Zmienne losowe – transformacje, charakterystyki liczbowe, twierdzenia graniczne

dr Mariusz Grządziel

Katedra Matematyki, Uniwersytet Przyrodniczy we Wrocławiu

r. akad. 2017/2018; semestr letni

Charakterystyki liczbowe zmiennych losowych o skończonym zbiorze wartości

Jeżeli zmienna losowa X przybiera skończoną liczbę wartości $x_1, x_2, \dots, x_n$ z prawdopodobieństwami $p_1, p_2, \dots, p_n$, to jej wartością oczekiwaną nazywamy wielkość

$$EX = \sum_{k=1}^{n} x_k p_k.$$

Interpretacja: środek ciężkości układu punktów na prostej o współrzędnych $x_1, x_2, \dots, x_n$ z prawdopodobieństwami $p_1, p_2, \dots, p_n$.

Jeżeli $Y = g(X)$, gdzie g jest funkcją ciągłą, to

$$E(Y) = \sum_{k=1}^{n} g(x_k) p_k.$$

Własności wartości oczekiwanej

Łatwo widać, że wartość oczekiwana zmiennej losowej $Y = aX + b$ jest równa

$$E(aX + b) = aEX + b;$$

zakładamy, że EX istnieje.
Jeśli wartości oczekiwane zmiennych losowych X_1 i X_2 istnieją i są równe, odpowiednio, μ_1 i μ_2, to

$$E(X_1 + X_2) = \mu_1 + \mu_2.$$

Wariancja zmiennej losowej I

Definicja 1

Wariancję zmiennej losowej X określamy wzorem

$$VarX = E(X - \mu)^2,$$

gdzie $\mu = EX$.

Wariancja jest równa wartości oczekiwanej kwadratu odchylenia wartości zmiennej losowej od swojej wartości przeciętnej.

Stwierdzenie 1

Dla $a > 0$ mamy:

$$Var(aX + b) = a^2 Var(X).$$

Stwierdzenie 2

Gdy zmienna losowa X jest dyskretna, wariancja może być wyrażona wzorem

$$VarX = \sum_i (x_i - \mu)^2 p_i,$$

a gdy jest typu ciągłego:

$$Var(X) = \int_{-\infty}^{\infty} (x - \mu)^2 g(x) dx.$$

Wariancja rozkładu zmiennej losowej — interpretacja mechaniczna

Widzimy, że wariancja zmiennej losowej jest tym mniejsza, im bardziej masa jej rozkładu jest kupiona wokół środka ciężkości μ- wartości oczekiwanej.
Wariancja zmiennej losowej jest równa zeru wtedy i tylko wtedy, gdy cała masa jej rozkładu jest skupiona w jednym punkcie (tj. gdy jest zmienną losową dyskretną, która przybiera jedną wartość z prawdopodobieństwem 1).

Odchylenie standardowe

Odchylenie standardowe zmiennej losowej X, oznaczane przez DX, definiujemy jako pierwiastek kwadratowy wariancji X.
Odchylenie standardowe zmiennej losowej X często jest też oznaczane grecką literą σ.
Można pokazać, że dla $a \geqslant 0$:

$$D(aX + b) = aDX. \tag{1}$$

Przykład

Zmienna losowa dyskretna W ma rozkład dany tabelką:

x_i	0	2	5
p_i	0,3	0,5	0,2

$$
\begin{aligned}
E(W) &= (0) \times 0{,}3 + 2 \times 0{,}5 + 5 \times 0{,}2 = 2, \\
Var(W) &= [0-2]^2 \times 0{,}3 + [2-2]^2 \times 0{,}5 + [5-2]^2 \times 0{,}2 = 3, \\
DW &= \sqrt{3}.
\end{aligned}
$$

Wartość oczekiwana i wariancja zmiennej o rozkładzie dwumianowym

Niech X będzie zmienną losową o rozkładzie dwumianowym z parametrami $n \in \mathbb{N}$ i $p \in (0, 1)$. Można pokazać, że:

$$E(X) = np, \quad Var(X) = np(1 - p).$$

Wartość oczekiwana dla zmiennych losowych typu ciągłego

Wartość oczekiwana zmiennej losowej typu ciągłego X określona jest wzorem

$$E(X) = \int_{-\infty}^{\infty} xg(x)dx, \tag{2}$$

przy założeniu, że całka (2) jest bezwzględnie zbieżna, to jest

$$\int_{-\infty}^{\infty} |x|g(x)dx = h,$$

gdzie h jest pewną liczbą rzeczywistą.
Jeżeli całka (2) nie jest bezwzględnie zbieżna, wartość oczekiwana zmiennej losowej X nie istnieje.

Wartość oczekiwana i wariancja zmiennej losowej o rozkładzie jednostajnym na odcinku $[0, 1]$

Jeżeli zmienna losowa X ma rozkład jednostajny na odcinku $[0, 1]$ (czyli $U(0, 1)$), to

$$E(X) = \frac{1}{2}; \quad Var(X) = \frac{1}{12}.$$

Przypomnijmy, że gęstość u zmiennej losowej o rozkładzie $U(0, 1)$ dana jest wzorem

$$u(x) = \begin{cases} 1, & \text{jeśli } 0 \leqslant x \leqslant 1, \\ 0 & \text{jeśli } x < 0 \text{ lub } x > 1. \end{cases}$$

Wartość oczekiwana i wariancja zmiennej losowej o rozkładzie normalnym

Można pokazać, że jeśli $X \sim N(\mu, \sigma)$, to

$$E(X) = \mu, \quad Var(X) = \sigma^2, \quad D(X) = \sigma.$$

Wymaga to obliczenia całek trochę bardziej skomplikowanych niż dla przypadku odpowiadającego $U(0, 1)$.

Niezależność zmiennych losowych

Definicja 2

Mówimy, że zmienne losowe X i Y, określone na pewnej przestrzeni zdarzeń elementarnych $\mathcal{S}$, są niezależne, jeżeli

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

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

Intuicyjnie: niezależne zmienne losowe odpowiadają wynikom liczbowym „niezależnych” eksperymentów losowych.
Definicję pojęcie niezależności układu zmiennych losowych $X_1, X_2, \ldots, X_n$ w przypadku, gdy $n > 2$, można znaleźć w książce [GK00] na str. 56.

Wariancja niezależnych zmiennych losowych

Twierdzenie 1

Jeżeli zmienne losowe X i Y są niezależne i mają wariancje, to dla dowolnych liczb rzeczywistych a i b prawdziwa jest równość

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y).$$

Uwaga 1

Prawdziwe jest również uogólnienie tego twierdzenia na przypadek n niezależnych zmiennych losowych: Jeżeli zmienne losowe $X_1, X_2, \ldots, X_n$ są niezależne i mają wariancje, to dla dowolnych liczb rzeczywistych $a_1, \ldots, a_n$

$$Var(a_1 X_1 + \ldots + a_n X_n) = a_1^2 Var(X_1) + \ldots + a_n^2 Var(X_n).$$

Standaryzacja

Uwaga 2

Standaryzacja zmiennej losowej — operacja polegająca na przyporządkowaniu zmiennej losowej X zmiennej

$$\frac{X - \mu_X}{\sigma_X},$$

gdzie $\mu_X = E(X)$ i $\sigma_X = DX$.

Standaryzowane sumy i średnie zmiennych losowych

Stwierdzenie 3

Niech $X_1, X_2, \ldots, X_n$ będą niezależnymi zmiennymi losowymi o jednakowym rozkładzie o średniej μ i wariancji σ^2. Niech $\bar{X}$ oznacza średnią $\frac{1}{n}(X_1 + \ldots + X_n)$.
Wówczas

$$E(\bar{X}) = \mu \quad i \quad D(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

Stwierdzenie 4

Standaryzowana średnia dla zmiennych $X_1, X_2, \ldots, X_n$ ze Stwierdzenia 3 jest równa standaryzowanej sumie tych zmiennych losowych:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}}.$$

Centralne Twierdzenie Graniczne

Twierdzenie 2

Jeżeli $X_1, X_2, \ldots, X_n$ są niezależnymi zmiennymi losowymi o tym samym rozkładzie, $E(X_1) = \mu$, $D(X_1) = \sigma$, to zmienna losowa

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

gdzie $\bar{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$, ma w przybliżeniu standardowy rozkład normalny $N(0, 1)$, tj.

$$P\Big(a \leqslant \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leqslant b\Big) \to \Phi(b) - \Phi(a)$$

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

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

Niech $X_1, \ldots, X_n$ będą niezależnymi zmiennymi losowymi o rozkładzie dwumianowym z parametrami $n \in \mathbb{N}$ i $p \in (0, 1)$. Z Centralnego Twierdzenia Granicznego wynika, że $\bar{X}$ ma w przybliżeniu rozkład

$$N\Big(p, \sqrt{\frac{p(1-p)}{n}}\Big)$$

a $X_1 + \ldots + X_n$ ma w przybliżeniu rozkład

$$N\Big(np, \sqrt{np(1-p)}\Big).$$

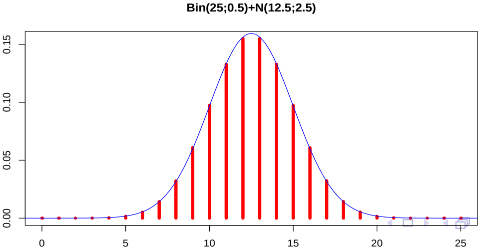
Przykład

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 $X_1, \ldots, X_{25}$ o rozkładzie $Bin(1; 0{,}5)$. Korzystając z Centralnego Twierdzenia Granicznego (Twierdzenia 2): $\bar{X} \sim N(p, \sigma/\sqrt{n})$, gdzie $n = 25, p = 0{,}5, \sigma = \sqrt{p(1-p)} = 0{,}5$, czyli

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

więc

$$Y \text{ ma w przybliżeniu rozkład } N(12{,}5; 2{,}5)$$



Ciągi niezależnych zmiennych losowych o niekoniecznie jednakowych rozkładach

Istnieją odpowiedniki Centralnego Twierdzenia Granicznego (Twierdzenia 2) dla przypadku ciągu niezależnych zmiennych losowych o niekoniecznie jednakowych rozkładach. W twierdzeniach tych warunek identyczności rozkładów zastępowany jest innymi (słabszymi) warunkami.

Ciągi niezależnych zmiennych losowych o niekoniecznie jednakowych rozkładach

Istnieją odpowiedniki Centralnego Twierdzenia Granicznego (Twierdzenia 2) dla przypadku ciągu niezależnych zmiennych losowych o niekoniecznie jednakowych rozkładach. W twierdzeniach tych warunek identyczności rozkładów zastępowany jest innymi (słabszymi) warunkami.

Rozkład normalny – zastosowania

Najwcześniejsze zastosowania: przybliżenie rozkładu dwumianowego rozkładem normalnym (tw. de Moivre'a – Laplace'a — lata 30-te XVIII-go wieku).

Centralne Twierdzenie Graniczne (CTG)— wersja klasyczna oraz wersje dla zmiennych losowych o niekoniecznie jednakowych rozkładach: pierwsze 4 dekady XX wieku.

Przykład

Uzasadnienie dla normalności zmiennej losowej opisującej sumę wydatków 20 kolejnych klientów w sklepie — wersja klasyczna CTG (Twierdzenie 2).

Twierdzenia graniczne dla zmiennych losowych o niekoniecznie identycznych rozkładach

Twierdzenie 3

Niech $X_1, X_2, \ldots, X_n, \ldots$ będzie ciągiem niezależnych zmiennych losowych i niech $S_n = X_1 + X_2 + \ldots + X_2$, $n \in \mathbb{N}$. Oznaczmy

$$E(X_k) = \mu_k, \quad Var(X_k) = \sigma_k^2, \quad k = 1, 2, \ldots$$
$$E(S_n) = m_n, \quad Var(S_n) = s_n^2.$$

Załóżmy, że $s_n^2 \to \infty$. Jeżeli istnieje stała A taka, że $|X_n| < A$ dla każdego n, wtedy dla $a < b$

$$P\left(a \leqslant \frac{S_n - m_n}{s_n} \leqslant b\right) \to \Phi(b) - \Phi(a).$$

Twierdzenie to jest wnioskiem z twierdzenia Lindeberga–Fellera (bardziej ogólnego, ale i „trudniejszego w zapisie"), por. [CS06, str. 344–345] i [Krz97, str. 242].

Przykład — rozrzut pocisków artyleryjskich

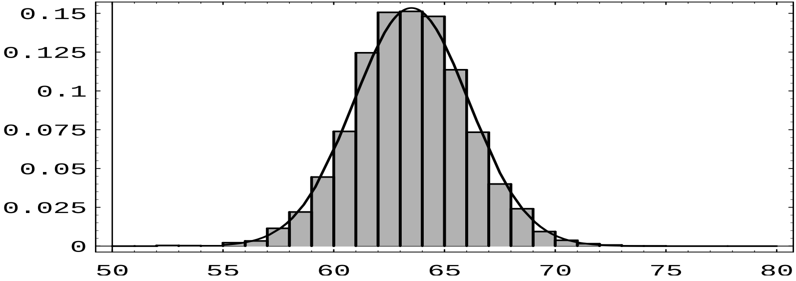
A. Borowkow, Rachunek prawdopodobieństwa, PWN 1977, str. 113–114:

> *(...) na trajektorię lotu pocisku wpływa duża liczba niezależnych czynników, przy czym wpływ każdego z nich oddzielnie wzięty jest niewielki. Są to odchylenia w ilości ładunku wybuchowego, odchylenia w wadze i rozmiarach pocisku, odchylenia w wilgotności i temperaturze powietrza, kierunek i siła wiatru na różnych wysokościach itd. W rezultacie odchylenie pocisku zadziwiająco dobrze opisuje rozkład normalny.*

Uzasadnieniem powyższego faktu mogą być centralne twierdzenie graniczne dla dowolnych ciągów zmiennych losowych.

Przykład

Histogram probabilistyczny dla danych dotyczących wzrostu 9593 kobiet w wieku od 21 do 74 lat. Dane pochodzą z badań wykonanych w ramach projektu *The Health and Nutrition Examination Survey I (HANES I)*. Do badań wybrano losową próbę z populacji obywateli Stanów Zjednoczonych. Badania były wykonane w latach pomiędzy 1971 i 1974. Wykres pochodzi z książki L. Snella i Ch. Grinsteada *Grinstead and Snell's Introduction to Probability* dostępnej na stronie
`www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/book.html`
Książka została udostępniona pod `https://www.gnu.org/copyleft/fdl.html`



Prace cytowane

[GK00] Gajek, L. i Kałuszka M., Wnioskowanie statystyczne. Modele i metody. WNT 2000.

[GS06] Grinstead Ch. i Snell, L., *Grinstead and Snell's Introduction to Probability*, The Chance Project 2006, książka dostępne na stronie `www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/book.html`

[Krz97] Krzyśko, M., Wykłady z teorii prawdopodobieństwa. Rektor Uniwersytetu im. A. Mickiewicza, Poznań 1997.