Statystyczna analiza danych SAD-2020/2021

Wykład 6

Dwuwymiarowe zmienne losowe

$$X: S \to (-\infty, \infty), \qquad Y: S \to (-\infty, \infty)$$

- (X, Y) – para zmiennych losowych (dwuwymiarowa zmienna losowa, wektor losowy 2-wymiarowy)
- Zmienną losową X nazywamy brzegową zmienną losową, podobnie Y nazywamy brzegową zmienną losową

Przykłady:

➢ Wzrost, Waga losowo wybranej osoby z pewnej populacji

Dwuwymiarowe zmienne losowe

- Miesięczny dochód pracownika, Miesięczny dochód firmy
- Temperatura, Zużycie energii na klimatyzację

Problemy:

- Wyznaczenie rozkładu prawdopodobieństwa łącznego zmiennej l. (X,Y)
- Wyznaczenie rozkładów prawdopodobieństwa brzegowych zmiennych losowych na podstawie rozkładu łącznego

Dwuwymiarowe zmienne losowe

- Charakterystyki liczbowe określające stopień współzależności między zmiennymi $X,\ Y$
- Określenie niezależności zmiennych losowych
- Warunkowe rozkłady prawdopodobieństwa jednej ze zmiennych pod warunkiem, że druga przyjęła konkretną wartość

Przykład. Wyznaczyć funkcję prawdopodobieństwa łącznego pary (X, Y), gdzie X przyjmuje wartość 1 (0) jeśli wypadł orzeł (reszka) w pierwszym rzucie monetą, a Y jest liczbą orłów w pierwszym i drugim rzucie monetą.

Dwuwymiarowy wektor losowy

Niech 1 – wyrzucenie orła, 0 – wyrzucenie reszki

s	$X(s)$	$Y(s)$
$(0,0)$	0	0
$(0,1)$	0	1
$(1,0)$	1	1
$(1,1)$	1	2

Dwuwymiarowy wektor losowy

Funkcja prawdopodobieństwa łącznego:

$$f(x,y) := P(X = x, Y = y)$$

określona jest tablicą kontyngencyjną

y / x	0	1	2
0	1/4	1/4	0
1	0	1/4	1/4

Dwuwymiarowe zmienne losowe

Rozkład łączny pary zmiennych losowych (X, Y) określonych na tej samej przestrzeni zdarzeń elementarnych:

$P((X,Y) \in A)$, A - dowolny podzbiór zbioru par wartości zmiennych X, Y.

Definicja. Dystrybuantą zmiennej losowej (X, Y) nazywamy funkcję

$$F(x, y) = P(X \le x, Y \le y),$$

gdzie $-\infty < x < \infty, -\infty < y < \infty.$

Dwuwymiarowe zmienne losowe

Twierdzenie. Łączny rozkład prawdopodobieństwa zmiennej losowej (X,Y) określony jest jednoznacznie przez jej dystrybuantę.

Dyskretne zmienne losowe

Funkcja prawdopodobieństwa (łącznego) dwuwymiarowej zmiennej losowej dyskretnej:

$$f(x,y) = P(X = x, Y = y).$$

Dwuwymiarowe zmienne losowe

Własności:

- $f(x,y) \geq 0,$ dla dowolnej pary wartości (x,y),
- $\sum_{x}\sum_{y} f(x,y) = 1,$
- $P((X,Y)\in A) = \sum_{(x,y)\in A} f(x,y),$
- $F(x,y) = \sum_{s\leq x}\sum_{t\leq y} f(s,t).$

Dwuwymiarowe zmienne losowe

Notacja.

Niech $x_1, x_2, \dots, x_r$ będą wartościami zmiennej losowej X, a $y_1, y_2, \dots, y_s$ będą wartościami zmiennej losowej Y

$$p_{ij} := P\left(X = x_i, Y = y_j\right), \quad i = 1, \dots, r, j = 1, \dots, s$$

Tablica kontyngencyjna

j i	$1 \quad 2 \quad \ldots\ldots \qquad\qquad \ldots. \quad s$
1	$p_{11}\ p_{12}\ \ \ldots\ \ldots\ \ \ldots. \qquad\qquad p_{1s}$
2	$p_{21}\ p_{22}\ \ldots\ \ldots..\ \ \ldots. \qquad\qquad p_{2s}$
.	
.	
.	
r	$p_{r1}\ p_{r2}\ \ \ldots\ \ldots\ \ldots\ \ldots. \qquad\qquad p_{rs}$

Całka podwójna

$D := \{(x,y)\colon\ \ a \le x \le b, c \le y \le d\}$ – prostokąt

$$D = [a,b] \times [c,d]$$

$$\iint_D f(x,y)dxdy = \int_a^b \left(\int_c^d f(x,y)dy \right) dx$$

Obliczanie całki podwójnej

Przykład. Niech $D = [0,1] \times [2,3]$

$$\iint_D xydxdy = \int_0^1 \left(\int_2^3 xydy \right) dx =$$

$$= \int_0^1 x \left(\int_2^3 ydy \right) dx = \int_0^1 x \left(\frac{y^2}{2} \Big|_2^3 \right) dx =$$

$$= \left(\frac{3^2}{2} - \frac{2^2}{2} \right) \int_0^1 xdx = \frac{5}{2} \frac{x^2}{2} \Big|_0^1 = \frac{5}{4}$$

Obliczanie całki podwójnej

Niech $D = (-\infty, x] \times (-\infty, y]$

$$\iint_D f(x,y)dxdy = \int_{-\infty}^{x}\left(\int_{-\infty}^{y} f(x,y)dy\right)dx =$$

$$= \lim_{a\to\infty, c\to\infty} \int_{a}^{x}\left(\int_{c}^{y} f(x,y)dy\right)dx$$

Wektor losowy typu ciągłego

Dwuwymiarowa zmienna losowa (X, Y) jest ciągłą zmienną losową, jeśli jej **łączny rozkład prawdopodobieństwa** określony jest przez **funkcję gęstości (łączną gęstość prawdopodobieństwa):**

$$P((X,Y)\in A)=\iint\limits_{A} f(x,y)\,dxdy$$

Dla $A=(-\infty,x]\times(-\infty,y]$:

$$\boxed{F(x,y) = P(X\le x, Y\le y)=\int_{-\infty}^{x}\int_{-\infty}^{y} f(s,t)\,dtds.}$$

Łączna gęstość prawdopodobieństwa

Własności gęstości

- $f(x,y) \geq 0$

- $\int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1$

- $f(x,y) = \dfrac{\partial^2}{\partial x \partial y} F(x,y), \quad -\infty < x < \infty, \ -\infty < y < \infty.$

Łączna gęstość prawdopodobieństwa

Przykład

$$f(x,y) = \begin{cases} Cx & \text{dla} \quad (x,y) \in [0,1] \times [0,2] \\ 0 & \text{dla} \quad (x,y) \notin [0,1] \times [0,2] \end{cases}$$

Wyznaczyć: a) stałą C, $F\left(\frac{1}{2}, \frac{1}{2}\right)$, $F\left(\frac{3}{2}, 1\right)$

b) $P\left(X > \frac{1}{2}, Y \leq \frac{3}{4}\right)$

$$1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)\, dxdy = \iint_{[0,1]\times[0,2]} Cxdxdy$$

Łączna gęstość prawdopodobieństwa

$$\iint_{[0,1]\times[0,2]} Cx\,dx\,dy = \int_0^1 \left(\int_0^2 Cx\,dy \right) dx = \int_0^1 Cx \left(\int_0^2 dy \right) dx$$

$$= C \cdot 2 \cdot \frac{x^2}{2}\Big|_0^1 = 2C\left(\frac{1^2}{2} - \frac{0^2}{2}\right) = C \quad \Longrightarrow \quad C = 1$$

$$F\left(\frac{1}{2}, \frac{3}{2}\right) = P\left(X \le \frac{1}{2}, Y \le \frac{3}{2}\right) = \int_{-\infty}^{\frac{1}{2}} \left(\int_{-\infty}^{\frac{3}{2}} f(x,y)\,dy \right) dx =$$

Łączna gęstość prawdopodobieństwa

$$F\left(\frac{1}{2},\frac{3}{2}\right)=\int_0^{\frac{1}{2}}\left(\int_0^{\frac{3}{2}} x\,dy\right)dx=\int_0^{\frac{1}{2}} x\cdot\frac{3}{2}\,dx=\frac{3}{2}\frac{x^2}{2}\bigg|_0^1=\frac{3}{4}$$

$$F\left(\frac{3}{2},1\right)=\int_0^1\left(\int_0^1 x\,dy\right)dx=\int_0^1 x\left(\int_0^1 dy\right)dx=\frac{x^2}{2}\bigg|_0^1=\frac{1}{2}$$

Łączna gęstość prawdopodobieństwa

$$P\left(X > \frac{1}{2}, Y \le \frac{3}{4}\right) = \int_{\frac{1}{2}}^{1} \left(\int_{0}^{\frac{3}{4}} x dy \right) dx = \int_{\frac{1}{2}}^{1} x \cdot \frac{3}{4} dx =$$

$$= \frac{3}{4} \frac{x^2}{2} \bigg|_{\frac{1}{2}}^{1} = \frac{3}{4} \left(\frac{1^2}{2} - \frac{\left(\frac{1}{2}\right)^2}{2} \right) = \frac{9}{32}$$

Brzegowe rozkłady prawdopodobieństwa

Niech (X,Y) będzie dwuwymiarową zmienną losową o rozkładzie prawdopodobieństwa określonym przez funkcję $f(x,y)$ (funkcja prawdopodobieństwa lub gęstość).

Rozkład brzegowy = rozkład prawdopodobieństwa

zmiennej losowej *X* lub zmiennej losowej *Y.*

Brzegowe rozkłady prawdopodobieństwa

- dla **dyskretnej** zmiennej (X, Y) , **brzegowe funkcje prawdopodobieństwa** są postaci

$$f_X(x) = P(X = x) = \sum_y f(x, y)$$

$$f_Y(y) = P(Y = y) = \sum_x f(x, y)$$

- dla **ciągłej** zmiennej (X, Y) , **brzegowe gęstości** są postaci

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy, \qquad f_Y(y) = \int_{-\infty}^{\infty} f(x, y)dx.$$

Brzegowa funkcja prawdopodobieństwa

Przykład. (c.d. str. 4-6)

$x \backslash y$	0	1	2	$f_X(x) = P(X=x)$
0	$1/4$	$1/4$	0	$1/2$
1	0	$1/4$	$1/4$	$1/2$
$f_Y(y) = P(Y=y)$	$1/4$	$2/4$	$1/4$	1

np

$$f_X(0) = f(0,0) + f(0,1) + f(0,2) = 1/2$$

$$f_Y(1) = f(0,1) + f(1,1) = 2/4$$

Rozkłady brzegowe w tablicy kontyngencyjnej

$$p_{ij} := P\left(X = x_i, Y = y_j\right), \quad i = 1, \dots, r, j = 1, \dots, s$$

$$p_{i\cdot} := P(X = x_i)$$

$$p_{\cdot j} := P\left(Y = y_j\right)$$

$$p_{i\cdot} = p_{i1} + p_{i2} + \cdots + p_{is}, \quad i = 1,2, \dots, r$$

$$p_{\cdot j} = p_{1j} + p_{2j} + \cdots + p_{rj}, j = 1,2, \dots, s$$

Rozkłady brzegowe w tablicy kontyngencyjnej

$$p_{ij} := P(X = x_i, Y = y_j), \quad i = 1, \dots, r, j = 1, \dots, s$$

j / i	1 2 …… …. s	$p_{i\cdot}$
1	p_{11} p_{12} … … …. p_{1s}	$p_{1\cdot}$
2	p_{21} p_{22} … … .. …. p_{2s}	$p_{2\cdot}$
.		
.		
r	p_{r1} p_{r2} … … … …. p_{rs}	$p_{r\cdot}$
$p_{\cdot j}$	$p_{\cdot 1}$ $p_{\cdot 2}$ … … …. $p_{\cdot s}$	1

Brzegowe gęstości prawdopodobieństwa

Przykład. $f(x,y) = \begin{cases} x & \text{dla} \quad (x,y) \in [0,1] \times [0,2] \\ 0 & \text{dla} \quad (x,y) \notin [0,1] \times [0,2] \end{cases}$

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy$$

Należy rozważyć 2 przypadki

1) $x \notin [0,1] \implies f_X(x) = \int_{-\infty}^{\infty} 0dy = 0$

2) $x \in [0,1] \implies f_X(x) = \int_{0}^{2} xdy = 2x$

Brzegowe gęstości prawdopodobieństwa

Przykład. $$f(x,y) = \begin{cases} x & \text{dla } (x,y) \in [0,1] \times [0,2] \\ 0 & \text{dla } (x,y) \notin [0,1] \times [0,2] \end{cases}$$

$$f_X(x) = \begin{cases} 2x & dla \;\; x \in [0,1] \\ 0 & dla \;\; x \notin [0,1] \end{cases}$$

Podobnie znajdziemy gęstość zmiennej brzegowej Y.

$$f_Y(y) = \begin{cases} 1/2 & dla \;\; y \in [0,2] \\ 0 & dla \;\; y \notin [0,2] \end{cases}$$

$$Y \sim U(0,2)$$

Warunkowe rozkłady prawdopodobieństwa

- Niech $f(x,y)$ funkcja prawdopodobieństwa **dyskretnej** zmiennej losowej (X,Y), y – ustalone, $f_Y(y) > 0$.

Rozkład **warunkowy** zmiennej losowej X pod warunkiem, że $Y = y$ określa **warunkowa funkcja prawdopodobieństwa**:

$$f(x|y) = \frac{f(x,y)}{f_Y(y)}, \quad x \text{ – dowolna wartość zmiennej } X.$$

$$f(x|y) = \frac{P(X=x, Y=y)}{P(Y=y)} = P(X=x|Y=y)$$

Warunkowe rozkłady prawdopodobieństwa

Analogicznie:

$$f(y|x) = \frac{f(x,y)}{f_X(x)} = P(Y=y|X=x), \text{ gdzie } f_X(x) > 0,$$

y - dowolna wartość Y

Notacja:

$$f(x|y) = f_{X|Y}(x|y)$$

$$f(y|x) = f_{Y|X}(y|x)$$

Warunkowa funkcja prawdopodobieństwa

Przykład. (kontynuacja, SADW06) Zmienne losowe X, Y oznaczają liczby punktów uzyskane w etapie I i II, odpowiednio, przez losowo wybranego uczestnika teleturnieju.

- Znaleźć **rozkład brzegowy zmiennej** ***Y***, liczby punktów uzyskanych w II etapie teleturnieju, przez losowo wybranego uczestnika.

- Wyznaczyć **rozkład warunkowy** ***Y*** **pod warunkiem**, że w I etapie uzyskano 2 punkty, tzn. $X = 2$.

Warunkowa funkcja prawdopodobieństwa

Y X	0	1	2
0	0,50	0,05	0,01
1	0,20	0,10	0,06
2	0,02	0,03	0,03

- $f_Y(y) = f(0,y) + f(1,y) + f(2,y)$. Stąd

y	0	1	2
$f_Y(y)$	0,72	0,18	0,1

- $f(y|2) = f_{Y|X}(y|2) = \dfrac{f(2,y)}{f_X(2)} = \dfrac{f(2,y)}{0{,}08} = ?$

y	0	1	2
$f(y\|2)$	$\frac{0{,}02}{0{,}08} = 2/8$	$\frac{0{,}03}{0{,}08} = 3/8$	$\frac{0{,}03}{0{,}08} = 3/8$

Warunkowa gęstość prawdopodobieństwa

Niech $f(x,y)$ - łączna gęstość **ciągłej** zmiennej losowej (X,Y), y – ustalone: $f_Y(y) > 0$

Warunkową gęstością prawdopodobieństwa zmiennej losowej X pod warunkiem, że $Y = \mathrm{y}$ nazywamy funkcję

$$f(x|y) = \frac{f(x,y)}{f_Y(y)}, \qquad x \in (-\infty, \infty)$$

Notacja: $f(x|y) = f_{X|Y}(x|y)$

Warunkowa gęstość prawdopodobieństwa

Niech $f(x,y)$ - łączna gęstość **ciągłej** zmiennej losowej (X,Y), x – ustalone: $f_X(x) > 0$

Warunkową gęstością prawdopodobieństwa zmiennej losowej Y pod warunkiem, że $X = x$ nazywamy funkcję

$$f(y|x) = \frac{f(x,y)}{f_X(x)}, \qquad y \in (-\infty, \infty)$$

Notacja: $f(y|x) = f_{Y|X}(y|x)$

Niezależne zmienne losowe

Definicja. Niech (X,Y) będzie dwuwymiarową zmienna losową o dystrybuancie $F(x,y)$ oraz dystrybuantach brzegowych $F_X(x)$, $F_Y(y)$, $x, y \in (-\infty, \infty)$.

Zmienne losowe ***X*, *Y* są niezależne**, jeśli

$$F(x,y) = F_X(x) F_Y(y),$$

dla wszystkich wartości *x*, *y*.

Niezależne zmienne losowe

Zmienne losowe X, Y są niezależne $\Leftrightarrow$

dla dowolnych $x \in (-\infty, \infty)$, $y \in (-\infty, \infty)$ zdarzenia

$\{X \leq x\}$, $\{Y \leq y\}$ są niezależne

Zmienne losowe są zależne, jeśli nie są niezależne

Niezależne zmienne losowe

Twierdzenie.

Zmienne losowe ***X*, *Y*** **są niezależne** wtedy i tylko wtedy gdy dla wszystkich wartości *x*, *y*.

$$f(x,y) = f_X(x) f_Y(y)$$

Wniosek. Poniższe warunki są równoważne:

- Zmienne losowe ***X*, *Y*** **są niezależne**.
- Rozkłady warunkowe są takie jak rozkłady brzegowe

Niezależne zmienne losowe

Wniosek. Poniższe warunki są równoważne:

- Zmienne losowe ***X*, *Y* są niezależne**.
- $f_{X|Y}(x|y) = f_X(x)$, x - dowolne, y -takie że $f_Y(y) > 0$.
- $f_{Y|X}(y|x) = f_Y(y)$, y - dowolne, x- takie że $f_X(x) > 0$.

Przykład. (kontynuacja) Czy liczby punktów uzyskane w I i II etapie teleturnieju przez losowo wybranego uczestnika są niezależnymi zmiennymi losowymi ?

Y / X	0	1	2
0	0,50	0,05	0,01
1	0,20	0,10	0,06
2	0,02	0,03	0,03

$f_X(0) = f(0{,}0) + f(0{,}1) + f(0{,}2)$ = 0,5 + 0,05 + 0,01 = 0,56.

$f_Y(0) = f(0{,}0) + f(1{,}0) + f(2{,}0)$ = 0,5 + 0,2 + 0,02 = 0,72.

$$f(0{,}0) = 0{,}5 \neq 0{,}56 \times 0{,}72 = f_X(0) f_Y(0)$$

Zmienne losowe X, Y są **zależne.**

Przykład. Czy *X*, *Y* są **niezależnymi** zmiennymi losowymi, jeśli ich łączna gęstość ma postać:

$$f(x,y)=\begin{cases}3(x-y)^2/8 & \text{gdy } -1\le x\le 1, -1\le y\le 1\\ 0 & \textit{przeciwnie}\end{cases}$$

Dla $x, y \in [-1,1]$:

$f_X(x)=(3x^2+1)/4$ oraz $f_Y(y)=(3y^2+1)/4$.

$$f(x,y)\neq f_X(x)f_Y(y).$$

X, *Y* są zależne.

Przykład. Czasy poprawnej pracy dwu podzespołów są **niezależnymi** zmiennymi losowymi ***X, Y*** o rozkładach **wykładniczych** z parametrami λ_1, λ_2, odpowiednio. Średnie czasy pracy podzespołów wynoszą 1000 (godzin) i 1200 (godzin). Obliczyć prawdopodobieństwo zdarzenia takiego, że żaden podzespół nie ulegnie awarii przed upływem 1500 godzin.

$E(X) = 1/\lambda_1 = 1000$ (godz.), $E(Y) = 1/\lambda_2 = 1200$ (godz.)

Stąd $\lambda_1 = 1/1000$ (1/godz.) $\lambda_2 = 1/1200$ (1/godz.).

$$P(X \geq 1500, Y \geq 1500) = P(X \geq 1500)\, P(Y \geq 1500) =$$

$$e^{-\lambda_1 1500} \times e^{-\lambda_2 1500} = e^{-1500/1000} \times e^{-1500/1200} =$$

$$= 0{,}2231 \times 0{,}2865 = 0{,}0639.$$

Wartość oczekiwana. Kowariancja.

- $$E[g(X,Y)] = \sum_x \sum_y g(x,y) f(x,y),$$

  gdy (*X*, *Y*) - **dyskretna,**

- $$E[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f(x,y) dxdy,$$

  gdy (*X*, *Y*) - **ciągła.**

Uwaga. Dla $g(X,Y) = X$ lub $g(X,Y) = Y$ otrzymujemy wartości oczekiwane brzegowych zmiennych losowych *X* lub *Y*, gdyż

Przykład.

$x \backslash y$	0	1	2
0	1/4	1/4	0
1	0	1/4	1/4

$$E(XY) = \sum_x \sum_y xyf(x,y) = 0 \cdot 0 \cdot \frac{1}{4} + 0 \cdot 1 \cdot \frac{1}{4} + 0 \cdot 2 \cdot 0 +$$

$$+1 \cdot 0 \cdot 0 + 1 \cdot 1 \cdot \frac{1}{4} + 1 \cdot 2 \cdot \frac{1}{4} = \frac{3}{4}$$

- w przypadku **dyskretnym:**

$$E(X) = \sum_x \sum_y x f(x,y) = \sum_x x \sum_y f(x,y) = \sum_x x f_X(x) = \mu_X \,.$$

$$E(Y) = \sum_x \sum_y y f(x,y) = \sum_y y \sum_x f(x,y) = \sum_y y f_Y(y) = \mu_Y$$

- w przypadku **ciągłym:**

$$E(X) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f(x,y)\,dx\,dy =$$

$$\int_{-\infty}^{\infty} x \left(\int_{-\infty}^{\infty} f(x,y)\,dy \right) dx = \int_{-\infty}^{\infty} x f_X(x)\,dx = \mu_X \,.$$

Analogicznie:

$$E(Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f(x, y) dx dy = \int_{-\infty}^{\infty} y f_Y(y) dy = \mu_Y .$$

Stwierdzenie. Niech *c* będzie dowolną stałą, a $g(X,Y)$, $g_1(X,Y)$, $g_2(X,Y)$ zmiennymi losowymi jednowymiarowymi.

- $E[cg(X,Y) = cE[g(X,Y)]$
- $E[g_1(X,Y) + g_2(X,Y)] = E[g_1(X,Y)] + E[g_2(X,Y)]$

Stwierdzenie.

Jeśli zmienne losowe *X*, *Y* są niezależne, to

$$E(XY) = E(X)E(Y).$$

D. Niezależność zmiennych jest równoważna

$f(x, y) = f_X(x) f_Y(y)$. Stąd i z definicji wartości średniej:

- (zmienne dyskretne)

$$E[g(X,Y)] = \sum_x \sum_y g(x, y) f(x, y).$$

$$E(XY) = \sum_x \sum_y xyf(x,y) = \sum_x \sum_y xyf_X(x) f_Y(y) =$$

$$\sum_x xf_X(x) \sum_y yf_Y(y) = \sum_y yf_Y(y) \times \sum_x xf_X(x) =$$

$$E(Y)E(X) = E(X)E(Y).$$

- (zmienne ciągłe) Dowód analogiczny - sumowanie należy zastąpić całkowaniem.

Definicja. Niech X i Y będą zmiennymi losowymi o łącznej funkcji prawdopodobieństwa (gęstości) $f(x, y)$.

Kowariancją zmiennych X i Y nazywamy liczbę

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)].$$

Uwaga. Z definicji σ_{XY} oraz $E[g(X,Y)]$, przyjmując $g(x, y) = (x - \mu_X)(y - \mu_Y)$, otrzymujemy:

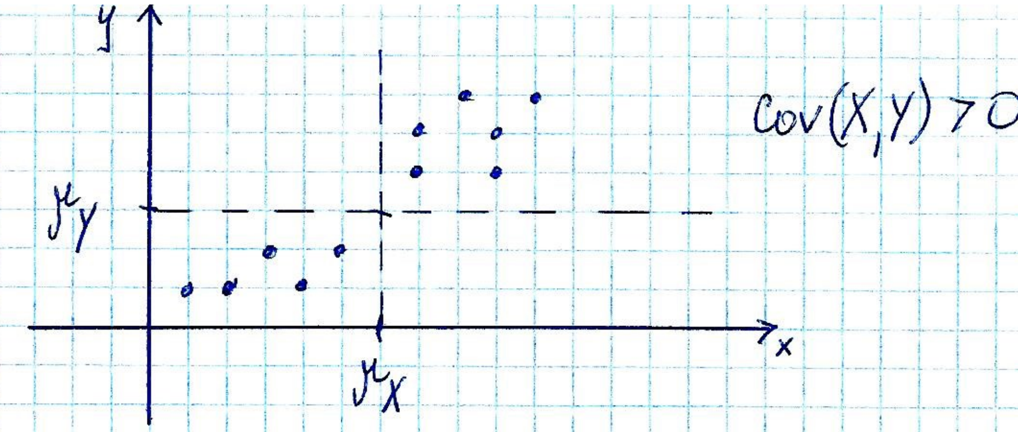
$$\sigma_{XY} = \sum_x \sum_y (x - \mu_X)(y - \mu_Y) f(x, y)$$, gdy (X, Y) - **dyskretna**

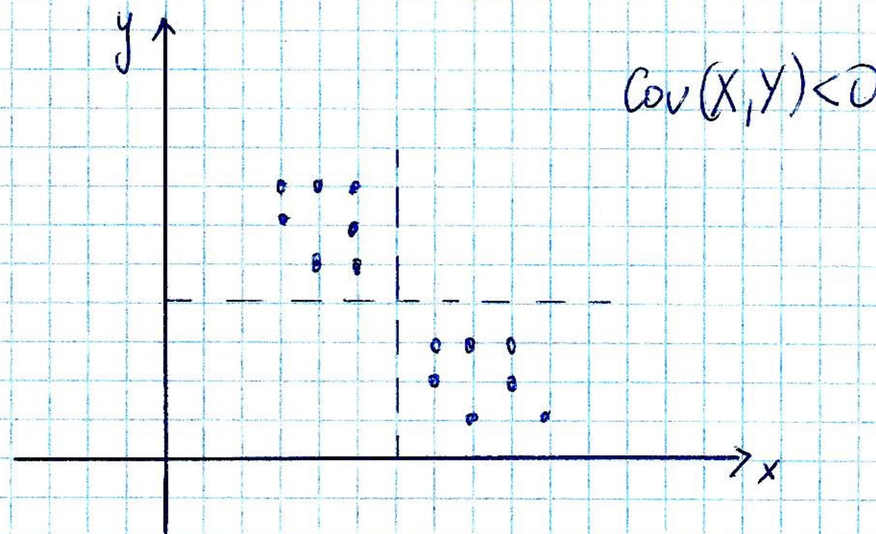
$$\sigma_{XY} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f(x, y) dx dy$$, gdy (X, Y) - **ciągła**

Interpretacja. Kowariancja określa pewną **współzależność** między zmiennymi losowymi:

- Jeśli „dużym” wartościom zmiennej *X* przewyższającym μ_X towarzyszą zwykle „duże” wartości zmiennej *Y* przewyższające μ_Y, a wartościom *X* mniejszym od μ_X towarzyszą zwykle wartości *Y* mniejsze od μ_Y, to $\sigma_{XY} > 0$
- Jeśli wartościom zmiennej *X* większym od μ_X towarzyszą zwykle wartości *Y* mniejsze od μ_Y wartościom *X* mniejszym od μ_X towarzyszą zwykle wartości *Y* większe od od μ_Y, to $\sigma_{XY} < 0$

Notacja: Zamiast $\boxed{\sigma_{XY}}$ często piszemy **Cov (*X*,*Y*).**


y
x
μ_Y
μ_X
Cov(X,Y) > 0


y
x
Cov(X,Y) < 0

Stwierdzenie.

$$\mathbf{Cov}(\boldsymbol{X},\boldsymbol{Y}) = E(XY) - \mu_X \mu_Y .$$

D. $\mathrm{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] =$

$$E(XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y) =$$

$$E(XY) - E(X\mu_Y) - E(Y\mu_X) + \mu_X\mu_Y =$$

$$E(XY) - \mu_X\mu_Y$$

Twierdzenie.

Jeśli zmienne losowe X i Y są niezależne, to

$$\mathrm{Cov}(X, Y) = 0$$

D. Dla niezależnych zmiennych losowych

$E(XY) = E(X)E(Y)$. Stąd oraz wzoru na kowariancję mamy:

$$\mathrm{Cov}(X,Y) = E(XY) - \mu_X \mu_Y =$$

$$= E(X)E(Y) - \mu_X \mu_Y = 0.$$

Uwaga. Twierdzenie odwrotne nie jest na ogół prawdziwe.

Przykład

Niech zmienna losowa X ma rozkład jednostajny na zbiorze $\{-1,0,1\}$, tzn

$$f_X(-1) = f_X(0) = f_X(1) = \frac{1}{3}$$

Niech $Y = X^2$. Między zmiennymi istnieje deterministyczna zależność, ale $Cov(X,Y) = 0$, ponieważ

$$Cov(X,Y) = E(XY) - E(X)E(Y) = E(X^3) - 0 = E(X) - 0 = 0$$

X, Y – niezależne $\Longrightarrow$ $Cov(X,Y) = 0$

$Cov(X,Y) = 0$ $\not\Rightarrow$ X, Y – niezależne, w przykładzie są zależne mimo że $Cov(X,Y) = 0$

Twierdzenie. Dla dowolnych stałych *a*, *b*

$$\boxed{\mathrm{Var}(aX+bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) + 2ab\mathrm{Cov}(X,Y)}$$

D. $E\{\,[(aX+bY)-(a\mu_X+b\mu_Y)]^2\,\} =$

$$E\{\,[a(X-\mu_X)+b(Y-\mu_Y)]^2\,\} = E\{\,[a(X-\mu_X))]^2\,\}$$

$$+\ E\,[2ab(X-\mu_X)(Y-\mu_Y)] \ +\ E\{\,[b(Y-\mu_Y)]^2\,\} =$$

$$a^2\mathrm{Var}(X) + 2ab\mathrm{Cov}(X,Y) + b^2\mathrm{Var}(Y)$$

Wniosek. Jeśli zmienne losowe X i Y są **niezależne**, to

$$\boxed{\mathrm{Var}(aX + bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y)}$$

Definicja. **Współczynnikiem korelacji** między zmiennymi losowymi X i Y nazywamy liczbę:

$$\boxed{\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}}$$

Przykład. $\rho = ?$

Y / X	0	1	2
0	0,50	0,05	0,01
1	0,20	0,10	0,06
2	0,02	0,03	0,03

- $E(X) = \sum_x \sum_y x f(x, y)$ = 0 x (0,5 + 0,05 + 0,01) + 1 x (0,2 + 0,1 + 0,06) + 2 x (0,02+0,03+0,03) = 0,52

- $E(Y) = \sum_x \sum_y y f(x, y)$ = 0 x (0,5 + 0,2 + 0,02) + + 1 x (0,05 + 0,1 + 0,03) + 2 x (0,01+0,06+0,03) = 0,38

- $E(XY) = \sum_x \sum_y xyf(x,y)$ = 0 + 0 + 0 + 0 + 1 x 1 x 0,1 +

  1 x 2 x 0,06 + 2 x 1 x 0,03 + 2 x 2 x 0,03 = 0,4

- Cov(X,Y) = 0,4 – 0,52 x 0,38 = 0,2024

- $E(X^2) =$ 0 + 1^2 x (0,2 + 0,1 + 0,06) +

  2^2 x (0,02 = 0,03 + 0,03) = 0,68

- $E(Y^2) =$ 1^2 x (0,05 + 0,1 + 0,03) +

  2^2 x (0,01 + 0,06 + 0,03) = 0,58

- Var(X) = $E(X^2) - [E(X)]^2$ = 0,68 - $0,52^2$ = 0,4096
- Var(Y) = $E(Y^2) - [E(Y)]^2$ = 0,58 - $0,38^2$ = 0,4356

$$\boxed{\rho = \frac{0,2024}{\sqrt{0,4096} \times \sqrt{0,4356}} = 0,47916}$$

Własności współczynnika korelacji

- $\boxed{-1 \leq \rho \leq 1}$
- Jeśli *a* i *b* są stałymi, ***Y* = *a* + *bX*,**

  to $\rho = \begin{cases} 1 & \text{gdy } b > 0 \\ -1 & \text{gdy } b < 0 \end{cases}$
- Jeśli $|\rho| = 1$, to między zmiennymi losowymi *X*, *Y*

  istnieje **liniowa zależność funkcyjna**
- Jeśli zmienne losowe *X* i *Y* są **niezależne,** to

$$\boxed{\rho = 0}$$

Interpretacja.

Współczynnik korelacji jest miarą zależności liniowej między zmiennymi losowymi

Dwuwymiarowy rozkład normalny

Zmienna losowa (X, Y) ma dwuwymiarowy rozkład normalny, jeśli ma gęstość postaci:

$$\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\, exp\left\{-\frac{1}{2(1-\rho^2)}\times q(x,y)\right\},$$

gdzie

$$q(x,y)=\frac{(x-\mu_X)^2}{\sigma_X^2}-2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}+\frac{(y-\mu_Y)^2}{\sigma_y^2},$$

$-\infty < x < \infty$, $-\infty < y < \infty$, stałe σ_X, σ_Y, ρ spełniają warunki

$\sigma_X > 0$, $\sigma_Y > 0$, $-1 < \rho < 1$

Notacja: $\boxed{(X,Y) \sim N(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)}$

Twierdzenie. Jeśli $\boxed{(X,Y) \sim N(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)}$, to

- ***X*** **~** $N(\mu_X, \sigma_X)$, ***Y*** **~** $N(\mu_Y, \sigma_Y)$
- Cov(*X*,*Y*) = $\rho \times \sigma_X \times \sigma_Y$
- ***X***, ***Y*** są **niezależne** wtedy i tylko wtedy gdy Cov(*X*,*Y*) = 0

Twierdzenie. Zmienna losowa (*X*,*Y*) ma dwuwymiarowy rozkład normalny wtedy i tylko wtedy gdy zmienna losowa ***aX* + *bY*** ma rozkład normalny, *a*, *b* są dowolnymi stałymi