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

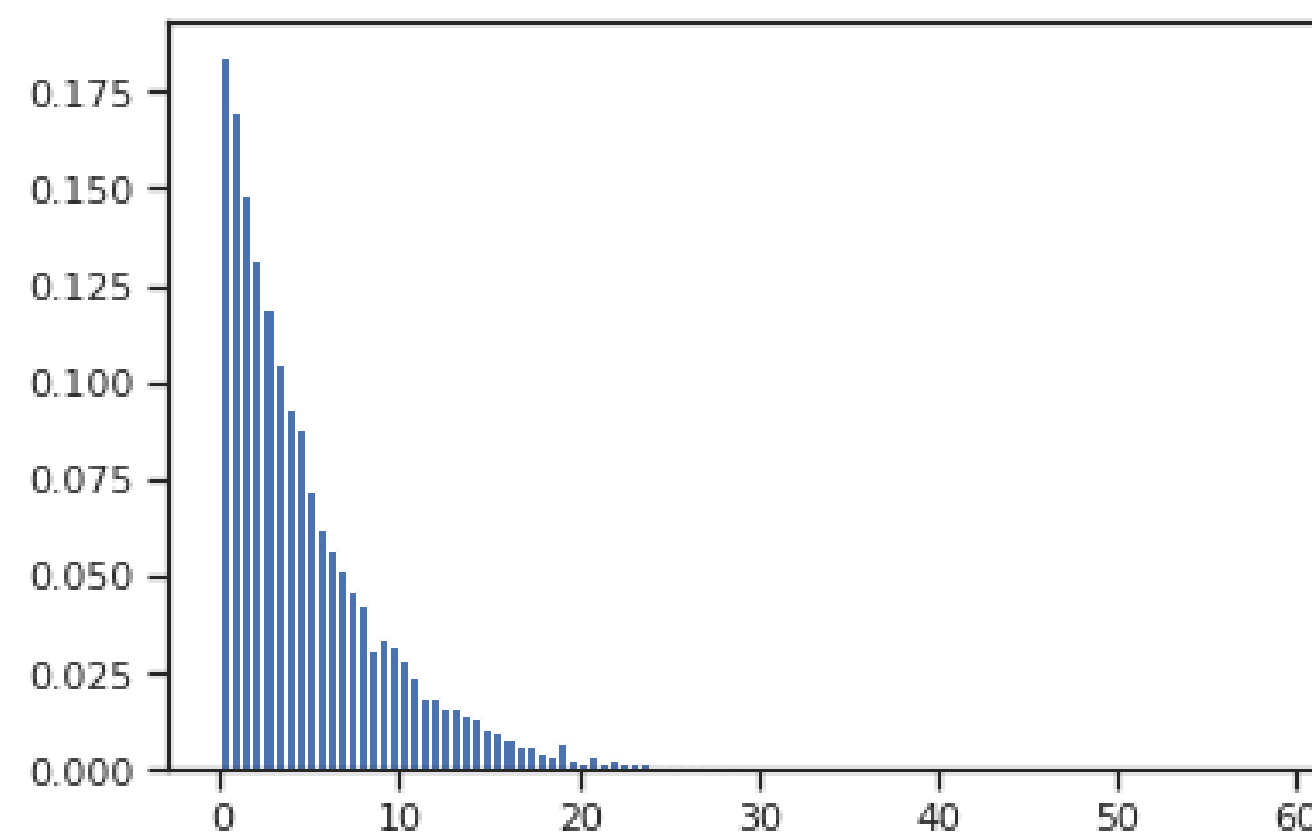
Экспоненциальное распределение задается плотностью вероятности:

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

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



# Распределение Стьюдента

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

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

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).

Спасибо за внимание. Вопросы?