

Wykład 13. Para zmiennych losowych

dr Mariusz Grządział

27 maja 2014

Rozkład dyskretnej zmiennej losowej

Zbiór wartości dyskretnej zmiennej losowej X — ciąg $x_1, x_2, \dots$, (skończony lub nieskończony).

Rozkład zmiennej losowej dyskretnej X jest określony przez nieujemne liczby $p_1, p_2, \dots$ spełniające warunki:

$$\sum p_i = 1, \quad (1)$$

$$p_i = P(X = x_i). \quad (2)$$

Funkcja przyporządkowująca $x_1, x_2, \dots$ prawdopodobieństwa $p_1, p_2, \dots$: funkcja prawdopodobieństwa.

Para dyskretnych zmiennych losowych

Niech X i Y będą dyskretnymi zmiennymi losowymi określonymi na przestrzeni zdarzeń losowych $\mathcal{S}$. Funkcja $p_{X,Y}$ jest określona przez

$$p_{X,Y}(x, y) = P(X = x, Y = y)$$

dla każdej pary wartości (x, y) , która może być przyjmowana przez parę zmiennych losowych (X, Y) z niezerowym prawdopodobieństwem.

Funkcję $p_{X,Y}$ będziemy nazywać funkcją prawdopodobieństwa pary dyskretnych zmiennych losowych X i Y .

Para dyskretnych zmiennych losowych – c.d.

Funkcja $p_{X,Y}(x, y)$ spełnia warunki:

- $p_{X,Y}(x, y) \geq 0$ dla wszystkich x i y ;
- $\sum_x \sum_y p_{X,Y}(x, y) = 1$;
- $P((X, Y) \in A) = \sum_{(x,y) \in A} p_{X,Y}(x, y)$, gdzie A jest danym podzbiorem par uporządkowanych (x, y) (par wartości, które mogą być przyjmowane przez (X, Y) z dodatnim prawdopodobieństwem).

Przykład

Na uczelni W studenci są oceniani za pomocą ocen $P1, P2$ i $P3$; ocena $P1$ jest najniższą a $P3$ najwyższą z wymienionych ocen. Pod koniec 1-go roku studenci bioinformatyki zdają egzaminy z matematyki i chemii. Oceny uzyskane przez studentów bioinformatyki na uczelni W w roku 2010 można przedstawić za pomocą tabelki:

	P1 (M)	P2 (M)	P3 (M)
P1 (C)	0	13	4
P2 (C)	8	84	41
P3 (C)	7	33	10

$P1 (M)$ oznacza ocenę $P1$ z matematyki, $P2 (C)$ oceną $P2$ z chemii itd. Przykładowo: liczba studentów, która otrzymała ocenę $P2$ z chemii i $P1$ z matematyki wynosi 8. Prawdopodobieństwo, że wybrany losowo student otrzymał ocenę $P2$ z chemii i $P1$ z matematyki wynosi $8/200 = 0,04$.

Przykład — c.d.

Prawdopodobieństwa uzyskania par ocen: z chemii i matematyki przedstawia poniższa tabela:

	$Y_B = 1$	$Y_B = 2$	$Y_B = 3$
$X_B = 1$	0	0,065	0,02
$X_B = 2$	0,04	0,42	0,205
$X_B = 3$	0,035	0,165	0,05

Dystrybuanta łączna dyskretnych zmiennych losowych

Dystrybuanta łączna dyskretnych zmiennych losowych X i Y nazywamy funkcję F określoną wzorem

$$F(x, y) = P(X \leq x, Y \leq y) = \sum_{s \leq x} \sum_{t \leq y} p_{X,Y}(s, t),$$

gdzie $p_{X,Y}$ jest funkcją prawdopodobieństwa dla pary zmiennych losowych X i Y .

Przykład Oznaczmy dystrybuantę łączną pary zmiennych losowych X_B i Y_B przez G . Mamy na przykład:

$$G(2, 1) = 0,04 \quad G(2, 2) = 0 + 0,065 + 0,04 + 0,42 = 0,525.$$

Para zmiennych losowych typu ciągłego

Parę zmiennych losowych X i Y , określonych na tej samej przestrzeni zdarzeń elementarnych, będziemy nazywać parą zmiennych losowych typu ciągłego, jeśli istnieje funkcja gęstości g , określona na $\mathbb{R}^2$, która spełnia własności:

- $g(x, y) \geq 0$ dla $(x, y) \in \mathbb{R}^2$;

-

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) dx dy = 1;$$

- Dla zbioru $A \in \mathbb{R}^2$

$$P((X, Y) \in A) = \int \int_{(x,y) \in A} g(x, y) dx dy.$$

(zakładamy, że dla zbioru A odpowiednia całka istnieje).

Dystrybuanta łączna pary ciągłych zmiennych losowej typu ciągłego

Dla X i Y typu ciągłego dystrybuantę F definiujemy wzorem

$$F(x, y) = P(X \leq x, Y \leq y) = \int_{-\infty}^x \int_{-\infty}^y g(x, y) dx dy.$$

Para niezależnych zmiennych losowych

Przypomnijmy:

Definicja 1. Mówimy, że zmienne losowe X i Y są niezależne, jeżeli

$$P(X \in [a, b] \wedge Y \in [c, d]) = P(X \in [a, b]) \times P(Y \in [c, d])$$

dla dowolnych przedziałów $[a, b]$ i $[c, d]$.

Przyjęty warunek jest równoważny następującemu warunkowi:

$$P(X \in (-\infty, a] \wedge Y \in (-\infty, b]) = P(X \in (-\infty, a]) \times P(Y \in (-\infty, b])$$

który z kolei możemy zapisać w postaci:

$$F(x, y) = F_X(x)F_Y(y), \tag{3}$$

gdzie F_X oznacza dystrybuantę zmiennej losowej X , F_Y oznacza dystrybuantę zmiennej losowej Y a F oznacza dystrybuantę łączną pary zmiennych losowych X i Y .

Niezależność zmiennych losowych dyskretnych

Jeżeli X i Y są dyskretnie, warunek (3) jest równoważny

$$p_{X,Y}(x, y) = p_X(x)p_Y(y), \quad (4)$$

gdzie $p_{X,Y}(x, y)$ jest funkcją prawdopodobieństwa dla pary (X, Y) , p_X i p_Y są funkcjami prawdopodobieństwa dla (odpowiednio) zmiennych X i Y a x i y są dowolnymi liczbami takimi, że $p_X(x) > 0$ oraz $p_Y(y) > 0$.

Niezależność zmiennych losowych typu ciągłego

Jeżeli X i Y są typu ciągłego, warunek (3) jest równoważny

$$g_{X,Y}(x, y) = g_X(x)g_Y(y),$$

gdzie $g_{X,Y}(x, y)$ jest funkcją gęstości dla pary (X, Y) , p_X i p_Y są funkcjami gęstości dla (odpowiednio) zmiennych X i Y ; a x i y są dowolnymi liczbami rzeczywistymi.

Przykład

Rozważmy jeszcze raz doświadczenie losowe polegające na wykonaniu przez zawodnika A dwóch rzutów osobistych (prawdopodobieństwo trafienia jest równe 0,9). Niech Y_1 oznacza wynik pierwszego rzutu (0, jeśli A chybił w pierwszym rzucie, 1 jeśli A trafił w pierwszym rzucie) a Y_2 wynik drugiego rzutu (0, jeśli A chybił w drugim rzucie, 1 jeśli A trafił w drugim rzucie). Łączna gęstość Y_1 i Y_2 jest przedstawiona przy użyciu tabelki

	$Y_2 = 0$	$Y_2 = 1$
$Y_1 = 0$	0,01	0,09
$Y_1 = 1$	0,09	0,81

Widzimy, że Y_1 i Y_2 spełniają warunek (4).

Sumy niezależnych zmiennych losowych

Gdy $X_1, X_2, \dots, X_n$ są niezależnymi zmiennymi losowymi, to jest możliwe określenie rozkładu (dystrybucji):

- sumy $Y = X_1 + X_2 + \dots + X_n$;
- średniej $\bar{X} = \frac{Y}{n}$.

Na przykład jeśli $X_1, X_2, \dots, X_n$ są niezależnymi zmiennymi losowymi o rozkładzie $Bin(1, p)$ (zero-jedynkowym z parametrem p), to Y ma rozkład $Bin(n, p)$.

Definicja 2. Mówimy, że zmienna losowa X ma rozkład normalny z parametrami μ i σ , gdzie $\mu \in \mathbb{R}$ i $\sigma > 0$, jeżeli gęstość jej rozkładu jest określona wzorem:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Skrótowy zapis: X ma rozkład $N(\mu, \sigma)$.

Centralne Twierdzenie Graniczne

Twierdzenie 1. Jeżeli $X_1, X_2, \dots, X_n$ są niezależnymi zmiennymi losowymi o tym samym rozkładzie, $E(X_1) = m$, $D(X_1) = \sigma$, to zmienna losowa

$$\frac{\bar{X} - m}{\sigma/\sqrt{n}},$$

gdzie $\bar{X} = \frac{1}{n}(X_1 + X_2 + \dots + X_n)$, ma w przybliżeniu standardowy rozkład normalny $N(0, 1)$; dokładniej:

$$P\left(a \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq b\right) \rightarrow \Phi(b) - \Phi(a)$$

dla dowolnych a i b , $a < b$.

Innymi słowy, rozkład $\bar{X}$ jest w przybliżeniu równy rozkładowi normalnemu $N(\mu, \sigma/\sqrt{n})$.

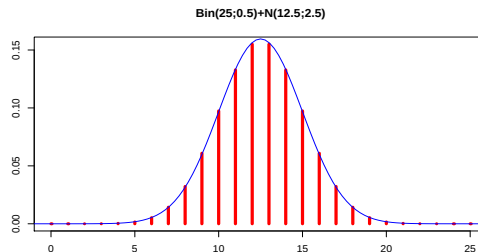
Przykład

Rzucamy monetą 25 razy. Niech Y oznacza liczbę orłów. X ma rozkład $Bin(25; 0,5)$. Zmienna Y może być przedstawiona jako suma 25 niezależnych zmiennych losowych $X_1, \dots, X_{25}$ o rozkładzie $Bin(1; 0,5)$ (zero-jedynkowym z parametrem $p = 0,5$). Korzystając z Centralnego Twierdzenia Granicznego: $\bar{X}$ ma w przybliżeniu rozkład $N(p, \sigma/\sqrt{n})$, gdzie $n = 25, p = 0,5, \sigma = \sqrt{p(1-p)} = 0,5$, czyli

$\bar{X}$ ma w przybliżeniu rozkład $N(0,5; 0,5/5)$

więc

Y ma w przybliżeniu rozkład $N(12,5; 2,5)$



Rysunek 1: Rozkład dwumianowy $Bin(25; 0,5)$ i odpowiadający mu rozkład normalny $N(12,5; 2,5)$

Suma niezależnych zmiennych losowych o rozkładzie normalnym

Twierdzenie 2. Jeśli $X_1, \dots, X_n$ są niezależnymi zmiennymi losowymi,

$$X_k \sim N(\mu_k, \sigma_k), \quad k = 1, 2, \dots, n$$

to zmienna losowa $Y = X_1 + \dots + X_n$ ma rozkład $N(\mu_s, \sigma_s)$, gdzie

$$\mu_s = \sum_{k=1}^n \mu_k \quad i \quad \sigma_s = \sqrt{\sum_{k=1}^n \sigma_k^2}.$$

Wniosek 1. Jeśli $X_1, \dots, X_n$ są niezależnymi zmiennymi losowymi o tym samym rozkładzie $N(\mu, \sigma)$, to

- średnia

$$\bar{X} = \frac{X_1 + \dots + X_n}{n}$$

ma rozkład $N(\mu, \sigma/\sqrt{n})$;

- zmienna losowa Z określona równością

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

ma rozkład $N(0, 1)$ (standardowy normalny).

Funkcja gamma

Funkcja Γ jest określona wzorem

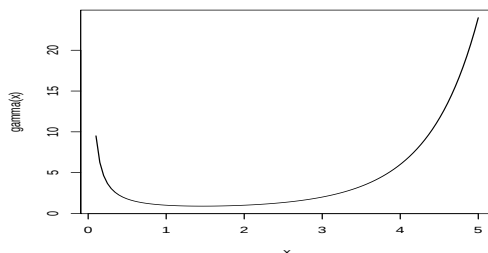
$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx, \quad t > 0;$$

Można uzasadnić, że:

1. $\Gamma(t) > 0$ dla $t > 0$;
2. $\Gamma(1) = 1$;

3. $\Gamma(t+1) = t\Gamma(t)$ dla $t > 0$;
4. $\Gamma(1/2) = \sqrt{\pi}$;
5. $\lim_{t \rightarrow 0+} \Gamma(t) = \infty$;
6. $\lim_{t \rightarrow \infty} \Gamma(t) = \infty$;

Funkcja gamma



Rysunek 2: Wykres funkcji Γ na odcinku $[0, 5]$.

Rozkład t-Studenta

Gęstość rozkładu t-Studenta z n stopniami swobody, $n \in \mathbb{N}$, określona jest wzorem

$$t_n(x) = \frac{\Gamma((n+1)/2)}{\Gamma(n/2)\sqrt{\pi n}} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}.$$

Można uzasadnić, że funkcja t_n jest

1. dodatnia na $\mathbb{R}$;
2. parzysta;

oraz że $t_n(x) \rightarrow \phi(x)$ gdy $n \rightarrow \infty$.

Rozkład t-Studenta — c.d.

Jeżeli $X_1, X_2, \dots, X_n$ są niezależnymi zmiennymi losowymi o rozkładzie $N(\mu, \sigma)$, to zmienna losowa T_n określona równością

$$T_n = \frac{\bar{X} - \mu}{S/\sqrt{n}},$$

gdzie $\bar{X} = \frac{1}{n}(X_1 + X_2 + \dots + X_n)$ oraz

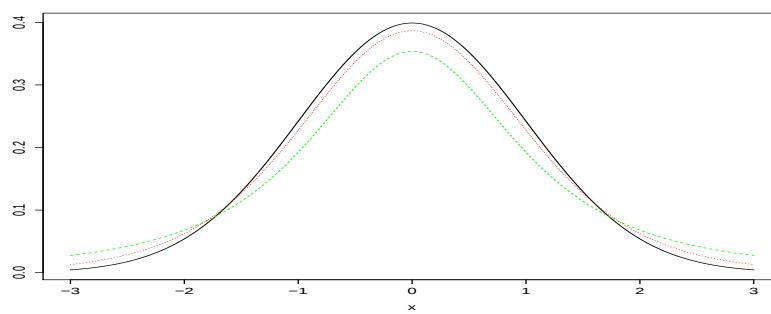
$$S = \left[\frac{(X_1 - \bar{X})^2 + \dots + (X_n - \bar{X})^2}{n-1} \right]^{1/2},$$

ma rozkład t-Studenta z $n-1$ stopniami swobody.

Rozkład t-Studenta i rozkład normalny

Lektura uzupełniająca

Koronacki, J., Mielniczuk, J. Statystyka dla studentów kierunków technicznych i przyrodniczych. WNT, Warszawa 2001, strony 122–126 i 138–142.



Rysunek 3: Gęstości rozkładów gamma: $N(0; 1)$ (linia ciągła czarna), t-Studenta z dwoma stopniami swobody („kreski” zielone), t-Studenta z ośmioma stopniami swobody („kropki” czerwone).