Wektory losowe

Joanna Czarnowska[1]

[1]Uniwersytet Gdański
Instytut Informatyki

Rozkład wektora losowego dyskretnego

Niech X i Y będą zmiennymi losowymi. Parę (X, Y) nazywamy wektorem losowym dwuwymiarowym.

Mówimy, że wektor jest typu dyskretnego, jeśli zmienne losowe X i Y są typu dyskretnego.

Definicja

Niech X będzie zmienną losową przyjmującą wartości $\{x_1, x_2, x_3, \ldots\}$, a Y zmienną losową o wartościach $\{y_1, y_2, y_3, \ldots\}$. Przez p_{kl} oznaczmy prawdopodobieństwo

$$p_{kl} = P(X = x_k, Y = y_l), \; k, l = 1, 2, \ldots.$$

Zbiór $\{(x_k, y_l), p_{kl}\}_{k,l \geqslant 1}$ nazywamy rozkładem wektora (X, Y).

Rozkład wektora dwuwymiarowego dyskretnego można przedstawić w tabeli.

Przykład 1. Rozkład wektora (X, Y).

$_Y\backslash^X$	-1	0	1	$P(Y=y)$
-1	$0,2$	0	$0,2$	
0	0	$0,2$	0	
1	$0,2$	0	$0,2$	
$P(X=x)$				

Z tabeli odczytujemy na przykład, że prawopodobieństwo zdarzenia w którym jednocześnie X i Y przyjmuje wartość -1 wynosi: $P(X=-1, Y=-1) = 0,2$.

Rozkłady brzegowe

Mając dany rozkład wektora (X, Y), możemy wyznaczyć rozkłady zmiennych losowych X i Y, czyli rozkłady brzegowe

$$P(X=x_i) = P(X=x_i, Y=y_1) + P(X=x_i, Y=y_2) + \ldots = \sum_{j=1}^{\infty} P(X=x_i, Y=y_j),$$

$$P(Y=y_j) = P(X=x_1, Y=y_j) + P(X=x_2, Y=y_j) + \ldots = \sum_{i=1}^{\infty} P(X=x_i, Y=y_j).$$

Rozkłady brzegowe – przykład

Rozkład łączny wektora (X, Y) i rozkłady brzegowe (na brzegach tabeli)

$_Y\backslash^X$	-1	0	1	$P(Y = y)$
-1	$0,2$	0	$0,2$	$0,4$
0	0	$0,2$	0	$0,2$
1	$0,2$	0	$0,2$	$0,4$
$P(X = x)$	$0,4$	$0,2$	$0,4$	

Rozkłady brzegowe, a rozkład wektora

Powstaje pytanie, czy znając rozkłady brzegowe, czyli rozkłady zmiennych X i Y można wyznaczyć rozkład wektora (X, Y)?

Odpowiedź brzmi, że w ogólnym przypadku – nie. Możemy to zrobić, gdy zmienne X i Y są niezależne.

Niezależność zmiennych losowych

Definicja

Dyskretne zmienne losowe X i Y nazywamy niezależnymi, jeśli

$$P(X = x, Y = y) = P(X = x) \cdot P(Y = y) \tag{1}$$

dla wszystkich $x, y \in \mathbb{R}$. Zmienne losowe które nie są niezależne, nazywamy zależnymi.

Niezależność zmiennych losowych – przykład

Przykład 1 c.d. Rozkład wektora (X, Y).

$_Y\backslash^X$	-1	0	1	$P(Y=y)$
-1	$0,2$	0	$0,2$	$0,4$
0	0	$0,2$	0	$0,2$
1	$0,2$	0	$0,2$	$0,4$
P(X=x)	$0,4$	$0,2$	$0,4$	

Zmienne losowe X i Y są zależne

$$P(X=-1, Y=-1) = 0,2, \quad P(X=-1)\cdot P(Y=-1) = 0,4\cdot 0,4 = 0,8,$$

a zatem $P(X=-1, Y=-1) \neq P(X=-1)\cdot P(Y=-1)$.

Niezależność zmiennych losowych – przykład

Przykład 2. W tabeli podany jest rozkład wektora (X, Y). Zmienne losowe X i Y są niezależne.

${}_Y\backslash{}^X$	1	0	$P(Y = y)$
1	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{4}$
2	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$
P(X=x)	$\frac{1}{3}$	$\frac{2}{3}$	

Wartość oczekiwana iloczynu zmiennych losowych

Wartość oczekiwaną iloczynu dyskretnych zmiennych losowych $X \in \{x_1, x_2, \dots\}$ i $Y \in \{y_1, y_2, \dots\}$ możemy obliczyć korzystając z zależności

$$E(XY) = \sum_{k,l \geqslant 1} x_k y_l \cdot P(X = x_k, Y = y_l).$$

Twierdzenie

Jeżeli zmienne losowe X i Y są niezależne, to

$$E(XY) = EX \cdot EY.$$

Niezależność zmiennych losowych – przykład

Przykład 2. c.d. Zmienne losowe X i Y są niezależne.

$Y \backslash X$	1	0	$P(Y=y)$
1	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{4}$
2	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$
P(X=x)	$\frac{1}{3}$	$\frac{2}{3}$	

$EX = 1/3,$
$EY = 1/4 + 6/4 = 7/4$
$E(XY) = 1/12 + 2/4 = 7/12 = EXEY$

Kowariancja. Współczynnik korelacji Pearsona

Do opisu zależności między zmiennymi losowymi, wykorzystuje się kowariancję oraz współczynnik korelacji Pearsona.

Oznaczmy $EX = \mu_1$ i $EY = \mu_2$.

Kowariancją zmiennych losowych X i Y nazywamy liczbę

$$\mathrm{cov}(X, Y) = E(X - \mu_1)(Y - \mu_2).$$

Siłę zależności mierzy współczynnik korelacj Pearsona, definiowany jako wartość oczekiwana iloczynu zmiennych losowych standaryzowanych

$$\rho(X, Y) = E\left(\frac{X-\mu_1}{\sigma_1}\frac{Y-\mu_2}{\sigma_2}\right) = \frac{\mathrm{cov}(X, Y)}{\sigma_1\sigma_2},$$

gdzie $\sigma_1^2 = \mathrm{Var}X$ i $\sigma_2^2 = \mathrm{Var}Y$.

Fakt

Dla dowolnych zmiennych losowych X i Y mamy

a) $\operatorname{cov}(X, Y) = E(XY) - EXEY$,

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

c) jeśli X i Y są zmiennymi niezależnymi, to $\operatorname{cov}(X, Y) = 0$ oraz $\rho(X, Y) = 0$.

Dowód a) Korzystając z własności wartości oczekiwanej, mamy

$$\operatorname{cov}(X, Y) = E(X - \mu_1)(Y - \mu_2)$$

$$= E(XY - \mu_2 X - \mu_1 Y + \mu_1\mu_2)$$
$$= E(XY) - E(\mu_2 X) - E(\mu_1 Y) + E(\mu_1\mu_2)$$
$$= E(XY) - \mu_2 EX - \mu_1 EY + E(\mu_1\mu_2)$$
$$= E(XY) - \mu_2\mu_1 - \mu_1\mu_2 + \mu_1\mu_2$$
$$= E(XY) - \mu_1\mu_2.$$

Współczynnik korelacji Pearsona

Współczynnik ρ przyjmuje wartości z przedziału $\langle -1, 1\rangle$.

- Jeżeli $\rho > 0$ mówimy, że zmienne są skorelowane dodatnio,
- jeżeli $\rho < 0$ – skorelowane ujemnie,
- w przypadku, gdy $\rho = 0$ mówimy, że zmienne są nieskorelowane.

Zachodzą zależności

- $\rho = 1 \iff P(Y{=}aX{+}b) = 1,$ dla pewnych stałych $a > 0, b \in \mathbb{R}$,
- $\rho = -1 \iff P(Y{=}aX{+}b),$ dla pewnych stałych $a < 0, b \in \mathbb{R}$.

Można też łatwo wykazać, że wariancja sumy zmiennych losowych jest równa sumie wariancji oraz podwojonej wartości kowariancji

$$Var(X+Y) = VarX + VarY + 2\mathrm{cov}(X, Y)$$

$$= \sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2.$$

Współczynnik korelacji Pearsona – przykład

Przykład 1. c.d. Zmienne losowe X i Y mają taki sam rozkład. Mamy

- $EX = EY = 0$, $E(X^2) = E(Y^2) = 0,8$, $E(XY) = 0$,
- $VarX = VarY = 0,8$.

Stąd $\rho = 0$. Zauważmy, że zmienne X i Y są zależne.

$Y \backslash X$	-1	0	1	$P(Y=y)$
-1	$0,2$	0	$0,2$	$0,4$
0	0	$0,2$	0	$0,2$
1	$0,2$	0	$0,2$	$0,4$
P(X=x)	$0,4$	$0,2$	$0,4$	

Uwaga.

- X, Y niezależne $\Rightarrow$ X, Y nieskorelowane.
- X, Y nieskorelowane $\not\Rightarrow$ X, Y niezależne.

Współczynnik korelacji Pearsona – przykład

Przykład 3. W tabeli podany jest rozkład łączny wektora (X, Y) oraz rozkłady zmiennych X i Y (rozkłady brzegowe).

$_Y\backslash^X$	-1	0	1	$P(Y=y)$
-1	$0,2$	0	$0,2$	$0,4$
0	0	$0,4$	0	$0,4$
1	0	0	$0,2$	$0,2$
P(X=x)	$0,2$	$0,4$	$0,4$	

Współczynnik korelacji $\rho \approx 0,43$.

Dystrybuanta

Definicja

Dystrybuantą wektora (X, Y) nazywamy funkcję $F : \mathbb{R}^2 \to \mathbb{R}$ taką, że

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

dla każdego $x, y \in \mathbb{R}$.

Rozkłady dwuwymiarowe ciągłe

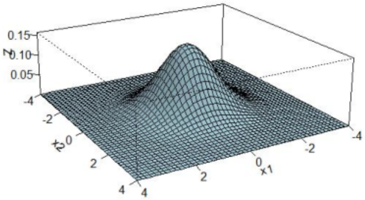
Mówimy, że wektor (X, Y) o dystrybuancie F jest typu ciągłego, jeśli istnieje funkcja nieujemna $f : \mathbb{R}^2 \to \mathbb{R}$ taka, że

$$F(x, y) = \int\limits_{-\infty}^{x} \int\limits_{-\infty}^{y} f(u, v) du dv,$$

dla każdego $x, y \in \mathbb{R}$.

Funkcję f nazywamy gęstością rozkładu wektora (X, Y). Mamy

$$f(x, y) = \frac{\partial^2 F}{\partial x \partial y}(x, y).$$



Fakt

Funkcja f jest gęstością rozkładu pewnego wektora (X, Y) wtedy i tylko wtedy, gdy

- $f(x, y) \geqslant 0$, dla każdego $(x, y) \in \mathbb{R}$,
- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dxdy = 1$.

Prawdopodobieństwo, że wektor (X, Y) przyjmuje wartości ze zbioru $D \subset \mathbb{R}^2$ obliczamy, korzystając z zależności

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

Przykład 4. Wektor (X, Y) ma rozkład o gęstości

$$f(x,y) = \begin{cases} e^{-x-y}, & \text{dla } x > 0 \text{ i } y > 0, \\ 0, & \text{dla pozostałych wartości.} \end{cases}$$

$$P((X,Y) \in (1,2) \times (1,3)) = \int_1^2 \int_1^3 e^{-x-y} dxdy$$

$$= (e^{-1} - e^{-2})(e^{-1} - e^{-3}) \approx 0{,}07.$$

Rozkłady brzegowe

Fakt

Jeśli wektor (X, Y) jest typu ciagłego o gestości f, to zmienne losowe X i Y też są typu ciągłego i ich gęstości możemy wyznaczyć z zależności

$$f_1(x) = \int_{-\infty}^{\infty} f(x,y)dy, \qquad f_2(y) = \int_{-\infty}^{\infty} f(x,y)dx,$$

gdzie f_1 jest gęstością zmiennej losowej X, a f_2 zmiennej losowej Y.

Przykład 4 c.d.

$$f_1(x) = \int_0^{\infty} e^{-x-y}dy = e^{-x}, \text{ dla } x > 0$$

$$f_2(x) = \int_0^{\infty} e^{-x-y}dx = e^{-y}, \text{ dla } y > 0$$

Niezależność zmiennych losowych

Definicja

Zmienne losowe X i Y nazywami niezależnymi, jeśli

$$F(x, y) = F_1(x) \cdot F_2(x),$$

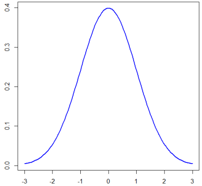
gdzie F_1, F_2 i F są odpowiednio dystrybuantami zmiennych X i Y oraz wektora (X, Y).

Dla rozkładów ciągłych, powyższa definicja jest równoważna warunkowi

$$f(x, y) = f_1(x) \cdot f_2(x), \tag{2}$$

gdzie f_1, f_2 i f są gęstościami zmiennych X i Y oraz wektora (X, Y), natomiast dla rozkładów dyskretnych, warunkowi (1).

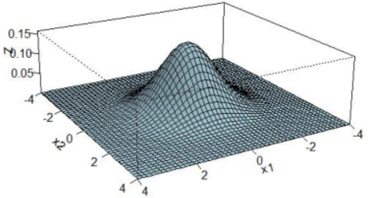
Zmienne losowe z Przykładu 4 spełniają warunek (2), zatem są niezależne.



Gęstość wektora $Z = (Z_1, Z_2)^T$, gdzie zmienne losowe Z_1, Z_2 są niezależne o jednakowym rozkładzie normalnym standardowym $N(0,1)$, jest zgodnie z (2) iloczynem gęstości rozkładów brzegowych

$$f(x_1, x_2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x_1^2}{2}} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{x_2^2}{2}} = \frac{1}{2\pi} e^{-\frac{x_1^2 + x_2^2}{2}}. \tag{3}$$

Rozkład o gęstości f, nazywamy rozkładem normalnym standardowym, dwuwymiarowym.



Rozkład normalny standardowy dwuwymiarowy

Opisując wektor normalny standardowy $\mathrm{Z} = (Z_1, Z_2)$, podajemy wektor wartości oczekiwanych oraz macierz kowariancji, która niesie informacje o zależnościach między zmiennymi.

Wektor warości oczekiwanych: $E(\mathrm{Z}) = (EZ_1, EZ_2) = (0, 0)$.

Macierz kowariancji:

$$\Sigma = \begin{bmatrix} \operatorname{cov}(Z_1, Z_1) & \operatorname{cov}(Z_1, Z_2) \\ \operatorname{cov}(Z_2, Z_1) & \operatorname{cov}(Z_2, Z_2) \end{bmatrix}$$

$$= \begin{bmatrix} \operatorname{Var}(Z_1) & \operatorname{cov}(Z_1, Z_2) \\ \operatorname{cov}(Z_2, Z_1) & \operatorname{Var}(Z_2) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \mathrm{I}_2$$

- Zauważ, że zmienne Z_1 i Z_2 są niezależne, zatem $\operatorname{cov}(Z_1, Z_2) = 0$.

Piszemy $Z \sim N(0, \mathrm{I}_2)$, gdzie I_2 oznacza macierz jednostkową w $\mathbb{R}^2$.

Rozkład normalny dwuwymiarowy

Niech $\mathrm{Z} = (Z_1, Z_2) \sim N_2(0, \mathrm{I}_2)$.

Mówimy, że wektor $\mathrm{X} = (X_1, X_2)$ ma rozkład normalny dwuwymiarowy (rozkład Gaussowski), jeżeli jest transormacją liniową

$$\begin{cases} X_1 = \mu_1 + a_{11} Z_1 + a_{1k} Z_2 \\ X_2 = \mu_2 + a_{21} Z_1 + a_{2k} Z_2 \end{cases} \tag{4}$$

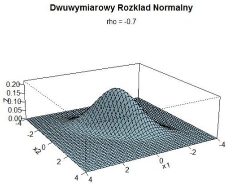
lub w postaci macierzowej

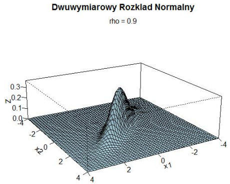
$$\mathrm{X} = \mu + A\mathrm{Z},$$

gdzie $\mu = (\mu_1, \mu_2) \in \mathbb{R}^2$ jest wektorem stałych oraz

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

gdzie A jest macierzą nieosobliwą.





Rozkład normalny dwuwymiarowy

Jeśli wektor X ma rozkład normalny, to wartość oczekiwana EX i macierz kowariancji Σ jednoznacznie wyznaczają rozkład tego wektora.

Wartość oczekiwana: $\mu = EX = (EX_1, EX_2) = (\mu_1, \mu_2)$

Macierz kowariancji:

$$\Sigma = \begin{bmatrix} \operatorname{cov}(X_1, X_1) & \operatorname{cov}(X_1, X_2) \\ \operatorname{cov}(X_2, X_1) & \operatorname{cov}(X_2, X_2) \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix},$$

gdzie σ_1, σ_2 są odchyleniami standardowymi, a ρ jest współczynnikiem korelacji.

Można wykazać też, że $\Sigma = AA^T$.

Piszemy $(X_1, X_2) \sim N(\mu, \Sigma)$ lub $(X_1, X_2) \sim N(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$.

Gęstość rozkładu normalnego dwuwymiarowego

Mając gęstość rozkładu dwuwymiarowego standardowego (3) oraz definicję (4) można wykazać, że gęstość rozkładu normalnego dwuwymiarowego dana jest wzorem

$$f(x_1,x_2)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}$$

$$\exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2}-2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}+\frac{(x_2-\mu_2)^2}{\sigma_2^2}\right]\right). \qquad (5)$$

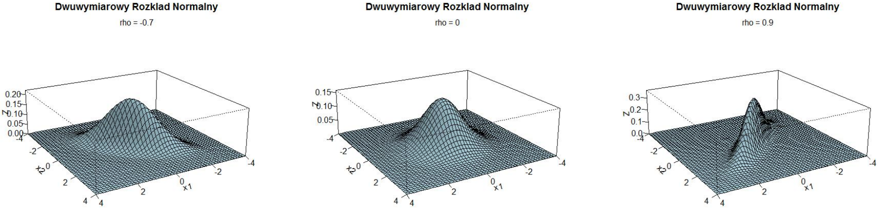
Na przykład gęstość rozkładu $N(0,0,1,1,0.9)$ dana jest wzorem

$$f(x_1,x_2)=\frac{5\sqrt{19}}{19\pi}\exp\left(-\frac{50}{19}\left(x_1^2-\frac{9}{5}x_1x_2+x_2^2\right)\right).$$

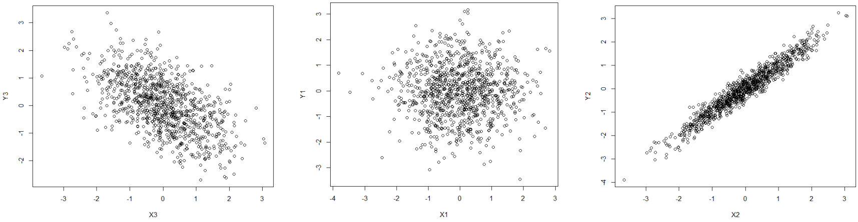
Wykres gęstości na kolejnym slajdzie.

Rozkład normalny dwuwymiarowy – wybrane gęstości

Gęstości rozkładów $N(0, 0, 1, 1, -0{,}7,)$, $N(0, 0, 1, 1, 0)$, $N(0, 0, 1, 1, 0{,}9)$.



Próby wygenerowane z powyższych rozkładów.



Fakt

Niech $(X_1, X_2) \sim N(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$. Wtedy

- rozkłady brzegowe są też rozkładami normalnymi
  $X_1 \sim N(\mu_1, \sigma_1), \qquad X_2 \sim N(\mu_2, \sigma_2)$,
- zmienne X_1 i X_2 są niezależne wtedy i tylko wtedy, gdy są nieskorelowane.

Dla $\rho = 0$ z (5) mamy

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right]\right)$$

$$= \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{1}{2}\frac{(x_1-\mu_1)^2}{\sigma_1^2}\right) \cdot \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{1}{2}\frac{(x_2-\mu_2)^2}{\sigma_2^2}\right) = f_1(x_1) \cdot f_2(x_2).$$

Rozkład normalny d-wymiarowy

Analogicznie do rozkładu dwuwymiarowego (4), możemy zdefiniować rozkład normalny d-wymiarowy o wartości oczekiwanej $EX = (EX_1, \ldots, EX_d)$ i macierzy kowariancji

$$\Sigma = \begin{bmatrix} \mathrm{cov}(X_1, X_1) & \mathrm{cov}(X_1, X_2) & \cdots & \mathrm{cov}(X_1, X_d) \\ \mathrm{cov}(X_2, X_1) & \mathrm{cov}(X_2, X_2) & \cdots & \mathrm{cov}(X_2, X_d) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(X_d, X_1) & \mathrm{cov}(X_d, X_2) & \cdots & \mathrm{cov}(X_d, X_d) \end{bmatrix}$$

które też jednoznacznie wyznaczają ten rozkład.

- Gęstość rozkładu normalnego d-wymiarowego

$$f(x_1, \ldots, x_d) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left[-\frac{1}{2}(\mathrm{x} - \mu)^T \Sigma^{-1} (\mathrm{x} - \mu)\right], \qquad (6)$$

  gdzie $\mathrm{x} = (x_1, \ldots, x_d)^T$, $\mu = (EX_1, \ldots, EX_d)^T$, a $|\Sigma|$ jest wyznacznikiem macierzy kowariancji.

Rozkład normalny d-wymiarowy

- Jeśli wektor $(X_1, \ldots, X_d)$ ma rozkład normalny d-wymiarowy (6), to rozkłady brzegowe też są rozkładami normalnymi $X_i \sim N(\mu_i, \sigma_i^2)$.
- Wektor $(X_1, X_2, \ldots, X_d)$ ma rozkład normalny d-wymiarowy wtedy i tylko wtedy, gdy każda kombinacja liniowa

$$\mathrm{Y} = a_1 X_1 + a_2 X_2 + \ldots + a_d X_d,$$

ma rozkład normalny. Czyli, dla dowolnego wektora $\mathrm{a} \in \mathbb{R}^d$, zmienna losowa $Y = \mathrm{a}^T \mathrm{X}$ ma rozkład normalny.