Analiza danych wielowymiarowych

Martyna Śpiewak

Bootcamp Data Science

Wielowymiarowe zmienne losowe

Istnieje wiele doświadczeń losowych, których wyniki przedstawione są za pomocą par liczb rzeczywistych

- gdy badamy prędkość i drogę zatrzymania się samochodu,
- gdy badamy wzrost i wagę człowieka,
- gdy badamy długość i wytrzymałość włókna bawełny,

lub trójek liczb rzeczywistych

- gdy badamy ciśnienie, objętość i temperaturę gazu.

Do modelowania takich doświadczeń służą zmienne losowe **dwuwymiarowe**, **trójwymiarowe** lub ogólniej zmienne losowe ***n*-wymiarowe.**

Dwuwymiarowe zmienne losowe

Niech X i Y będą zmiennymi losowymi określonymi niekoniecznie na tej samej przestrzeni probabilistycznej.

Parę (X, Y) nazywamy **dwuwymiarową zmienną losową** lub **dwuwymiarowym wektorem losowym**, a X oraz Y jej współrzędnymi.

Dystrybuanta

Dystrybuantą zmiennej losowej (X, Y) nazywamy funkcję rzeczywistą F, która jest określona dla wszystkich liczb rzeczywistych $x, y \in \mathbb{R}$ wzorem

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

Dystrybuanta F posiada następujące własności:

W1 $\forall (x, y) \in \mathbb{R}^2 \quad 0 \leq F(x, y) \leq 1$;

W2 F jest funkcją niemalejąca ze względu na każdy argument;

W3 F jest funkcją co najmniej prawostronnie ciągłą ze względu na każdy z argumentów;

W4 $\forall x \in \mathbb{R} \quad \lim_{y \to -\infty} F(x, y) = 0$, $\forall y \in \mathbb{R} \quad \lim_{x \to -\infty} F(x, y) = 0$ oraz $\lim_{x,y \to +\infty} F(x, y) = 1$.

Zmienna losowa typu dyskretnego

Zmienna losowa (X, Y) jest **typu skokowego**, jeżeli przyjmuje co najwyżej przeliczalną liczbę wartości (x_i, y_k) oraz

$$P(X = x_i, Y = y_k) = p_{ik} \geq 0 \quad \text{dla} \quad i, k = 1, 2, \ldots$$

przy czym

$$\sum_i \sum_k p_{ik} = 1.$$

Jeżeli dwuwymiarowa zmienna (X, Y) przyjmuje skończoną liczbę wartości, to rozkład jest zadawany przez **tabele dwudzielną**, gdzie

$$p_{i\cdot} = P(X = x_i) = \sum_k P(X = x_i, Y = y_k)$$

$$p_{\cdot k} = P(Y = y_k) = \sum_i P(X = x_i, Y = y_k).$$

Zmienna losowa typu dyskretnego

		x_i				
		x_1	x_2	$\dots$	x_m	$p_{\cdot k}$
y_k	y_1	p_{11}	p_{21}	$\dots$	p_{m1}	$p_{\cdot 1}$
	y_2	p_{12}	p_{22}	$\dots$	p_{m2}	$p_{\cdot 2}$
	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
	y_s	p_{1s}	p_{2s}	$\dots$	p_{ms}	$p_{\cdot s}$
	$p_{i\cdot}$	$p_{1\cdot}$	$p_{2\cdot}$	$\dots$	$p_{m\cdot}$	$\sum = 1$

Rozkład dyskretny — przykład

Pewien handlarz trudni się sprzedażą używanych samochodów i motocykli. Niech X oznacza liczbę samochodów, natomiast Y liczbę motocykli, jaką udaje się handlarzowi sprzedać w ciągu jednego dnia. Rozkład prawdopodobieństwa zmiennej losowej (X, Y) jest następujący:

Rozkład wektora losowego:

		X		
		0	1	2
Y	0	0.1	0.2	0.1
	1	0.1	0.2	0.3

Rozkład dyskretny — przykład

Pewien handlarz trudni się sprzedażą używanych samochodów i motocykli. Niech X oznacza liczbę samochodów, natomiast Y liczbę motocykli, jaką udaje się handlarzowi sprzedać w ciągu jednego dnia. Rozkład prawdopodobieństwa zmiennej losowej (X, Y) jest następujący:

Rozkład wektora losowego:

		X		
		0	1	2
Y	0	0.1	0.2	0.1
	1	0.1	0.2	0.3

- $P(X = 0, Y = 1)$

Rozkład dyskretny — przykład

Pewien handlarz trudni się sprzedażą używanych samochodów i motocykli. Niech X oznacza liczbę samochodów, natomiast Y liczbę motocykli, jaką udaje się handlarzowi sprzedać w ciągu jednego dnia. Rozkład prawdopodobieństwa zmiennej losowej (X, Y) jest następujący:

Rozkład wektora losowego:

		X		
		0	1	2
Y	0	0.1	0.2	0.1
	1	0.1	0.2	0.3

- $P(X = 0, Y = 1) = 0.1$;

Rozkład dyskretny — przykład

Pewien handlarz trudni się sprzedażą używanych samochodów i motocykli. Niech X oznacza liczbę samochodów, natomiast Y liczbę motocykli, jaką udaje się handlarzowi sprzedać w ciągu jednego dnia. Rozkład prawdopodobieństwa zmiennej losowej (X, Y) jest następujący:

Rozkład wektora losowego:

		X		
		0	1	2
Y	0	0.1	0.2	0.1
	1	0.1	0.2	0.3

- $P(X = 0, Y = 1) = 0.1$;
- $P(X = 1)$

Rozkład dyskretny — przykład

Pewien handlarz trudni się sprzedażą używanych samochodów i motocykli. Niech X oznacza liczbę samochodów, natomiast Y liczbę motocykli, jaką udaje się handlarzowi sprzedać w ciągu jednego dnia. Rozkład prawdopodobieństwa zmiennej losowej (X, Y) jest następujący:

Rozkład wektora losowego:

		X		
		0	1	2
Y	0	0.1	0.2	0.1
	1	0.1	0.2	0.3

- $P(X = 0, Y = 1) = 0.1$;
- $P(X = 1) = P(X = 1, Y = 0) + P(X = 1, Y = 1) = 0.2 + 0.2 = 0.4$;

Rozkład dyskretny — przykład

Pewien handlarz trudni się sprzedażą używanych samochodów i motocykli. Niech X oznacza liczbę samochodów, natomiast Y liczbę motocykli, jaką udaje się handlarzowi sprzedać w ciągu jednego dnia. Rozkład prawdopodobieństwa zmiennej losowej (X, Y) jest następujący:

Rozkład wektora losowego:

		X		
		0	1	2
Y	0	0.1	0.2	0.1
	1	0.1	0.2	0.3

- $P(X = 0, Y = 1) = 0.1$;
- $P(X = 1) = P(X = 1, Y = 0) + P(X = 1, Y = 1) = 0.2 + 0.2 = 0.4$;
- $F(1, 1) =$

Rozkład dyskretny — przykład

Pewien handlarz trudni się sprzedażą używanych samochodów i motocykli. Niech X oznacza liczbę samochodów, natomiast Y liczbę motocykli, jaką udaje się handlarzowi sprzedać w ciągu jednego dnia. Rozkład prawdopodobieństwa zmiennej losowej (X, Y) jest następujący:

Rozkład wektora losowego:

		X		
		0	1	2
Y	0	0.1	0.2	0.1
	1	0.1	0.2	0.3

- $P(X = 0, Y = 1) = 0.1$;
- $P(X = 1) = P(X = 1, Y = 0) + P(X = 1, Y = 1) = 0.2 + 0.2 = 0.4$;
- $F(1, 1) = P(X \leq 1, Y \leq 1)$

Rozkład dyskretny — przykład

Pewien handlarz trudni się sprzedażą używanych samochodów i motocykli. Niech X oznacza liczbę samochodów, natomiast Y liczbę motocykli, jaką udaje się handlarzowi sprzedać w ciągu jednego dnia. Rozkład prawdopodobieństwa zmiennej losowej (X, Y) jest następujący:

Rozkład wektora losowego:

		X		
		0	1	2
Y	0	0.1	0.2	0.1
	1	0.1	0.2	0.3

- $P(X = 0, Y = 1) = 0.1$;
- $P(X = 1) = P(X = 1, Y = 0) + P(X = 1, Y = 1) = 0.2 + 0.2 = 0.4$;
- $F(1, 1) = P(X \leq 1, Y \leq 1) = P(X = 0, Y = 0) + P(X = 0, Y = 1) + P(X = 1, Y = 0) + P(X = 1, Y = 1) = 0.1 + 0.1 + 0.2 + 0.2 = 0.6$.

Zmienna losowa typu dyskretnego — rozkład wielomianowy

Zmienna losowa $(X_1, X_2, \ldots, X_k)$ ma rozkład wielomianowy, jeżeli

$$P(X_1 = n_1, \ldots, X_k = n_k) = \frac{n!}{n_1! \cdot \ldots \cdot n_k!} p_1^{n_1} \cdot \ldots \cdot p_k^{n_k},$$

gdzie $p_i \in (0, 1)$, $i = 1, 2, \ldots, k$, $p_1 + p_2 + \ldots + p_k = 1$, $n_1 + n_2 + \ldots + n_k = n$.

Rozkład wielomianowy jest uogólnieniem rozkładu dwumianowego i opisuje rozkład wyników przy n-krotnym powtórzeniu doświadczenia o k możliwych rezultatach.

X_i oznacza liczbę wyników i-tego typu w serii.

Rozkład wielomianowy można wykorzystać do obliczenia prawdopodobieństwa w sytuacjach, w których istnieją więcej niż dwa możliwe wyniki.

Rozkład wielomianowy — przykład

Na podstawie historycznych rozgrywek dwóch znanych szachistów ustalono, że prawdopodobieństwo wygranej przez gracza A wynosi 0.40, prawdopodobieństwo wygranej przez gracza B wynosi 0.35, a prawdopodobieństwo, że gra zakończy się remisem, wynosi 0.25 . Jeśli ci dwaj zawodnicy rozegrali 12 partii szachów, jakie jest prawdopodobieństwo, że gracz A wygra 7 gier, gracz B wygra 2 rozgrywki, a pozostałe 3 gry zakończą się remisem?

- X_1 — zmienna losowa opisująca liczbę wygranych partii szachowych przez gracza *A*;
- X_2 — zmienna losowa opisująca liczbę wygranych partii szachowych przez gracza *B*;
- X_3 — zmienna losowa opisująca liczbę partii zakończonych remisem;

Rozkład wielomianowy — przykład

- n jest całkowitą liczbą zdarzeń: $n = 12$;
- n_1 oznacza liczbę wygranych gracza A: $n_1 = 7$;
- n_2 oznacza liczbę wygranych gracza B: $n_2 = 2$;
- n_3 oznacza liczbę rozgrywek zakończonych remisem: $n_3 = 3$;
- p_1 oznacza prawdopodobieństwo wygranej przez gracza A: $p_1 = 0.40$;
- p_2 oznacza prawdopodobieństwo wygranej przez gracza B: $p_1 = 0.35$;
- p_3 oznacza prawdopodobieństwo remisu: $p_3 = 0.25$;

Rozkład wielomianowy — przykład

- n jest całkowitą liczbą zdarzeń: $n = 12$;
- n_1 oznacza liczbę wygranych gracza A: $n_1 = 7$;
- n_2 oznacza liczbę wygranych gracza B: $n_2 = 2$;
- n_3 oznacza liczbę rozgrywek zakończonych remisem: $n_3 = 3$;
- p_1 oznacza prawdopodobieństwo wygranej przez gracza A: $p_1 = 0.40$;
- p_2 oznacza prawdopodobieństwo wygranej przez gracza B: $p_1 = 0.35$;
- p_3 oznacza prawdopodobieństwo remisu: $p_3 = 0.25$;

$$P(X_1 = 7, X_2 = 2, X_3 = 3) = \frac{12!}{7! \cdot 2! \cdot 3!}(0.4)^7 \cdot (0.35)^2 \cdot (0.25)^3 = 0.0248.$$

Zmienna losowa typu ciągłego

Zmienna losowa (X, Y) jest **typu ciągłego**, jeżeli istnieje nieujemna funkcja f, zwana *gęstością*, taka, że dystrybuantę tej zmiennej losowej można przedstawić w postaci

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, t) du dt \quad \text{dla} \quad (x, y) \in \mathbb{R}^2.$$

Jeżeli zmienna losowa (X, Y) jest typu ciągłego o gęstości f, to

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(u, t) du dt = 1 \tag{1}$$

oraz w punktach ciągłości gęstości f zachodzi

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

Zmienna losowa typu ciągłego — przykład

Zmienna losowa (X, Y) ma rozkład ciągły o gęstości f określonej wzorem

$$f(x,y) = \begin{cases} Cx & \text{dla} \quad 0 < x, y < 1 \\ 0 & \text{wpp} \end{cases}$$

a) Wyznaczyć stałą C.

Zmienna losowa typu ciągłego — przykład

Zmienna losowa (X, Y) ma rozkład ciągły o gęstości f określonej wzorem

$$f(x,y) = \begin{cases} Cx & \text{dla} \quad 0 < x, y < 1 \\ 0 & \text{wpp} \end{cases}$$

a) Wyznaczyć stałą C.

$$1 = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x,y)dxdy = \int_0^1 \left(\int_0^1 Cxdx \right) dy$$
$$= \int_0^1 \frac{C}{2} dy = \frac{C}{2} \implies C = 2.$$

b) Obliczyć wartość dystrybuanty $F\left(\frac{1}{2}, 2\right)$.

$$\begin{aligned} F\left(\frac{1}{2}, 2\right) &= \int_{-\infty}^{2} \int_{-\infty}^{\frac{1}{2}} f(x,y)\,dx\,dy \\ &= \int_{0}^{2} \left(\int_{0}^{\frac{1}{2}} 2x\,dx \right) dy \\ &= \int_{0}^{2} \frac{1}{4}\,dy = \frac{1}{2}. \end{aligned}$$

Dwuwymiarowy rozkład normalny

Zmienna losowa (X, Y) ma rozkład dwuwymiarowy normalny z parametrami μ_X, μ_Y, $\sigma_X > 0$, $\sigma_Y > 0$ oraz $\rho \in (-1, 1)$, jeżeli jej gęstość f wyraża się wzorem

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right\}.$$

Dwuwymiarowy rozkład normalny

Zmienna losowa (X, Y) ma rozkład dwuwymiarowy normalny z parametrami μ_X, μ_Y, $\sigma_X > 0$, $\sigma_Y > 0$ oraz $\rho \in (-1, 1)$, jeżeli jej gęstość f wyraża się wzorem

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right\}.$$

Jeżeli dwuwymiarowa zmienna losowa (X, Y) ma rozkład normalny, to zmienne losowe X i Y mają, odpowiednio, jednowymiarowe rozkłady normalne $\mathcal{N}(\mu_X, \sigma_X)$ i $\mathcal{N}(\mu_Y, \sigma_Y)$.

Dystrybuanty brzegowe

Twierdzenie. Jeśli F jest dystrybuantą zmiennej losowej dwuwymiarowej (X, Y), to funkcja

$$F_X(x) = F(x, \infty) \quad \text{dla } x \in \mathbb{R}$$

jest dystrybuantą zmiennej losowej X, zaś funkcja

$$F_Y(y) = F(\infty, y) \quad \text{dla } y \in \mathbb{R}$$

jest dystrybuantą zmiennej losowej Y.

Z twierdzenia wynika, że rozkład prawdopodobieństwa zmiennej losowej dwuwymiarowej (X, Y) wyznacza rozkłady prawdopodobieństwa zmiennych losowych X i Y.

Rozkłady brzegowe zmiennych losowych X i Y nie wyznaczają rozkładu dwuwymiarowej zmiennej losowej (X, Y).

Rozkład brzegowy dla zmiennej losowej typu skokowego — przykład

Rozkład zmiennej losowej (X, Y)
z wyznaczonymi rozkładami
brzegowymi:

		X			
		0	1	2	$p_{\cdot i}$
Y	0	0.1	0.2	0.1	0.4
	1	0.1	0.2	0.3	0.6
	$p_{i\cdot}$	0.2	0.4	0.4	1

Rozkład brzegowy dla zmiennej losowej typu skokowego — przykład

Rozkład zmiennej losowej (X, Y) z wyznaczonymi rozkładami brzegowymi:

		X			
		0	1	2	$p_{\cdot i}$
Y	0	0.1	0.2	0.1	0.4
	1	0.1	0.2	0.3	0.6
	$p_{i\cdot}$	0.2	0.4	0.4	1

Rozkład brzegowy zmiennej X:

x_i	0	1	2
p_i	0.2	0.4	0.4

Rozkład brzegowy zmiennej Y:

y_i	0	1
p_i	0.4	0.6

Rozkład brzegowy dla zmiennej losowej typu ciągłego

Podobnie jak w przypadku zmiennej losowej typu skokowego, można wyznaczyć gęstości brzegowe dla poszczególnych współrzędnych. Funkcje

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

są gęstościami, odpowiednio, zmiennych losowych X i Y.

Rozkład brzegowy dla zmiennej losowej typu ciągłego — przykład

- Gęstość brzegowa dla zmiennej losowej X:

$$f_X(x) = \int_{-\infty}^{+\infty} f(x,y)dy = \int_0^1 2x dy = 2x \implies f_X(x) = \begin{cases} 2x & \text{dla} \quad 0 < x < 1 \\ 0 & \text{wpp} \end{cases}$$

Rozkład brzegowy dla zmiennej losowej typu ciągłego — przykład

- Gęstość brzegowa dla zmiennej losowej X:

$$f_X(x) = \int_{-\infty}^{+\infty} f(x,y)dy = \int_0^1 2x dy = 2x \implies f_X(x) = \begin{cases} 2x & \text{dla} \quad 0 < x < 1 \\ 0 & \text{wpp} \end{cases}$$

- Gęstość brzegowa dla zmiennej losowej Y:

$$f_Y(y) = \int_{-\infty}^{+\infty} f(x,y)dx = \int_0^1 2x dx = 1 \implies f_Y(y) = \begin{cases} 1 & \text{dla} \quad 0 < y < 1 \\ 0 & \text{wpp} \end{cases}$$

Rozkład warunkowy

Rozkład zmiennej losowej $(X|Y = y)$ nazywamy **rozkładem warunkowym** zmiennej losowej X przy ustalonej wartości zmiennej losowej Y.

- Jeśli wektor losowy (X, Y) ma rozkład dyskretny oraz $P(Y = y) > 0$, to rozkład warunkowy zmiennej losowej X pod warunkiem, że $Y = y$ określamy wzorem

$$P(X \in A|Y = y) = \frac{P(X \in A, Y = y)}{P(Y = y)}.$$

- Jeśli wektor (X, Y) ma rozkład ciągły z gęstością $f(x, y)$ oraz $f_Y(y) > 0$, to gęstością rozkładu warunkowego X pod warunkiem $Y = y$ nazywamy funkcję określoną wzorem

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)}.$$

Rozkład warunkowy dla zmiennej losowej typu dyskretnego — przykład

Chcemy wyznaczać rozkład $X|Y=1$, tj. rozkład liczby zakupionych samochodów pod warunkiem sprzedaży dokładnie 1 motocykla dziennie.

		X			
		0	1	2	$p_{\cdot i}$
Y	0	0.1	0.2	0.1	0.4
	1	0.1	0.2	0.3	0.6
	$p_{i\cdot}$	0.2	0.4	0.4	1

Rozkład warunkowy dla zmiennej losowej typu dyskretnego — przykład

Chcemy wyznaczać rozkład $X|Y=1$, tj. rozkład liczby zakupionych samochodów pod warunkiem sprzedaży dokładnie 1 motocykla dziennie.

		X			
		0	1	2	$p_{\cdot i}$
Y	0	0.1	0.2	0.1	0.4
	1	0.1	0.2	0.3	0.6
	$p_{i\cdot}$	0.2	0.4	0.4	1

Musimy wyznaczyć następujące prawdopodobieństwa:

$$P(X=0|Y=1) = \frac{P(X=0,Y=1)}{P(Y=1} = \frac{0.1}{0.6} = \frac{1}{6},$$

$$P(X=1|Y=1) = \frac{P(X=1,Y=1)}{P(Y=1} = \frac{0.2}{0.6} = \frac{2}{6},$$

$$P(X=1|Y=1) = \frac{P(X=2,Y=1)}{P(Y=1} = \frac{0.3}{0.6} = \frac{3}{6}.$$

Rozkład warunkowy dla zmiennej losowej typu ciągłego — przykład

Chcemy wyznaczyć rozkład warunkowy $X|Y$ dla wektora losowego (X, Y) z daną gęstością:

$$f(x,y) = \begin{cases} \frac{3(x-y)^2}{8} & \text{dla} \quad -1 \leq x, y \leq 1 \\ 0 & \text{wpp} \end{cases}$$

Rozkład warunkowy dla zmiennej losowej typu ciągłego — przykład

Chcemy wyznaczyć rozkład warunkowy $X|Y$ dla wektora losowego (X, Y) z daną gęstością:

$$f(x,y) = \begin{cases} \frac{3(x-y)^2}{8} & \text{dla} \quad -1 \leq x, y \leq 1 \\ 0 & \text{wpp} \end{cases}$$

Musimy wyznaczyć gęstość brzegową f_Y:

$$f_Y(y) = \int_{-\infty}^{+\infty} f(x,y)dx = \frac{3}{8}\int_{-1}^{1}(x^2 - 2xy + y^2)dx = \frac{3y^2+1}{4} \quad \text{dla} \quad -1 \leq y \leq 1.$$

Rozkład warunkowy dla zmiennej losowej typu ciągłego — przykład

Wówczas warunkową gęstością prawdopodobieństwa zmiennej losowej X pod warunkiem, że $Y = y$ jest postaci:

$$f(x|y) = \frac{f(x,y)}{f_Y(y)} = \frac{3(x-y)^2}{8} : \frac{3y^2+1}{4} = \frac{3(x-y)^2}{6y^2+2} \quad \text{dla} \quad -1 \leq x \leq 1.$$

Niezależność zmiennych losowych

Zmienne losowe X i Y nazywamy **niezależnymi zmiennymi losowymi**, jeżeli dla dowolnych zbiorów A i B na prostej zachodzi równość

$$P(X \in A, Y \in B) = P(X \in A) \cdot P(Y \in B).$$

1. Zmienne losowe dyskretne X i Y są niezależne, wtedy i tylko wtedy, gdy

$$P(X = x_i, Y = y_k) = P(X = x_i) \cdot P(Y = y_k) \quad \text{dla} \quad \forall i, k = 1, 2 \dots$$

2. Warunkiem koniecznymi i wystarczającym niezależności zmiennych losowych X i Y o gęstościach brzegowych f_X i f_Y jest

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

gdzie f jest gęstością dwuwymiarowej zmiennej losowej (X, Y).

Niezależność zmiennych losowych typu dyskretnego — przykład

		X			
		0	1	2	$p_{\cdot i}$
Y	0	0.1	0.2	0.1	0.4
	1	0.1	0.2	0.3	0.6
	$p_{i\cdot}$	0.2	0.4	0.4	1

Niezależność zmiennych losowych typu dyskretnego — przykład

		X			
		0	1	2	$p_{\cdot i}$
Y	0	0.1	0.2	0.1	0.4
	1	0.1	0.2	0.3	0.6
	$p_{i\cdot}$	0.2	0.4	0.4	1

Aby sprawdzić, czy zmienne losowe X i Y są niezależne, należy ocenić czy zachodzi równość

$$P(X = x_i, Y = y_k) = P(X = x_i) \cdot P(Y = y_k)$$

dla każdej pary (i, j).

Niezależność zmiennych losowych typu dyskretnego — przykład

		X			
		0	1	2	$p_{\cdot i}$
Y	0	0.1	0.2	0.1	0.4
	1	0.1	0.2	0.3	0.6
	$p_{i\cdot}$	0.2	0.4	0.4	1

Aby sprawdzić, czy zmienne losowe X i Y są niezależne, należy ocenić czy zachodzi równość

$$P(X = x_i, Y = y_k) = P(X = x_i) \cdot P(Y = y_k)$$

dla każdej pary (i, j).

$$L = P(X = 0, Y = 0) = 0.1$$
$$P = P(X = 0)(Y = 0) = 0.2 \cdot 0.4 = 0.08$$

$L \neq P$, czyli zmienne X i Y nie są niezależne.

Niezależność zmiennych losowych typu ciągłego — przykład

Chcemy sprawdzić niezależność zmiennych X i Y, których rozkład łączny dany jest gęstością:

$$f(x,y) = \begin{cases} \frac{3(x-y)^2}{8} & \text{dla} \quad -1 \leq x,y \leq 1 \\ 0 & \text{wpp} \end{cases}$$

Wiemy, że

$$f_Y(y) = \frac{3y^2+1}{4} \quad \text{dla} \quad -1 \leq y \leq 1.$$

Wyznaczymy postać $f_X(x)$:

$$f_X(x) = \int_{-\infty}^{+\infty} f(x,y)dy = \frac{3}{8}\int_{-1}^{1}(x^2 - 2xy + y^2)dy = \frac{3x^2+1}{4} \quad \text{dla} \quad -1 \leq x \leq 1.$$

Niezależność zmiennych losowych typu ciągłego — przykład

Sprawdzamy równość $f(x,y) = f_X(x) \cdot f_Y(y)$:

$$L = f(x,y) = \frac{3x^2 - 6xy + 3y^2}{8},$$
$$P = f_X(x) \cdot f_Y(y) = \frac{3x^2+1}{4} \cdot \frac{3y^2+1}{4} = \frac{9x^2y^2 + 3x^2 + 3y^2 + 1}{16}.$$

$L \neq P$, stąd zmienne losowe X i Y nie są niezależne.

Kowariancja

Kowariancją zmiennych losowych X i Y nazywamy liczbę $\mathrm{Cov}(X, Y)$ określoną wzorem

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

Kowariancja może być traktowana jako pewna miara zgodności dwóch zmiennych losowych, będąc średnią wartością iloczynu odchyleń obu zmiennych od ich wartości oczekiwanych.

Bezpośrednio z definicji wynika, że

$$\mathrm{Cov}(X, Y) = \mathbb{E}(XY) - (\mathbb{E}X)(\mathbb{E}Y).$$

Kowariancja

Gdy zmienna losowa (X, Y) jest typu dyskretnego, to

$$\mathbb{E}(XY) = \sum_i \sum_k x_i y_k P(X = x_i, Y = y_k).$$

Gdy zmienna losowa (X, Y) jest typu ciągłego, to

$$\mathbb{E}(XY) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xyf(x, y)dxdy.$$

Własności:

- Jeżeli $\mathrm{Cov}(X, Y) = 0$, to zmienne losowe X i Y nazywamy nieskorelowanymi. Wynika stąd, że zmienne losowe niezależne są jednocześnie nieskorelowane, ale zmienne losowe nieskorelowane mogą być zależne.
- $\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \pm 2\mathrm{Cov}(X, Y)$.
- $\mathrm{Cov}(aX + b, cY + d) = ac\mathrm{Cov}(X, Y)$.

Macierz kowariancji

Niech $(X_1, X_2, \ldots, X_n)$ będzie wektorem losowym (o składowych całkowalnych z kwadratem), wtedy **macierz kowariancji** jest określona następująco:

$$\sum = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \ldots & \sigma_{2n} \\ \vdots & \ldots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \ldots & \sigma_n^2 \end{bmatrix}$$

gdzie

- σ_i^2 jest wariancją zmiennej losowej X_i;
- $\sigma_{ij} = \mathrm{Cov}(X_i, X_j)$ jest kowariancją między zmiennymi losowym X_i oraz X_j.

Współczynnik korelacji

Współczynnikiem korelacji zmiennych losowych X i Y nazywamy liczbę

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

Własności:

- $|\rho(X, Y)| \leq 1$.
- Jeżeli zmienne losowe X i Y są niezależne, to $\rho(X, Y) = 0$.
- Dla dowolnych liczb rzeczywistych a, b, c, d zachodzi

$$|\rho(aX + b, cY + d)| = |\rho(X, Y)|.$$

- Zmienne losowe X i Y są zależne liniowo wtedy i tylko wtedy, gdy $|\rho(X, Y)| = 1$

N-wymiarowy rozkład normalny

Wektor losowy $(X_1, \ldots, X_n)$ ma rozkład N-wymiarowy rozkład normalny z macierzą kowariancji $\sum$ oraz wektorem średnich μ, jeżeli jej gęstość f wyraża się wzorem

$$f(x_1, \ldots, x_n) = \frac{1}{\sqrt{(2\pi)^n |\sum|}} \exp\left\{ -\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \right\}.$$

Centralne twierdzenie graniczne Lindeberga-Levy'ego

Jeżeli zmienne losowe $X_1, X_2, \ldots$ są niezależne o jednakowych rozkładach z parametrami $\mathbb{E}X_k = \mu$, $\mathrm{Var}X_k = \sigma^2$ dla $k = 1, 2, \ldots$, to

$$\lim_{n\to\infty} P\left(a < \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b\right) = \Phi(b) - \Phi(a),$$

gdzie Φ jest dystrybuantą rozkładu normalnego $\mathcal{N}(0, 1)$.

Centralne twierdzenie graniczne Lindeberga-Levy'ego — przykład

Rozkład prawdopodobieństwa miesięcznych zarobków absolwentów szkół wyższych jest normalny ze średnią 5000zł i odchyleniem standardowym 1000zł. Jakie jest prawdopodobieństwo, że miesięczne wynagrodzenie losowo wybranego absolwenta przekracza 6000zł? Jakie jest prawdopodobieństwo, że średnia płaca wyliczona dla 16 osobowej grupy absolwentów przekroczy 6000zł?

Centralne twierdzenie graniczne Lindeberga-Levy'ego — przykład

Rozkład prawdopodobieństwa miesięcznych zarobków absolwentów szkół wyższych jest normalny ze średnią 5000zł i odchyleniem standardowym 1000zł. Jakie jest prawdopodobieństwo, że miesięczne wynagrodzenie losowo wybranego absolwenta przekracza 6000zł? Jakie jest prawdopodobieństwo, że średnia płaca wyliczona dla 16 osobowej grupy absolwentów przekroczy 6000zł?

X_i — zmienna losowa opisująca miesięczne zarobki i-tego absolwenta szkoły wyższej;

Centralne twierdzenie graniczne Lindeberga-Levy'ego — przykład

Rozkład prawdopodobieństwa miesięcznych zarobków absolwentów szkół wyższych jest normalny ze średnią 5000zł i odchyleniem standardowym 1000zł. Jakie jest prawdopodobieństwo, że miesięczne wynagrodzenie losowo wybranego absolwenta przekracza 6000zł? Jakie jest prawdopodobieństwo, że średnia płaca wyliczona dla 16 osobowej grupy absolwentów przekroczy 6000zł?

X_i — zmienna losowa opisująca miesięczne zarobki i-tego absolwenta szkoły wyższej;

$X_i \sim \mathcal{N}(\mu = 5000, \sigma = 1000)$.

Centralne twierdzenie graniczne Lindeberga-Levy'ego — przykład

Rozkład prawdopodobieństwa miesięcznych zarobków absolwentów szkół wyższych jest normalny ze średnią 5000zł i odchyleniem standardowym 1000zł. Jakie jest prawdopodobieństwo, że miesięczne wynagrodzenie losowo wybranego absolwenta przekracza 6000zł? Jakie jest prawdopodobieństwo, że średnia płaca wyliczona dla 16 osobowej grupy absolwentów przekroczy 6000zł?

X_i — zmienna losowa opisująca miesięczne zarobki i-tego absolwenta szkoły wyższej;

$X_i \sim \mathcal{N}(\mu = 5000, \sigma = 1000)$.

$P(X_i > 6000) =$

Centralne twierdzenie graniczne Lindeberga-Levy'ego — przykład

Rozkład prawdopodobieństwa miesięcznych zarobków absolwentów szkół wyższych jest normalny ze średnią 5000zł i odchyleniem standardowym 1000zł. Jakie jest prawdopodobieństwo, że miesięczne wynagrodzenie losowo wybranego absolwenta przekracza 6000zł? Jakie jest prawdopodobieństwo, że średnia płaca wyliczona dla 16 osobowej grupy absolwentów przekroczy 6000zł?

X_i — zmienna losowa opisująca miesięczne zarobki i-tego absolwenta szkoły wyższej;

$X_i \sim \mathcal{N}(\mu = 5000, \sigma = 1000)$.

$$P(X_i > 6000) = 1 - P(X_i \leq 6000) =$$

Centralne twierdzenie graniczne Lindeberga-Levy'ego — przykład

Rozkład prawdopodobieństwa miesięcznych zarobków absolwentów szkół wyższych jest normalny ze średnią 5000zł i odchyleniem standardowym 1000zł. Jakie jest prawdopodobieństwo, że miesięczne wynagrodzenie losowo wybranego absolwenta przekracza 6000zł? Jakie jest prawdopodobieństwo, że średnia płaca wyliczona dla 16 osobowej grupy absolwentów przekroczy 6000zł?

X_i — zmienna losowa opisująca miesięczne zarobki i-tego absolwenta szkoły wyższej;

$X_i \sim \mathcal{N}(\mu = 5000, \sigma = 1000)$.

$$P(X_i > 6000) = 1 - P(X_i \leq 6000) = 1 - F_{\mathcal{N}(\mu=5000,\sigma=1000)}(6000) = 0.1586553.$$

Centralne twierdzenie graniczne Lindeberga-Levy'ego — przykład

$\overline{X}_{16}$ — zmienna losowa, opisująca średnią miesięczną płacę na 16 losowo wybranych absolwentów szkół wyższych.

Zauważmy, że prawdziwa jest zależność $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{S_n}{n}$.

Centralne twierdzenie graniczne Lindeberga-Levy'ego — przykład

$\overline{X}_{16}$ — zmienna losowa, opisująca średnią miesięczną płacę na 16 losowo wybranych absolwentów szkół wyższych.

Zauważmy, że prawdziwa jest zależność $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{S_n}{n}$.

$$
\begin{aligned}
P(\overline{X}_{16} > 6000) &= P\left(\frac{S_{16}}{16} > 6000\right) = P(S_{16} > 16 \cdot 6000) \\
&= 1 - P(S_{16} \leq 16 \cdot 6000) \\
&= 1 - P\Big(\frac{S_{16} - n\mu}{\sqrt{n}\sigma} \leq \frac{16 \cdot 6000 - n\mu}{\sqrt{n}\sigma}\Big) \\
&\cong 1 - \Phi\Big(\frac{16 \cdot 6000 - n\mu}{\sqrt{n}\sigma}\Big) = 1 - \Phi\Big(\frac{16 \cdot 6000 - 16 \cdot 5000}{\sqrt{16} \cdot 1000}\Big) \\
&= 1 - \Phi(4) \approx 0
\end{aligned}
$$

Centralne twierdzenie graniczne Moivre'a-Laplace'a

Jeżeli zmienne losowe $X_1, X_2, \ldots$ są niezależne o jednakowych rozkładach dwupunktowych Bern(p), to

$$\lim_{n\to\infty} P\left(a < \frac{S_n - np}{\sqrt{np(1-p)}} \leq b\right) = \Phi(b) - \Phi(a).$$

Centralne twierdzenie graniczne Moivre'a-Laplace'a — przykład

W hotelu jest 100 pokoi. Właściciel hotelu polecił przyjmować rezerwacje na więcej niż 100 pokoi, ponieważ z doświadczenia wie, że jedynie 90% dokonywanych wcześniej rezerwacji jest później wykorzystywanych. Jakie jest prawdopodobieństwo, że przy przyjęciu 104 rezerwacji w hotelu zabraknie wolnych miejsc?

Centralne twierdzenie graniczne Moivre'a-Laplace'a — przykład

W hotelu jest 100 pokoi. Właściciel hotelu polecił przyjmować rezerwacje na więcej niż 100 pokoi, ponieważ z doświadczenia wie, że jedynie 90% dokonywanych wcześniej rezerwacji jest później wykorzystywanych. Jakie jest prawdopodobieństwo, że przy przyjęciu 104 rezerwacji w hotelu zabraknie wolnych miejsc?

X_i — zmienna losowe opisujące, że i-ta rezerwacja zostanie zrealizowana;

Centralne twierdzenie graniczne Moivre'a-Laplace'a — przykład

W hotelu jest 100 pokoi. Właściciel hotelu polecił przyjmować rezerwacje na więcej niż 100 pokoi, ponieważ z doświadczenia wie, że jedynie 90% dokonywanych wcześniej rezerwacji jest później wykorzystywanych. Jakie jest prawdopodobieństwo, że przy przyjęciu 104 rezerwacji w hotelu zabraknie wolnych miejsc?

X_i — zmienna losowe opisujące, że i-ta rezerwacja zostanie zrealizowana;
$X_i \sim \text{Bern}(p = 0.9)$.

Centralne twierdzenie graniczne Moivre'a-Laplace'a — przykład

W hotelu jest 100 pokoi. Właściciel hotelu polecił przyjmować rezerwacje na więcej niż 100 pokoi, ponieważ z doświadczenia wie, że jedynie 90% dokonywanych wcześniej rezerwacji jest później wykorzystywanych. Jakie jest prawdopodobieństwo, że przy przyjęciu 104 rezerwacji w hotelu zabraknie wolnych miejsc?

X_i — zmienna losowe opisujące, że i-ta rezerwacja zostanie zrealizowana;
$X_i \sim \text{Bern}(p = 0.9)$.

$S_{104} = \sum_{i=1}^{104} X_i$ — zmienna losowa, opisująca liczbę wykorzystanych rezerwacji spośród 104 przyjętych rezerwacji;

Centralne twierdzenie graniczne Moivre'a-Laplace'a — przykład

W hotelu jest 100 pokoi. Właściciel hotelu polecił przyjmować rezerwacje na więcej niż 100 pokoi, ponieważ z doświadczenia wie, że jedynie 90% dokonywanych wcześniej rezerwacji jest później wykorzystywanych. Jakie jest prawdopodobieństwo, że przy przyjęciu 104 rezerwacji w hotelu zabraknie wolnych miejsc?

X_i — zmienna losowe opisujące, że i-ta rezerwacja zostanie zrealizowana;
$X_i \sim \text{Bern}(p = 0.9)$.

$S_{104} = \sum_{i=1}^{104} X_i$ — zmienna losowa, opisująca liczbę wykorzystanych rezerwacji spośród 104 przyjętych rezerwacji;

$S_{104} \sim \text{Bin}(n = 104, p = 0.9)$ przy założeniu, że zmienne X_i są niezależne.

Centralne twierdzenie graniczne Moivre'a-Laplace'a — przykład

$$
\begin{aligned}
P(S_{104} > 100) &= 1 - P(S_{104} \leq 100) \\
&= 1 - P\Big(\frac{S_{104} - pn}{\sqrt{np(1-p)}} \leq \frac{100 - np}{\sqrt{np(1-p)}}\Big) \\
&\cong 1 - \Phi\Big(\frac{100 - np}{\sqrt{np(1-p)}}\Big) \\
&= 1 - \Phi\Big(\frac{100 - 104 \cdot 0.9}{\sqrt{104 \cdot 0.9 \cdot 0.1}}\Big) \\
&= 1 - \Phi(2.091905) \\
&= 1 - 0.9817765 = 0.01822351.
\end{aligned}
$$

Materiały na podstawie

- Grzegorzewski P., Bobecka K., Dembińska A., Pusz J., Rachunek prawdopodobieństwa i statystyka, WSISiZ, Warszawa, wyd. V - 2008.
- Jacek Jakubowski, Rafał Sztencel, Rachunek prawdopodobieństwa dla prawie każdego, Script, Warszawa, 2006.