Лекция 9
Ковариация и коэффициент корреляции. Их свойства.

<u>Опред.</u> <u>Ковариацией</u> случайных величин X и Y называется число

$$Cov(X,Y) = M\{(X-MX)(Y-MY)\},$$

если математические ожидания справа существуют.

Правую часть последнего равенства можно преобразовать к виду $M\{XY\}-MX\cdot MY$, так что

$$Cov(X,Y) = M\{XY\} - MX\cdot MY.$$

<u>Опред.</u> Если $Cov(X,Y)=0$, то СВ X и Y называются <u>некоррелированными</u>.

Заметим, что если X и Y – независимы, то они некоррелированы.

Непосредственно из определения следуют <u>свойства ковариации</u>:

1) $Cov(\alpha X, Y) = Cov(X, \alpha Y) = \alpha\cdot Cov(X,Y)$, где α – постоянная;

2) $Cov(X+Y, Z) = Cov(X,Z) + Cov(Y,Z)$;

3) $Cov(X, Y+Z) = Cov(X,Y) + Cov(X,Z)$;

4) $Cov(X,X) = D\{X\}$, т.е. дисперсию формально можно рассматривать как ковариацию между случайной величиной и ей самой.

Найдём дисперсию суммы двух произвольных СВ X и Y:

$$D\{X+Y\} = M\{(X+Y-M\{X+Y\})^2\} =$$
$$= M\{((X-MX)+(Y-MY))^2\} =$$
$$= M\{(X-MX)^2\} + M\{(Y-MY)^2\} + 2M\{(X-MX)(Y-MY)\} =$$
$$= DX + DY + 2\,Cov(X,Y) .$$

В случае суммы n случайных величин аналогично устанавливается, что

$$D\{X_1+X_2+\ldots+X_n\} = \sum_{i=1}^{n} DX_i + 2\sum_{i<j} Cov(X_i,X_j) .$$

В частности, если все дисперсии DX_i и ~~коэффициенты~~ ковариации $Cov(X_i,X_j)$ одинаковы:

$$DX_i = d \ \ \forall i=1,\ldots,n , \quad Cov(X_i,X_j) = K \ \ \forall i\neq j ,$$

то

$$D\{X_1+X_2+\ldots+X_n\} = n\cdot d + n(n-1)K .$$

<u>Опред.</u> Коэффициентом корреляции случайных величин X и Y называется число

$$r_{xy} = \frac{Cov(X,Y)}{\sigma_x \sigma_y} ,$$

где $\sigma_x^2 = DX$, $\sigma_y^2 = DY$. (r_{xy} определён только при $\sigma_x > 0$ и $\sigma_y > 0$.)

<u>Утверждение</u>. $|r_{xy}| \le 1$.

~~Док-во~~:

$$D\{X+Y\} = DX + DY + 2 r_{xy}\sqrt{DX}\sqrt{DY} ,$$

откуда

$$\frac{D\{X+Y\}}{\sqrt{DX}\sqrt{DY}} = \frac{\sqrt{DX}}{\sqrt{DY}} + \frac{\sqrt{DY}}{\sqrt{DX}} + 2r_{xy} \quad (*)$$

Пусть $t = \frac{\sqrt{DX}}{\sqrt{DY}} \in (0;+\infty)$. Из $(*)$ следует, что каково бы ни было $t \in (0;+\infty)$

$$t + \frac{1}{t} + 2r_{xy} \geq 0$$

или

$$r_{xy} \geq -\frac{1}{2}\left(t+\frac{1}{t}\right) \quad \forall t \in (0;+\infty),$$

так что

$$r_{xy} \geq \max_{t\in(0;+\infty)} \left\{-\frac{1}{2}\left(t+\frac{1}{t}\right)\right\} = -1.$$

Далее

$$D\{X-Y\} = DX + DY - 2r_{xy}\sqrt{DX}\sqrt{DY}$$

$$\frac{D\{X-Y\}}{\sqrt{DX}\sqrt{DY}} = \frac{\sqrt{DX}}{\sqrt{DY}} + \frac{\sqrt{DY}}{\sqrt{DX}} - 2r_{xy} \quad . (**)$$

Пусть $t = \frac{\sqrt{DX}}{\sqrt{DY}} \in (0;+\infty)$. Из $(**)$ следует

$$t + \frac{1}{t} - 2r_{xy} \geq 0$$

или

$$r_{xy} \leq \frac{1}{2}\left(t+\frac{1}{t}\right) \quad \forall t \in (0;+\infty),$$

так что

$$r_{xy} \leq \min_{t\in(0;+\infty)} \left\{\frac{1}{2}\left(t+\frac{1}{t}\right)\right\} = 1.$$

Отметим также следующие св-ва коэффициента корреляции:

1) r_{xy} — безразмерная величина;

2) если СВ X и Y связаны линейной зависимостью

$$Y = \alpha X + \beta,$$

где $\alpha \neq 0$, β – произвольные постоянные, то

$$r_{xy} = 1 \text{ при } \alpha > 0;$$
$$r_{xy} = -1 \text{ при } \alpha < 0.$$

3) если X и Y независимые СВ, то $r_{xy} = 0$.

На содержательном уровне величина коэффициента r_{xy} характеризует тесноту связи между СВ X и Y: чем меньше $|r_{xy}|$, тем эта связь слабее. При этом знак r_{xy} допускает следующую интерпретацию: отрицательные значения r_{xy} означают, как правило, что в среднем 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 r_{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. $MX=0$) нормально распределенная случайная величина:

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

Пусть $Y=X^2$. Тогда

$$Cov(X,Y)=M\{X\cdot Y\}-MX\cdot MY=$$
$$=M\{X^3\}-0\cdot MY=\underbrace{\int_{-\infty}^{+\infty}x^3\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{x^2}{2\sigma^2}}dx}_{\text{интеграл от нечётной функции по симметричному промежутку}}=0,$$

так что и $r_{xy}=0$.

Однако СВ 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 СВ X. Пусть для определенности $x>0$, $y>0$, $x<\sqrt{y}$. Тогда

$$P\{X<x,Y<y\}=P\{X<x,X^2<y\}=P\{X<x,-\sqrt{y}<X<\sqrt{y}\}=$$
$$=P\{-\sqrt{y}<X<x\}=F(x)-F(-\sqrt{y}), \qquad (1)$$

$$P\{X<x\}\cdot P\{Y<y\}=F(x)\cdot P\{X^2<y\}=F(x)\cdot P\{-\sqrt{y}<X<\sqrt{y}\}=$$
$$=F(x)\cdot\left(F(\sqrt{y})-F(-\sqrt{y})\right). \qquad (2)$$

Ясно, что (1) и (2) не совпадают, т.е. СВ X и Y не являются независимыми.

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

Приведём пример. Пусть СВ X имеет плотность

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

а СВ $Y=X^{2k-1}$, $k=2,3,\ldots$ Найдём r_{xy}.

Имеем

$$\mathrm{Cov}(X,Y)=M\{XY\}-MX\cdot MY=M\{X\cdot X^{2k-1}\}-0=$$
$$=M\{X^{2k}\}=$$

$$=\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}\,\sigma_x}x^{2k}e^{-\frac{x^2}{2\sigma_x^2}}dx=\left\{\begin{matrix}\text{замена:}\\ t=\frac{x}{\sigma_x}\end{matrix}\right\}=\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}\sigma_x^{2k}t^{2k}e^{-\frac{t^2}{2}}dt=$$

$$=\frac{\sigma_x^{2k}}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}t^{2k-1}d\left(-e^{-\frac{t^2}{2}}\right)=\frac{\sigma_x^{2k}}{\sqrt{2\pi}}\left\{-t^{2k-1}e^{-\frac{t^2}{2}}\Big|_{-\infty}^{+\infty}+\int_{-\infty}^{+\infty}(2k-1)t^{2k-2}e^{-\frac{t^2}{2}}dt\right\}$$

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

Если ввести обозначение

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

что

$$J_{2k} = (2k-1)\int\limits_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} t^{2k-2} e^{-\frac{t^2}{2}} dt = (2k-1) J_{2k-2},$$

т.е.

$$J_{2k} = (2k-1) J_{2k-2} = (2k-1)(2k-3) J_{2k-4} =$$

$$= \ldots = (2k-1)(2k-3)\ldots(2k-(2k-1)) J_0 =$$

$$= (2k-1)!! \int\limits_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = (2k-1)!!,$$

так что $M\{X^{2k}\} = \sigma_x^{2k} \cdot (2k-1)!!$ (*)

Далее,

$$\sigma_y^2 = D\{Y\} = M\{Y^2\} - (MY)^2 = M\{X^{4k-2}\} - \left(M\{X^{2k-1}\}\right)^2 =$$

$$= M\{X^{4k-2}\} = \sigma_x^{4k-2} \cdot (4k-3)!!$$

по аналогии с результатом (*)

Таким образом,

$$r_{xy} = \frac{\cancel{M}\, Cov(X,Y)}{\sigma_x \sigma_y} = \frac{M\{X^{2k}\}}{\sigma_x \sqrt{M\{X^{4k-2}\}}} =$$

$$= \frac{\sigma_x^{2k} \cdot (2k-1)!!}{\sigma_x \cdot \sigma_x^{2k-1} \sqrt{(4k-3)!!}} = \frac{(2k-1)!!}{\sqrt{(4k-3)!!}}.$$

При $k=2$ получим

$$r_{xy} = \frac{3!!}{\sqrt{5!!}} = \frac{3}{\sqrt{15}} \approx 0{,}774.$$

При $k=3$ получим

$$r_{xy} = \frac{5!!}{\sqrt{9!!}} = \frac{3 \cdot 5}{\sqrt{9 \cdot 7 \cdot 5 \cdot 3}} = \frac{5}{\sqrt{105}} \approx 0{,}487.$$