Теория вероятностей и статистика в Машинном Обучении

- **МОМЕНТЫ И КРИТИЧЕСКИЕ ГРАНИЦЫ СЛУЧАЙНОЙ ВЕЛИЧИНЫ**

- Меры связи случайных величин

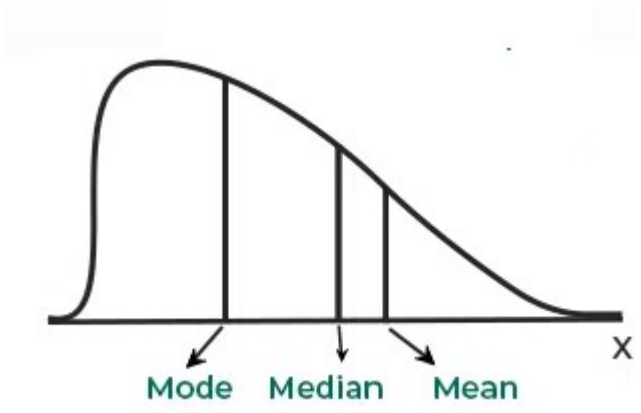
1. Коэффициент ковариации
2. Коэффициент корреляции Пирсона
3. Коэффициент корреляции Спирмена

- **ЦЕНТРАЛЬНАЯ ПРЕДЕЛЬНАЯ ТЕОРЕМА**

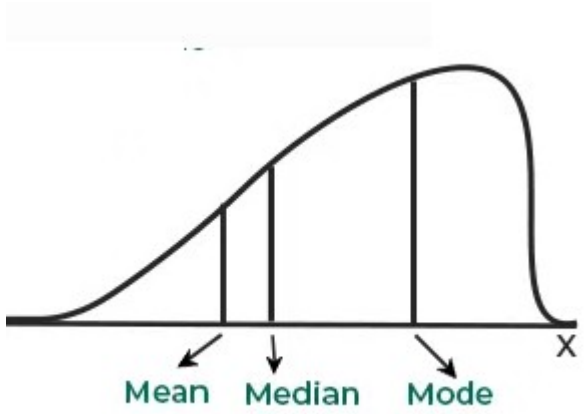
МОМЕНТЫ И КРИТИЧЕСКИЕ ГРАНИЦЫ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

- Математическое ожидание
- Дисперсия
- Асимметрия
- Эксцесс
- Квантили

Медина и мода случайной величины



Ассиметрия > 0



Ассиметрия < 0

Медина и мода случайной величины

В качестве показателя центра группирования значений случайной величины, наряду с математическим ожиданием, используются также медиана и мода.

Медианой x_{med} случайной величины X называется ее квантиль уровня 0,5 (или, что то же самое, ее 50%-ная точка):

$$x_{med} = x_{0,5} = \omega_{0,5}.$$

В качестве медианы дискретной случайной величины (поскольку квантиль дискретной случайной величины может быть определена не однозначно) обычно берут значение x_{med}, получаемое линейной аппроксимацией:

$$x_{med} = x_I + \frac{x_{I+1} - x_I}{p_{I+1}}\left(\frac{1}{2} - \sum_{i=1}^{I} p_i\right),$$

где *медианный отрезок* $[x_I; x_{I+1}]$ определяется из условий

$$\sum_{i=1}^{I} p_i \leqslant 0{,}5; \qquad \sum_{i=1}^{I+1} p_i > 0{,}5.$$

Модой абсолютно непрерывной случайной величины X называется точка локального максимума плотности распределения:

$$f_X(x_{mod}) = \max_{x \in \bar{Y}} f_X(x)$$

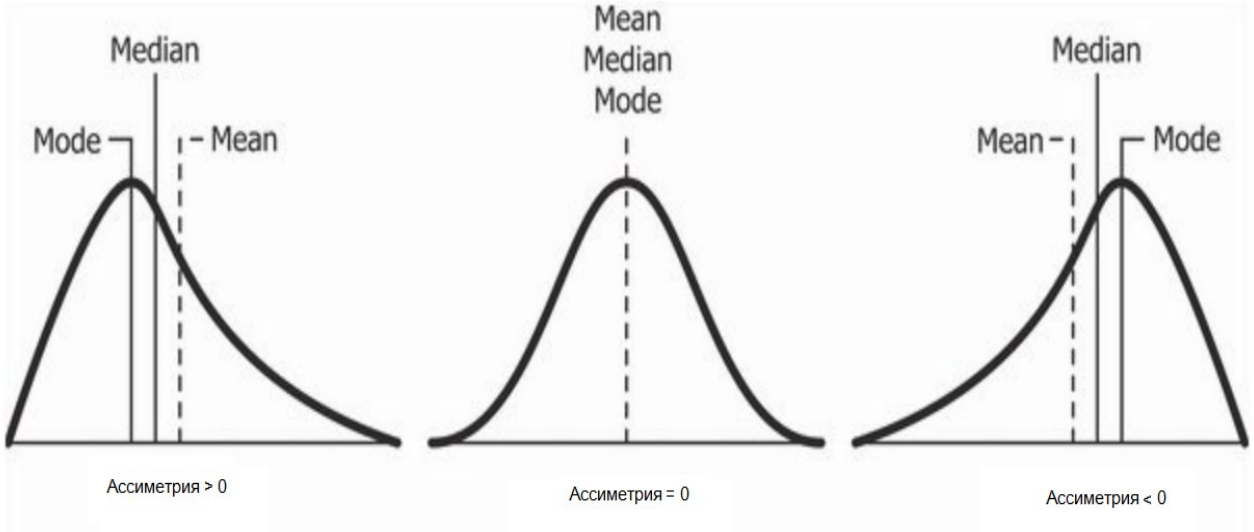
Модой дискретной случайной величины X называется значение этой случайной величины, соответствующее наибольшей вероятности:

$$x_{mod} = x_i, \text{ такое что } p_i = \max_j p_j.$$

Случайные величины, имеющие одну моду, называются *унимодальными*.

Бимодальное

Мультимодальное



Сумма под риском

Пусть случайная величина X вновь имеет смысл дохода от инвестиционной операции, которая продолжается в течение некоторого периода T, и мы хотим оценить риск этой операции.

В практической деятельности для этого часто рассчитывают показатель суммы под риском (*Value at Risk*, **VaR**): *$100\gamma\%$-ной суммой под риском* данной инвестиционной операции за период T называется такое число $\mathbf{VaR}_\gamma$, что с вероятностью γ убыток $(-X)$ от проведения операции в течение данного периода будет не больше, чем $\mathbf{VaR}_\gamma$:

$$\mathbf{P}\{-X \leqslant \mathbf{VaR}_\gamma\} = \gamma$$

или

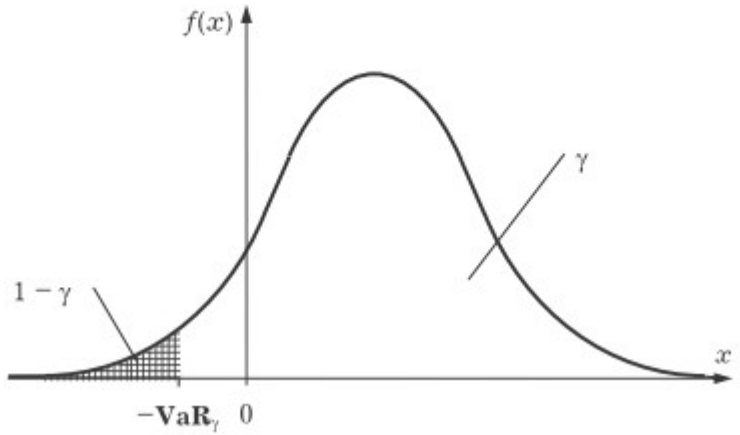
$$\mathbf{P}\{X \geqslant -\mathbf{VaR}_\gamma\} = \gamma.$$

$$-\mathbf{VaR}_\gamma = \omega_\gamma$$

или, окончательно,

$$\mathbf{VaR}_\gamma = -\omega_\gamma = -u_{1-\gamma},$$

$100\gamma\%$-ная сумма под риском — это число, противоположное $100\gamma\%$-ной точке случайной величины X

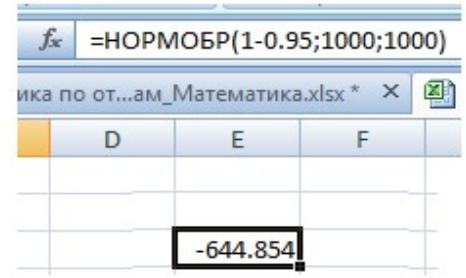


Текущая цена акции равна 15 руб. Стоимость акции через семь дней описывается случайной величиной $Y = \mathcal{N}(16;\ 1)$. Рассчитать недельную 95%-ную сумму под риском операции по покупке пакета 1000 таких акций и его последующей продаже через семь дней.

Доход от данной операции $X = 1000(Y - 15)$.
X будет при этом распределен по нормальному закону с параметрами $a = 1000$, $\sigma = 1000$.

Найдем 95%-ную точку случайной величины X.
$\omega_{0,95} = \text{НОРМ.ОБР}(1 - 0{,}95;\ 1000;\ 1000) = -644{,}85,$

поэтому $\mathbf{VaR}_{0,95} = -\omega_{0,95} = 644{,}85$ ден. ед.



МЕРЫ СВЯЗИ СЛУЧАЙНЫХ ВЕЛИЧИН

Для измерения «степени зависимости» случайных величин *ковариация* случайных величин X и Y:

$$\operatorname{cov}(X, Y) = \mathbf{E}\{[X - \mathbf{E}(X)][Y - \mathbf{E}(Y)]\}.$$

$$\operatorname{cov}(X, Y) = \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y).$$

Коммутативность ковариации. *Для любых случайных величин X и Y*

$$\operatorname{cov}(X, Y) = \operatorname{cov}(Y, X).$$

Связь ковариации и дисперсии. *Для любой случайной величины X*

$$\operatorname{cov}(X, X) = \mathbf{Var}(X).$$

Линейность ковариации по каждому аргументу. *Для любых случайных величин X и Y и любых чисел $a,\ b \in \mathbb{R}$*

$$\operatorname{cov}(aX, bY) = ab \operatorname{cov}(X, Y).$$

Аддитивность ковариации по каждому агрументу. *Для любых случайных величин $X,\ Y$ и Z*

$$\operatorname{cov}(X + Y, Z) = \operatorname{cov}(X, Z) + \operatorname{cov}(Y, Z),$$
$$\operatorname{cov}(X,\ Y + Z) = \operatorname{cov}(X, Y) + \operatorname{cov}(X, Z).$$

Формула для ковариации независимых случайных величин. *Для независимых случайных величин X и Y*

$$\operatorname{cov}(X, Y) = 0.$$

Формула для дисперсии суммы произвольных случайных величин. *Дисперсия суммы произвольных (зависимых или независимых) случайных величин X и Y рассчитывается по формуле*

$$\mathbf{Var}(X + Y) = \mathbf{Var}(X) + \mathbf{Var}(Y) + 2\operatorname{cov}(X, Y).$$

две случайные величины X и Y называются *независимыми*, если для всех $x,\ y \in \mathbb{R}$

$$\mathbf{P}\{(X \leqslant x) \cap (Y \leqslant y)\} = \mathbf{P}\{X \leqslant x\}\mathbf{P}\{Y \leqslant y\},$$

т. е. если для всех $x,\ y \in \mathbb{R}$ события $\{X \leqslant x\}$ и $\{Y \leqslant y\}$ независимы.

Коэффициент корреляции

Коэффициент корреляции случайных величин X и Y

$$R(X,Y)=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{\mathbf{E}\{[X-\mathbf{E}(X)][Y-\mathbf{E}(Y)]\}}{\sigma_X\sigma_Y}=\frac{\mathbf{E}(XY)-\mathbf{E}(X)\mathbf{E}(Y)}{\sigma_X\sigma_Y}.$$

Инвариантность коэффициента корреляции относительно линейных преобразований аргументов. *Для любых случайных величин X и Y и любых чисел a, b, c, $d \in \mathbb{R}$*

$$R(aX+b, cY+d) = R(X, Y).$$

Нормированность коэффициента корреляции. *Для любых случайных величин X и Y*

$$-1 \leqslant R(X, Y) \leqslant 1.$$

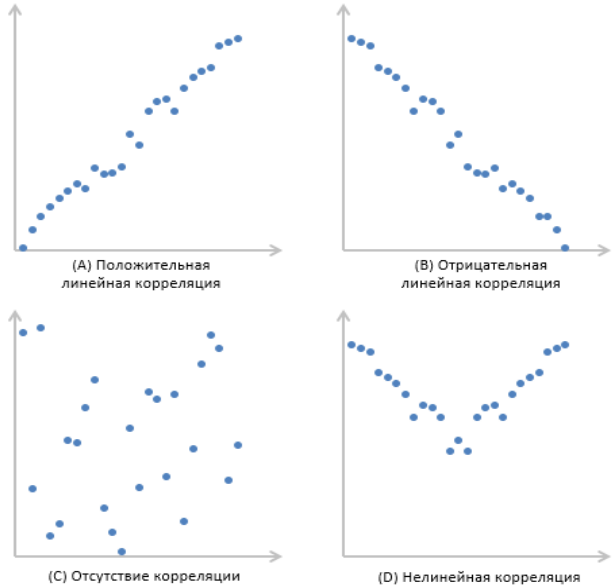
Формула для коэффициента корреляции независимых случайных величин. *Для независимых случайных величин X и Y*

$$R(X, Y) = 0;$$

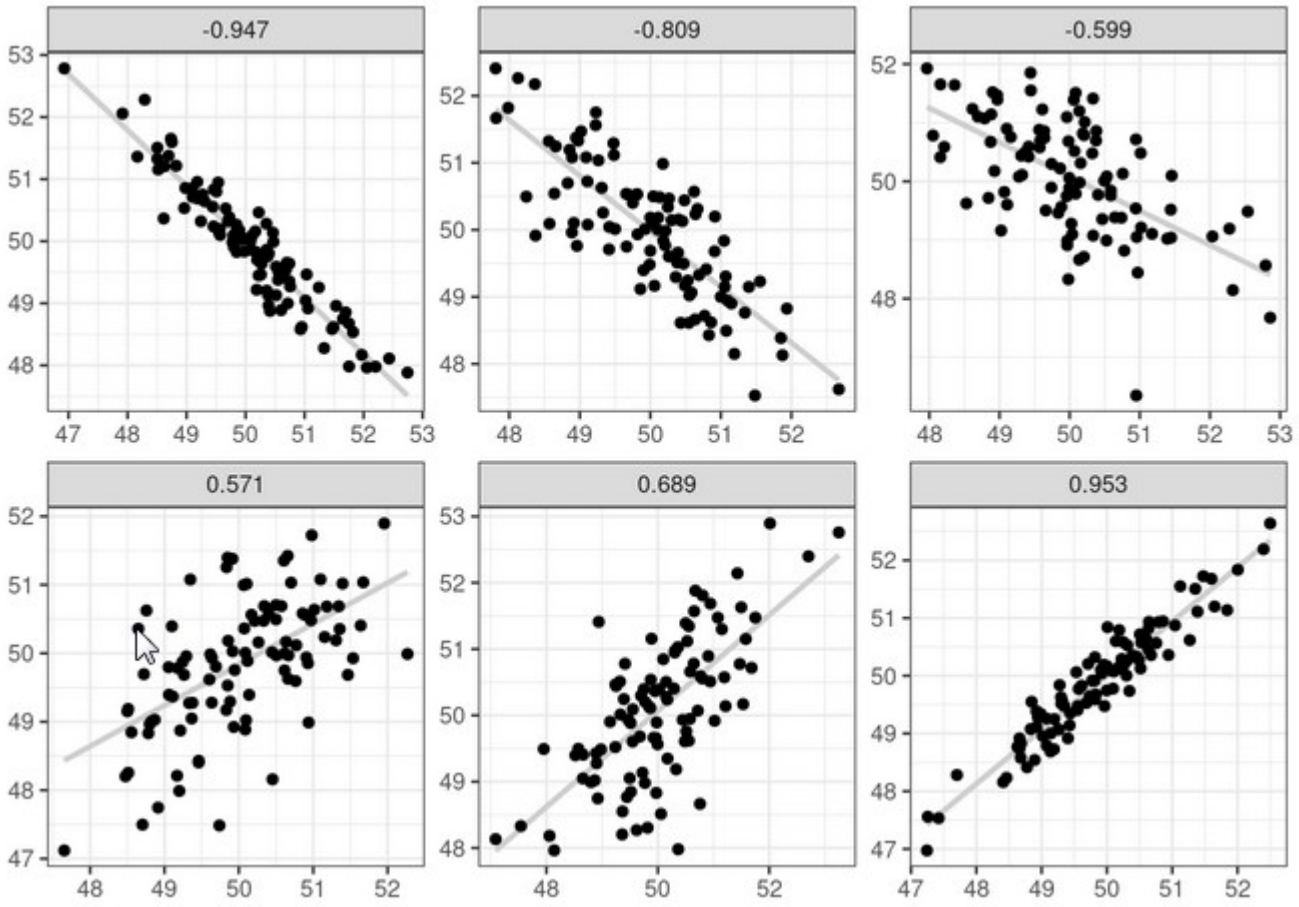
Формула для коэффициента корреляции линейно связанных случайных величин. *Для линейно связанных случайных величин X и $Y = aX + b$ (a, $b \in \mathbb{R}$, a 0)*

$$|R(X, Y)| = 1, \qquad (\star)$$

если для каких-либо случайных величин X и Y выполняется ($\star$), то существуют такие числа a, $b \in \mathbb{R}$, $a \neq 0$, что $\mathbf{P}\{Y = aX + b\} = 1$.



Примеры зависимых случайных величин



Интерпретация коэффициента корреляции производится исходя из уровня силы связи:

[0,2; 0,3) – **слабая** положительная связь,
[0,30; 0,7) – **умеренная** положительная связь,
[0,70; 1,0] – **сильная** положительная связь,

(-0,3; -0,2]– **слабая** отрицательная связь,
(-0,7; -0,3]; – **умеренная** отрицательная связь,
[-1,0; -0,7] – **сильная** отрицательная связь.

Вычисление коэффициента корреляции (Пирсона).

Применение

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Excel: **КОРРЕЛ(массив1;массив2)**

Коэффициент детерминации = $R2 = 1 - \frac{D[y|x]}{D[y]} = 1 - \frac{\sigma^2}{\sigma_y^2}$

обычно применяется при оценке качества **регрессионных моделей**

Доля дисперсии зависимой переменной, объясняемая рассматриваемой моделью зависимости (объясняющими переменными), или
1- *доля необъяснённой дисперсии*

1. Коэффициент корреляции Пирсона чувствителен к выбросам. Одно аномальное значение может существенно исказить коэффициент. Поэтому перед проведением анализа следует проверить и при необходимости удалить выбросы. Другой вариант – перейти к ранговому коэффициенту корреляции Спирмена. Рассчитывается также, только не по исходным значениям, а по их рангам

2. Синоним корреляции – это взаимосвязь или совместная вариация. Поэтому наличие корреляции ($r \neq 0$) еще не означает причинно-следственную связь между переменными. Вполне возможно, что совместная вариация обусловлена влиянием третьей переменной. Совместное изменение переменных без причинно-следственной связи называется **ложная корреляция**.

3. Отсутствие линейной корреляции ($r = 0$) не означает отсутствие взаимосвязи. Она может быть нелинейной. Частично эту проблему решает ранговая корреляция Спирмена, которая показывает совместный рост или снижение рангов, независимо от формы взаимосвязи.

Ранговый коэффициент корреляции

Коэффициент корреляции Спирмена:

$$r_s = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)} \quad (*)$$

$$r_s = \rho_{\mathrm{rg}_X, \mathrm{rg}_Y} = \frac{\mathrm{cov}(\mathrm{rg}_X, \mathrm{rg}_Y)}{\sigma_{\mathrm{rg}_X} \sigma_{\mathrm{rg}_Y}},$$

Расчет коэффициента Спирмена состоит из следующих этапов:

1) Ранжирование признаков по возрастанию.

Ранг – это порядковый номер. Если встречаются два одинаковых значения, им присваивают одинаковое значение ранга.

2) Определение разности рангов каждой пары сопоставляемых значений, d = dx – dy.

3) Возведение в квадрат разность di и нахождение общей суммы, ∑d2.

4) Вычисление коэффициента корреляции рангов по формуле (*)

Шкала Чеддока

Количественная мера тесноты связи	Качественная характеристика силы связи
0,1 - 0,3	Слабая
0,3 - 0,5	Умеренная
0,5 - 0,7	Заметная
0,7 - 0,9	Высокая
0,9 - 0,99	Весьма высокая

ЦЕНТРАЛЬНАЯ ПРЕДЕЛЬНАЯ ТЕОРЕМА

В ряде задач приходится сталкиваться с ситуацией, когда исследуемая случайная величина является суммой большого числа независимых слагаемых, влияние каждого из которых на сумму очень мало. На основании центральной предельной теоремы часто можно до наблюдения того или иного явления сказать, что соответствующая случайная величина должна иметь нормальное распределение или близкое к нему.

Теорема Леви. *Если независимые случайные величины* $X_1, X_2, \ldots, X_n, \ldots$ *распределены по одному и тому же закону с математическим ожиданием* a *и средним квадратическим отклонением* σ, *то при* $n \to \infty$ *функция распределения случайной величины*

$$Z_n = \frac{\sum_{i=1}^{n} (X_i - a)}{\sigma\sqrt{n}}$$

сходится (в определенном смысле) к функции распределения стандартной нормальной случайной величины $\mathcal{N}(0;1)$. Это записывается так:

$$Z_n \Rightarrow \mathcal{N}(0;1).$$