

Юстина Иванова

Программист, data scientist

Центральная предельная теорема и
статистический анализ данных в python.
Виды распределений. Собственные вектора.

Спикер



Юстина Иванова,
Data scientist по
Компьютерному зрению
в компании ОЦРВ,
Выпускница МГТУ им. Баумана,
Выпускница магистратуры
University of Southampton

Нахождение зависимости случайных величин

Дисперсия — среднеквадратичное отклонение 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

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

Ковариация — это дисперсия, если две переменных — одна и та же x.

Ковариация не равна нулю — можно предположить зависимость.

Нормирование величин

Корелляция — нормированная ковариация

Корелляция — нормированная ковариация, определяет силу зависимости

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

Нормирование данных

Как нормировать:

- 1) центрирование данных;
- 3) масштабирование данных, чтобы они находились в промежутке [-1;1] или [0,1]

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).

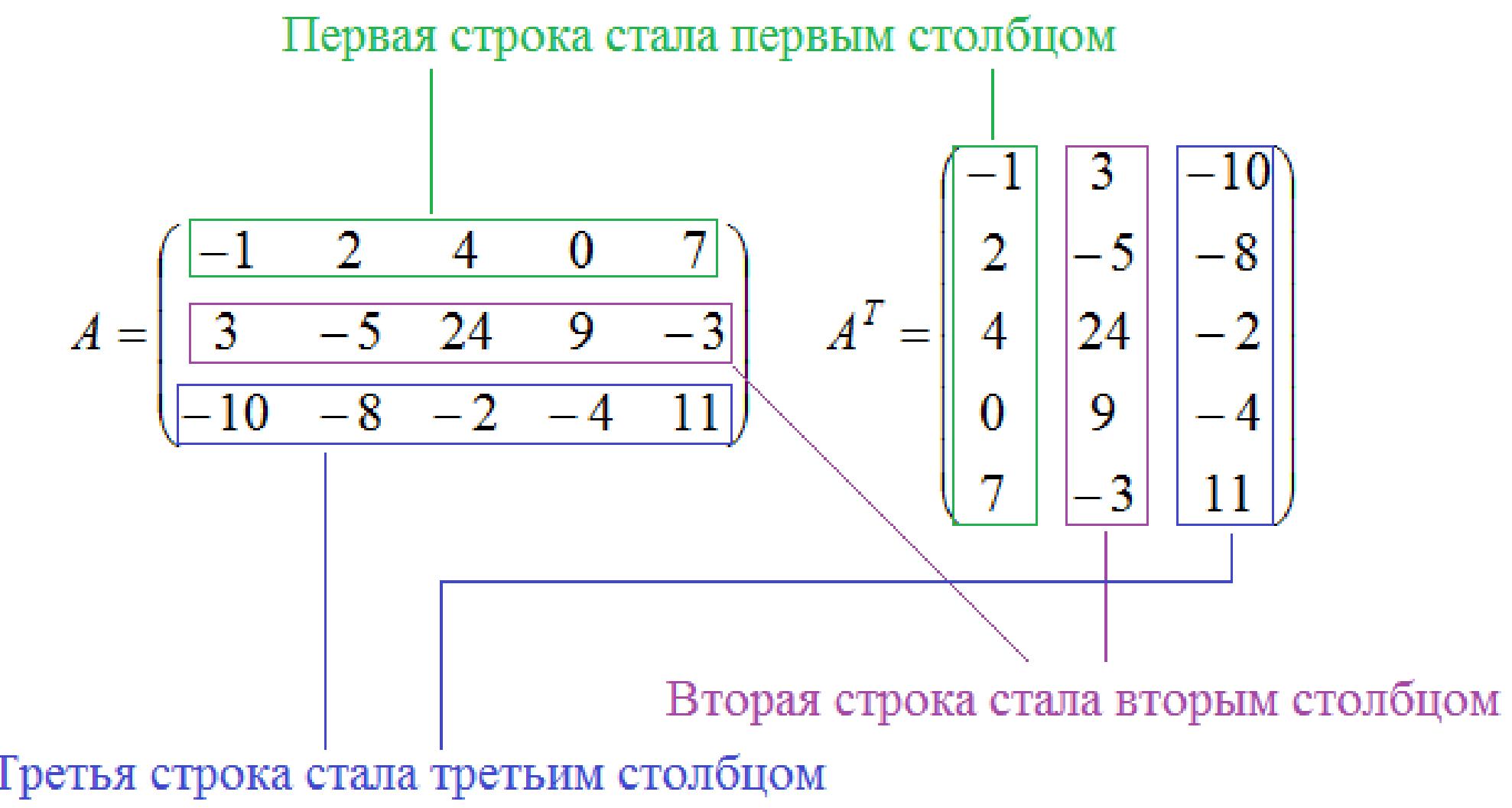
$$\Sigma = \begin{bmatrix} \sigma(X_1, X_1) & \sigma(X_1, X_2) & \dots & \sigma(X_1, X_n) \\ \sigma(X_2, X_1) & \sigma(X_2, X_2) & \dots & \sigma(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \sigma(X_n, X_1) & \sigma(X_n, X_2) & \dots & \sigma(X_n, X_n) \end{bmatrix}$$

Свойства матрицы кореляций

Матрица кореляций симметрична.

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

Транспонирование матрицы



`numpy.transpose()`
`numpy.ndarray.T()`
`numpy.matrix.transpose()`

Геометрический смысл ковариационной матрицы.

\vec{v} собств. вектор
 D данные
 $\vec{v}^T D$ проекция данных
на собств. вектор
 $\vec{v}^T \Sigma \vec{v}$ дисперсия
строимированных
данных

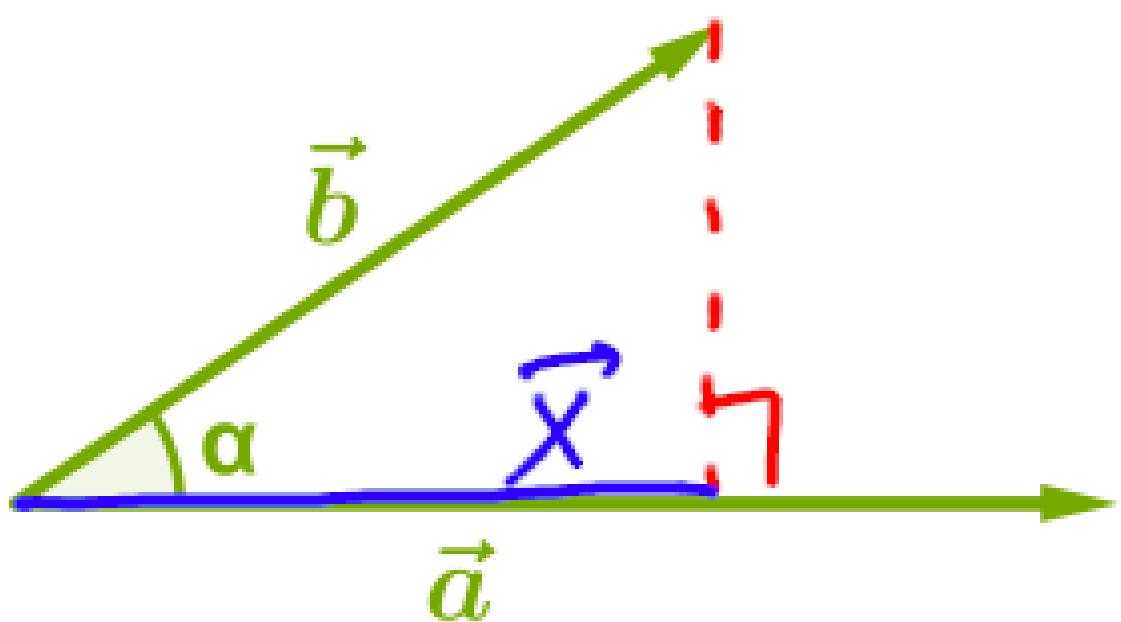
Ковариационная матрица
позволяет подсчитать
собственные вектора и
собственные значения.

Нам необходимо найти такой
вектор, при проецировании
данных на который вариация
максимальна. Этот вектор
называется собственный вектор.

Проектирование данных на вектор.

Скалярное произведение.

$$\vec{a} \cdot \vec{b} = |\vec{a}| \cdot |\vec{b}| \cdot \cos \alpha$$



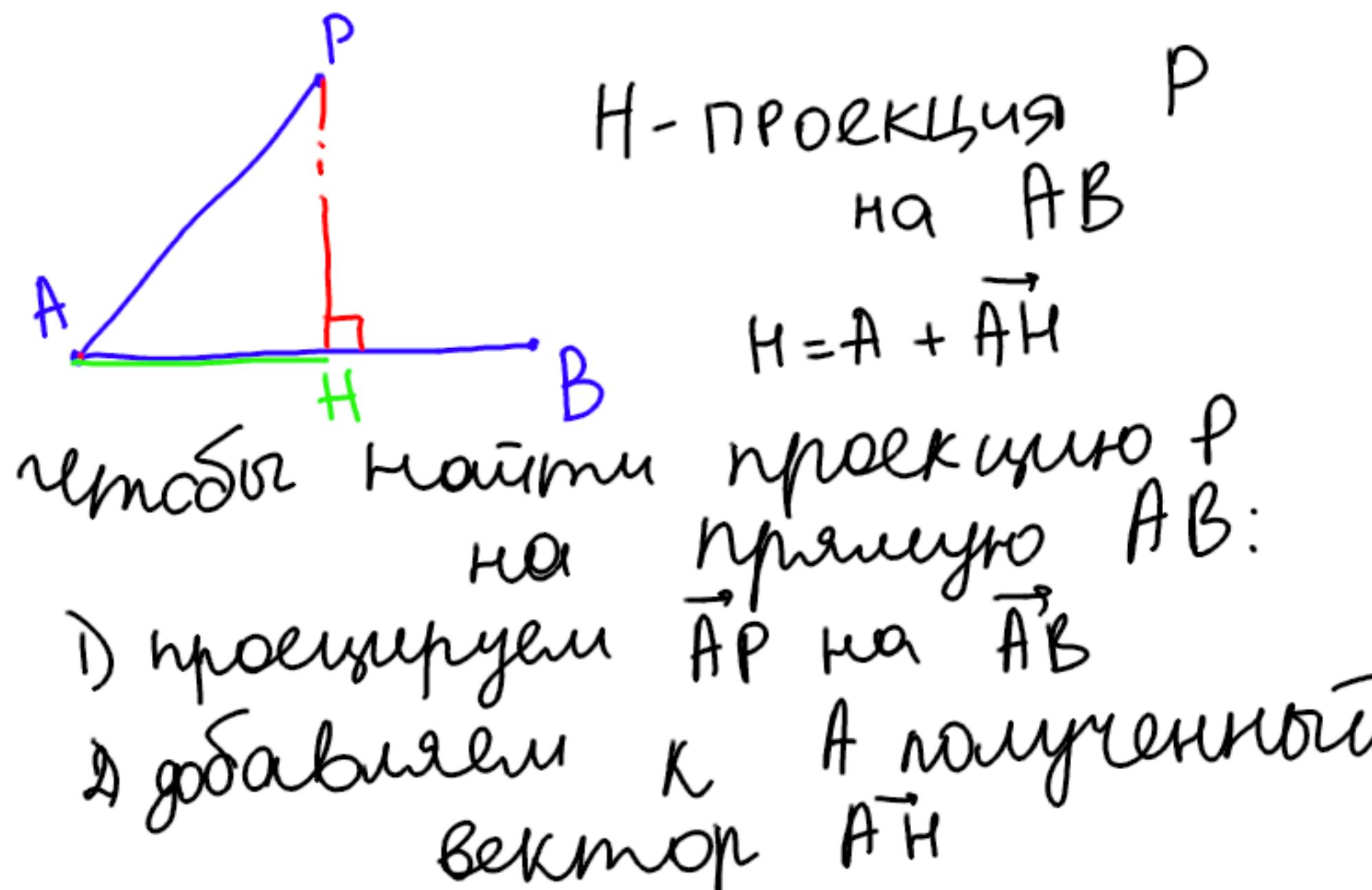
$$\frac{|\vec{x}|}{|\vec{b}|} = \cos \alpha$$

$$|\vec{x}| = |\vec{b}| \cos \alpha = |\vec{b}| \cdot \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| \cdot |\vec{b}|}$$

Формула скалярного произведения:

$$\vec{a} \cdot \vec{b} = a_1 \cdot b_1 + a_2 \cdot b_2 + \dots$$

Проектирование данных на вектор. Скалярное произведение.



Формула проекции точки P на прямую:

$$A + \text{dot}(AP, AB) / \text{dot}(AB, AB)^* AB$$

Скалярное произведение.

Скалярное произведение двух векторов можно посчитать через перемножение матриц.

Если \mathbf{A} — это матрица размера $n_1 \times k_1$, а \mathbf{B} — матрица размера $n_2 \times k_2$ и, если $k_1 = n_2$, тогда их произведение \mathbf{AB} есть матрица размера $n_1 \times k_2$, чья (i, j) -я запись имеет вид:

$$a_{i1}b_{1i} + a_{i2}b_{2i} + \dots + a_{ik}b_{ki},$$

что в точности соответствует скалярному произведению i -й строки матрицы \mathbf{A} (представленной как вектор) на j -й столбец матрицы \mathbf{B} (тоже представленной как вектор):

Скалярное произведение в python:

Numpy.dot()

Некоторые ловушки корреляции

Корреляция, равная нулю, означает отсутствие линейной связи между двумя переменными. Однако могут быть совсем другие виды зависимостей. Например, если:

```
x = [-2, -1, 0, 1, 2]  
y = [ 2,  1, 0, 1, 2]
```

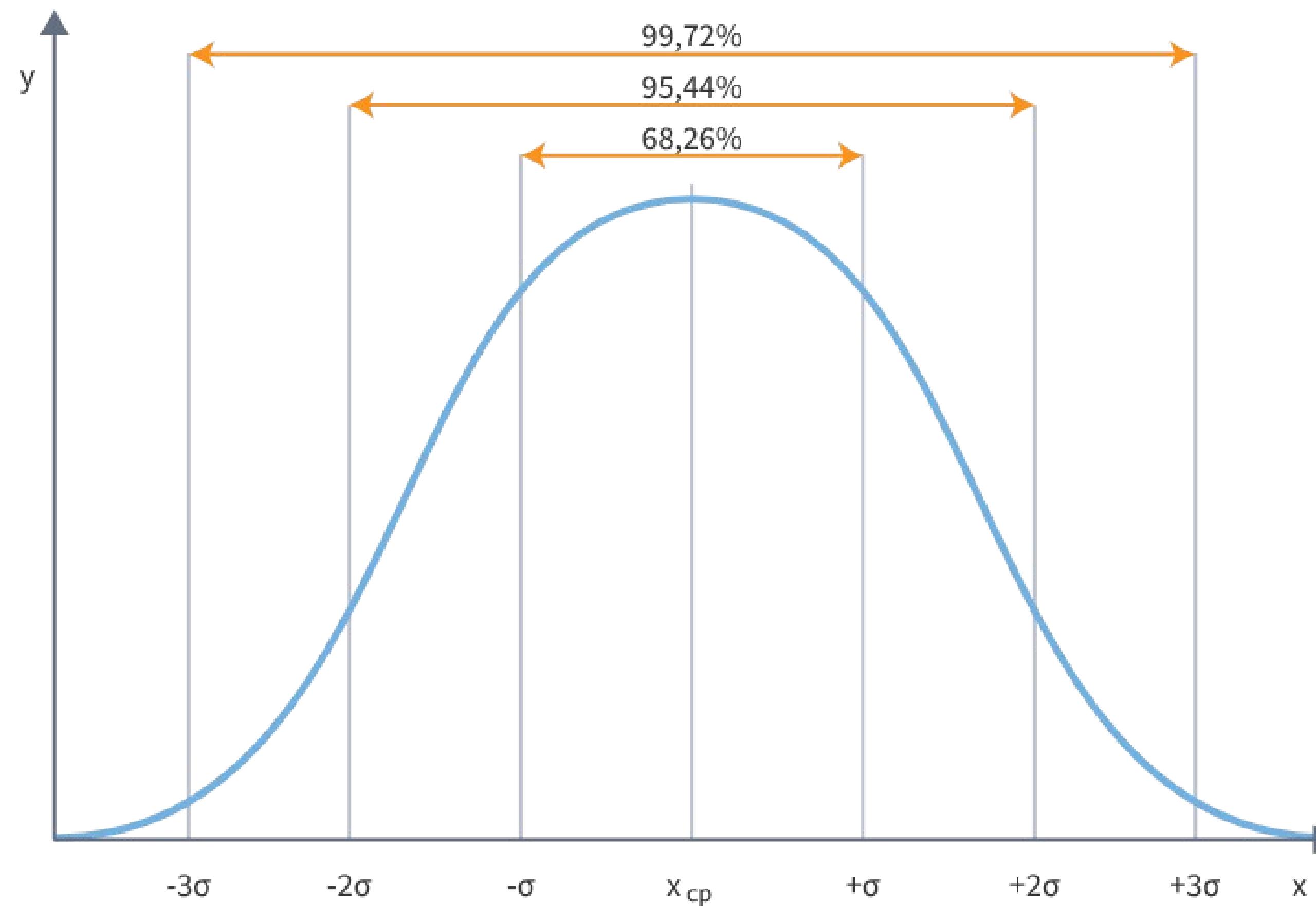
то переменные x и y имеют нулевую корреляцию, но связь между ними определенно существует — каждый элемент y равен абсолютному значению соответствующего элемента x . Но в них отсутствует связь, при которой, зная о соотношении переменной x_i со средним $\text{mean}(x)$, можно получить информацию о соотношении переменной y_i со средним $\text{mean}(y)$. Это как раз тот тип связи, который корреляция пытается установить.

Кроме того, корреляция ничего не говорит о величине связи. Переменные:

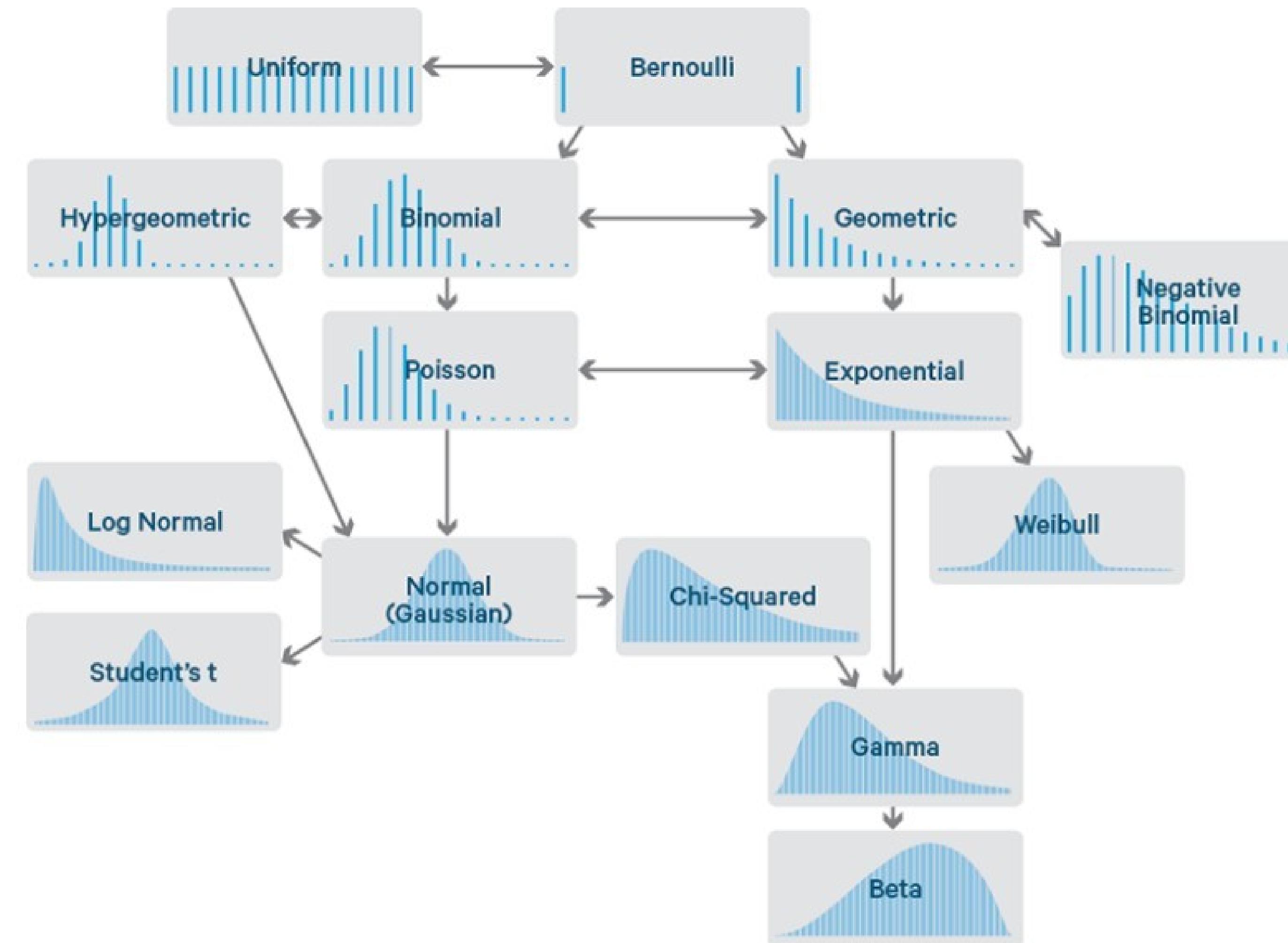
```
x = [-2, 1, 0, 1, 2]  
y = [99.98, 99.99, 100, 100.01, 100.02]
```

имеют очень хорошую корреляцию, но в зависимости от того, что вы измеряете, вполне возможно, что эта связь не представляет особого интереса.

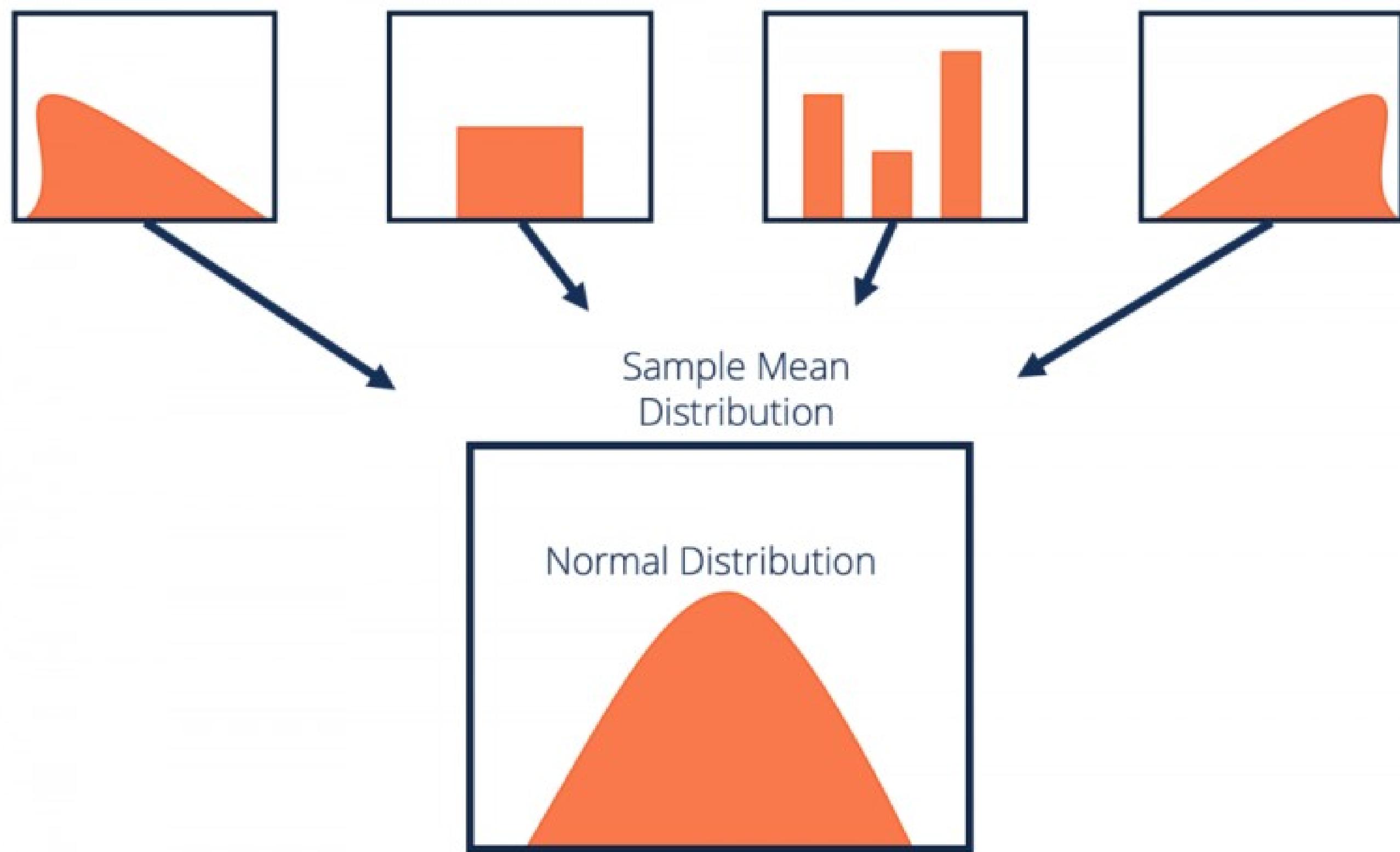
Правило трех сигм



Виды распределений



Центральная предельная теорема



Доверительные интервалы

Доверительным называют интервал, который покрывает неизвестный параметр с заданной надёжностью.

Выборочное среднее имеет нормальное распределение, если объем выборки большой, поэтому можно применить знания о нормальном распределении при рассмотрении выборочного среднего.

В частности, **95% распределения выборочных средних находится в пределах 1,96 стандартных отклонений (SD) среднего популяции.**

Когда у нас есть только одна выборка, мы называем это стандартной ошибкой среднеквадратичного отклонения (SEM) и вычисляем 95% доверительного интервала для среднего следующим образом:

$$\bar{x} - (1,96 \times SEM); \quad \bar{x} + (1,96 \times SEM).$$

Дискретные и непрерывные распределения

Различают дискретные и непрерывные вероятностные распределения. Дискретное распределение характеризуется тем, что оно сосредоточено в конечном или счетном числе точек. Непрерывное распределение "размазано" по некоторому вещественному интервалу.

Дискретное распределение:

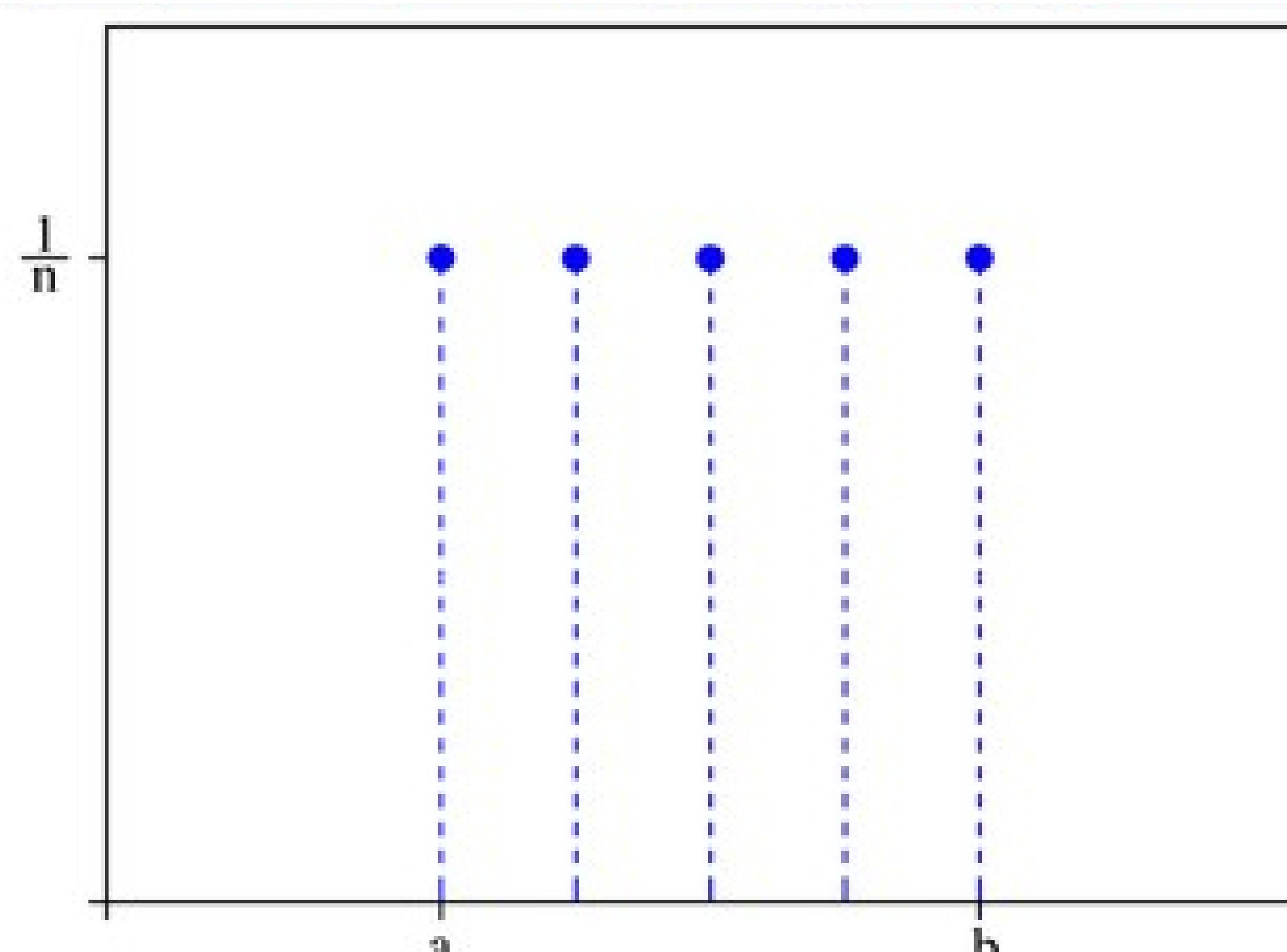
- При [подбрасывании монеты](#) случайная величина принимает значение 1, если выпал «орёл», или 0, если выпала «решка». Вероятность выпадения одного из двух значений равна $1/2$, одинакова для обоих значений, поэтому случайная величина имеет дискретное равномерное распределение.
- При бросании [игральной кости](#) случайная величина — число точек на грани принимает одно из 6-и возможных значений: {1, 2, 3, 4, 5, 6}. Вероятность выпадения одной точки из шести равна $1/6$, одинакова для каждой точки, поэтому случайная величина имеет дискретное равномерное распределение.

Дискретные и непрерывные распределения

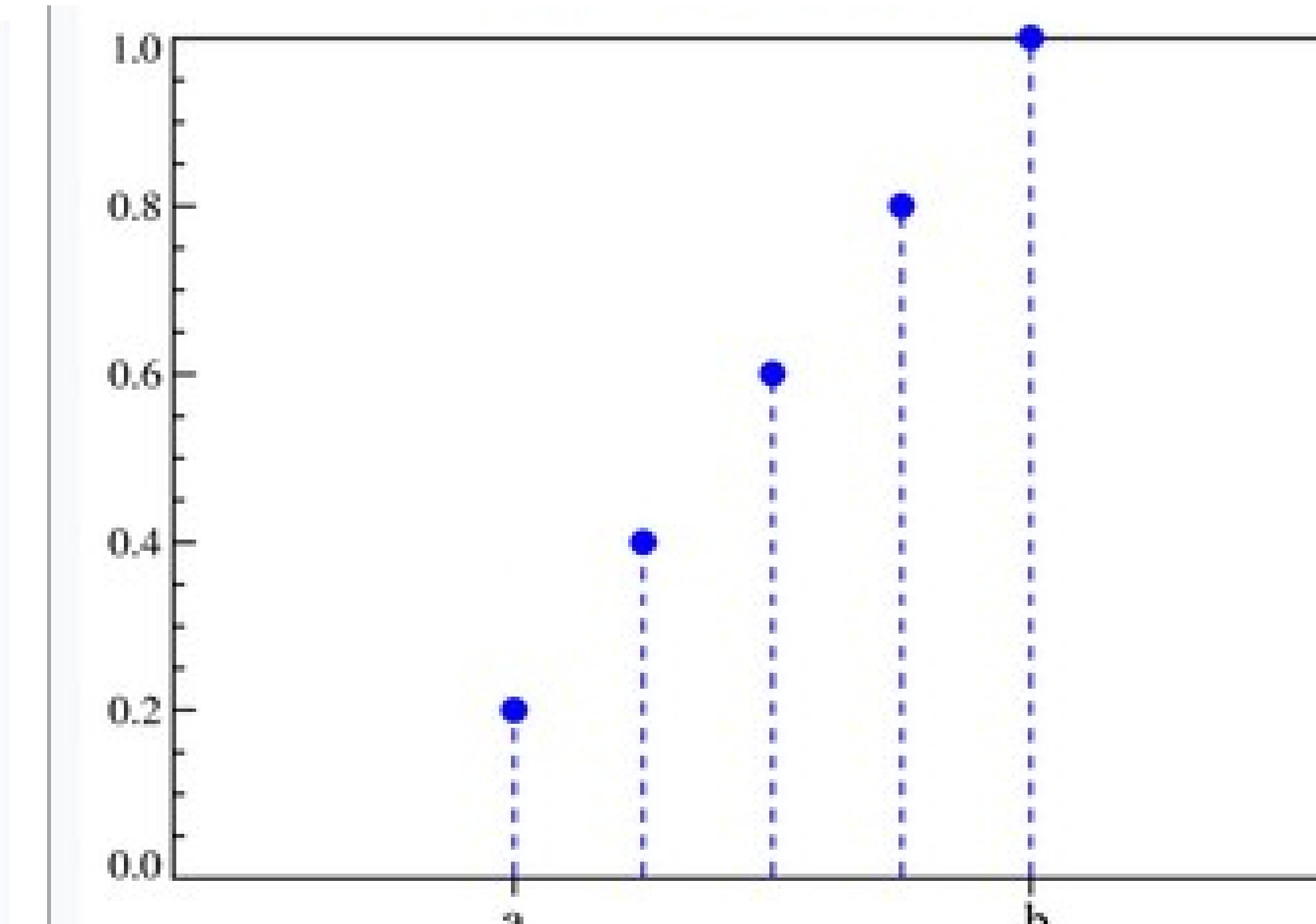
Дискретной случайной величиной называется случайная величина, которая в результате испытания принимает отдельные значения с определёнными вероятностями. Число возможных значений дискретной случайной величины может быть конечным и бесконечным. Примеры дискретной случайной величины: запись показаний спидометра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

Дискретное равномерное распределение



$n=5$, где $n=b-a+1$
Функция вероятности



$n=5$, где $n=b-a+1$.
Функция распределения

Математическое ожидание случайной величины

Математическим ожиданием (средним значением) случайной величины X , заданной на дискретном вероятностном пространстве, называется число $m=M[X]=\sum x_i p_i$, если ряд сходится абсолютно.

ПРИМЕР №1.

x_i	1	3	4	7	9
p_i	0.1	0.2	0.1	0.3	0.3

Математическое ожидание находим по формуле $m = \sum x_i p_i$.

Математическое ожидание $M[X]$.

$$M[X] = 1 \cdot 0.1 + 3 \cdot 0.2 + 4 \cdot 0.1 + 7 \cdot 0.3 + 9 \cdot 0.3 = 5.9$$

Дисперсию находим по формуле $d = \sum x_i^2 p_i - M[X]^2$.

Дисперсия $D[X]$.

$$D[X] = 1^2 \cdot 0.1 + 3^2 \cdot 0.2 + 4^2 \cdot 0.1 + 7^2 \cdot 0.3 + 9^2 \cdot 0.3 - 5.9^2 = 7.69$$

Среднее квадратическое отклонение $\sigma(x)$.

$$\sigma = \sqrt{D[X]} = \sqrt{7.69} = 2.78$$

Найти мат ожидание случайной величины

x_i	-1	2	5	10	20
p_i	0.1	0.2	0.3	0.3	0.1

Мат ожидание

Дисперсия

Среднее квадратичное отклонение

Найти мат ожидание случайной величины

x_i	-1	2	5	10	20
p_i	0.1	0.2	0.3	0.3	0.1

$$M(X) = \sum_{i=1}^n x_i \cdot p_i = -1 \cdot 0.1 + 2 \cdot 0.2 + 5 \cdot 0.3 + 10 \cdot 0.3 + 20 \cdot 0.1 = 6.8.$$

Дисперсию находим по формуле $d = \sum x_i^2 p_i - M[x]^2$.

$$D[X] = (-1)^2 \cdot 0.1 + 2^2 \cdot 0.2 + 5^2 \cdot 0.3 + 10^2 \cdot 0.3 + 20^2 \cdot 0.1 - 6.8^2 =$$

Среднее квадратическое отклонение $\sigma(x)$.

$$\sigma = \sqrt{D[X]} :$$

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 Гиннесса считало таковой использование статистического аппарата в своей работе), статья Госсета вышла в 1908 году в журнале «Биометрика» под псевдонимом «Student» (Студент).

Контакты спикера

E-mail: yustiks@gmail.com.