Лекция 9. Ковариация и коэффициент корреляции. Их свойства

Определение 1. *Ковариацией случайных величин X и Y называется число*

$$\operatorname{Cov}(X,Y) = M\left[(X - M[X])(Y - M[Y])\right],$$

если математические ожидания справа существуют.

Правую часть последнего равенства можно преобразовать к виду

$$M[XY] - M[X] \cdot M[Y]$$

так что

$$\operatorname{Cov}(X,Y) = M[XY] - M[X] \cdot M[Y].$$

Определение 2. *Если $\operatorname{Cov}(X,Y) = 0$, то случайные величины X и Y называются некоррелированными.*

Заметим, что если X и Y — независимы, то они некоррелированы.
Непосредственно из определения следуют свойства ковариации:

1. $\operatorname{Cov}(\lambda X, Y) = \operatorname{Cov}(X, \lambda Y) = \lambda \cdot \operatorname{Cov}(X,Y)$, где λ — постоянная;
2. $\operatorname{Cov}(X+Y, Z) = \operatorname{Cov}(X,Z) + \operatorname{Cov}(Y,Z)$;
3. $\operatorname{Cov}(X, Y+Z) = \operatorname{Cov}(X,Y) + \operatorname{Cov}(X,Z)$;
4. $\operatorname{Cov}(X,X) = D[X]$, т.е. дисперсию формально можно рассматривать как ковариацию между случайной величиной и ей самой.

Найдем дисперсию суммы двух произвольных случайных величин X и Y:

$$\begin{aligned} D[X+Y] &= M\left[(X+Y-M[X+Y])^2\right] = M\left[((X-M[X]) + (Y-M[Y]))^2\right] \\ &= M\left[(X-M[X])^2\right] + M\left[(Y-M[Y])^2\right] + 2M\left[(X-M[X])(Y-M[Y])\right] \\ &= D[X] + D[Y] + 2\operatorname{Cov}(X,Y). \end{aligned}$$

В случае суммы n случайных величин аналогично устанавливается, что

$$D\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} D[X_i] + 2\sum_{i<j} \operatorname{Cov}(X_i, X_j).$$

В частности, если все дисперсии $D[X_i]$ и коэффициенты ковариации $\operatorname{Cov}(X_i, X_j)$ одинаковы:

$$D[X_i] = d \quad \forall i = 1, \ldots, n, \qquad \operatorname{Cov}(X_i, X_j) = K \quad \forall i \neq j,$$

то

$$D\left[\sum_{i=1}^{n} X_i\right] = nd + n(n-1)K.$$

Определение 3. *Коэффициентом корреляции случайных величин X и Y называется число*

$$\rho_{XY} = \frac{\operatorname{Cov}(X,Y)}{\sigma_X \sigma_Y},$$

где $\sigma_X^2 = D[X]$, $\sigma_Y^2 = D[Y]$. (ρ_{XY} определён только при $\sigma_X > 0$ и $\sigma_Y > 0$.)

Утверждение 1. $|\rho_{XY}| \leq 1$.

Доказательство. Рассмотрим дисперсию суммы:

$$D[X+Y] = D[X] + D[Y] + 2\rho_{XY}\sqrt{D[X]}\sqrt{D[Y]}.$$

Отсюда

$$\frac{D[X+Y]}{\sqrt{D[X]}\sqrt{D[Y]}} = \frac{\sqrt{D[X]}}{\sqrt{D[Y]}} + \frac{\sqrt{D[Y]}}{\sqrt{D[X]}} + 2\rho_{XY}. \tag{1}$$

Пусть $t = \frac{\sqrt{D[X]}}{\sqrt{D[Y]}} \in (0, +\infty)$. Из (1) следует, что каково бы ни было $t \in (0, +\infty)$, выполняется

$$t + \frac{1}{t} + 2\rho_{XY} \geq 0 \quad \Rightarrow \quad \rho_{XY} \geq -\frac{1}{2}\left(t + \frac{1}{t}\right).$$

Минимум правой части равен -1, откуда $\rho_{XY} \geq -1$.

Аналогично, рассмотрим дисперсию разности:

$$D[X-Y] = D[X] + D[Y] - 2\rho_{XY}\sqrt{D[X]}\sqrt{D[Y]}.$$

Тогда

$$\frac{D[X-Y]}{\sqrt{D[X]}\sqrt{D[Y]}} = \frac{\sqrt{D[X]}}{\sqrt{D[Y]}} + \frac{\sqrt{D[Y]}}{\sqrt{D[X]}} - 2\rho_{XY}. \tag{2}$$

Из (2) следует, что

$$t + \frac{1}{t} - 2\rho_{XY} \geq 0 \quad \Rightarrow \quad \rho_{XY} \leq \frac{1}{2}\left(t + \frac{1}{t}\right).$$

Минимум правой части равен 1, откуда $\rho_{XY} \leq 1$.

Таким образом, $|\rho_{XY}| \leq 1$. □

Отметим также следующие свойства коэффициента корреляции:

1. ρ_{XY} — безразмерная величина;
2. Если случайные величины X и Y связаны линейной зависимостью $Y = \alpha X + \beta$, где $\alpha \neq 0$, β — произвольные постоянные, то

$$\rho_{XY} = \begin{cases} 1, & \alpha > 0, \\ -1, & \alpha < 0. \end{cases}$$

3. Если X и Y независимые случайные величины, то $\rho_{XY} = 0$.

На содержательном уровне величина коэффициента ρ_{XY} характеризует тесноту связи между случайными величинами X и Y: чем меньше $|\rho_{XY}|$, тем связь слабее. При этом знак ρ_{XY} допускает следующую интерпретацию: отрицательные значения ρ_{XY} означают, как правило, что увеличению Y соответствует уменьшение X и, наоборот, уменьшению Y соответствует увеличение X (такое положение вещей иногда характеризуют словами "X и Y изменяются в противофазе");

положительные значения ρ_{XY} означают, как правило, что в среднем увеличению Y соответствует увеличение и X, а уменьшению Y соответствует и уменьшение X ("X и Y изменяются в фазе, или синфазно").

4. Из некоррелированности случайных величин, вообще говоря, не следует их независимость.

   Покажем это на примере. Пусть X — центрированная (т.е. $M[X] = 0$) нормально распределённая случайная величина:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/(2\sigma^2)}, \quad x \in (-\infty, +\infty).$$

   Пусть $Y = X^2$. Тогда

$$\mathrm{Cov}(X,Y) = M[XY] - M[X] \cdot M[Y] = M\left[X^3\right] - 0 \cdot M[Y] =$$

$$= \int_{-\infty}^{+\infty} x^3 \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/(2\sigma^2)} dx = 0,$$

   так что $\rho_{XY} = 0$. Однако случайные величины X и Y, очевидно, не являются независимыми. Покажем это формально, т.е. покажем, что

$$P\{X < x, Y < y\} \neq P\{X < x\} \cdot P\{Y < y\}.$$

   Обозначим через $F(x)$ функцию распределения случайной величины X. Пусть дана определённость: $x > 0$, $y > 0$, $x < \sqrt{y}$. Тогда:

$$\begin{aligned} P\{X < x, Y < y\} &= P\{X < x, X^2 < y\} = P\{X < x, -\sqrt{y} < X < \sqrt{y}\} \\ &= P\{-\sqrt{y} < X < x\} = F(x) - F(-\sqrt{y}), \qquad (3) \\ P\{X < x\} \cdot P\{Y < y\} &= F(x) \cdot P\{X^2 < y\} = F(x) \cdot P\{-\sqrt{y} < X < \sqrt{y}\} \\ &= F(x) \cdot \left(F(\sqrt{y}) - F(-\sqrt{y})\right). \qquad (4) \end{aligned}$$

   Ясно, что (3) и (4) не совпадают, т.е. случайные величины X и Y не являются независимыми.

5. Отметим также, что жёсткая связь между случайными величинами X и Y (когда по реализации одной случайной величины можно однозначно определить реализацию другой случайной величины, что, например, имеет место, когда X и Y связаны функциональной зависимостью: $Y = \varphi(X)$, где $\varphi(\cdot)$ — заданная функция) далеко не всегда означает, что $|\rho_{XY}| = 1$; более того, в этом случае $|\rho_{XY}|$ может заметно отличаться от 1 (и даже быть нулём!).

Приведём пример. Пусть случайная величина X имеет плотность

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma_X} e^{-x^2/(2\sigma_X^2)}, \quad x \in (-\infty, +\infty),$$

а случайная величина $Y = X^{2k-1}$, $k = 2, 3, \ldots$. Найдём ρ_{XY}. Вычислим

$$\begin{aligned} \mathrm{Cov}(X,Y) &= M[XY] - M[X] \cdot M[Y] = M\left[X \cdot X^{2k-1}\right] - 0 = M\left[X^{2k}\right] \\ &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma_X} x^{2k} e^{-x^2/(2\sigma_X^2)} dx. \end{aligned}$$

Сделаем замену: $t = x/\sigma_X$, тогда

$$M\left[X^{2k}\right] = \frac{\sigma_X^{2k}}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} t^{2k} e^{-t^2/2} dt.$$

Введём обозначение:

$$J_{2k} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} t^{2k} e^{-t^2/2} dt.$$

Тогда

$$J_{2k} = (2k-1)J_{2k-2} = (2k-1)(2k-3)J_{2k-4} = \cdots = (2k-1)!! \cdot J_0,$$

где $J_0 = 1$. Таким образом,

$$M\left[X^{2k}\right] = \sigma_X^{2k} \cdot (2k-1)!!. \tag{5}$$

Далее,

$$\sigma_Y^2 = D\left[Y\right] = M\left[Y^2\right] - (M\left[Y\right])^2 = M\left[X^{4k-2}\right] - (M\left[X^{2k-1}\right])^2.$$

Поскольку $M\left[X^{2k-1}\right] = 0$ (нечётный момент центрированной нормальной величины), то

$$\sigma_Y^2 = M\left[X^{4k-2}\right] = \sigma_X^{4k-2} \cdot (4k-3)!!.$$

Таким образом,

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{M\left[X^{2k}\right]}{\sigma_X \sqrt{M\left[X^{4k-2}\right]}} = \frac{\sigma_X^{2k} \cdot (2k-1)!!}{\sigma_X \cdot \sigma_X^{2k-1} \sqrt{(4k-3)!!}} = \frac{(2k-1)!!}{\sqrt{(4k-3)!!}}.$$

При $k = 2$ получим:

$$\rho_{XY} = \frac{3!!}{\sqrt{5!!}} = \frac{3}{\sqrt{15}} \approx 0.774.$$

При $k = 3$ получим:

$$\rho_{XY} = \frac{5!!}{\sqrt{9!!}} = \frac{3 \cdot 5}{\sqrt{9 \cdot 7 \cdot 5 \cdot 3}} = \sqrt{\frac{5}{105}} \approx 0.487.$$