Rozkłady dyskretne

Joanna Czarnowska[1]

[1]Uniwersytet Gdański
Instytut Informatyki
jczarn@inf.ug.edu.pl

Modelowanie wielkości o charakterze losowym

Do opisu wielkości losowych wykorzystujemy zmienne losowe. Badane wielkości mogą być:

- przeliczalne – liczba wypadków, liczba wypłat z portfela polis, liczba awarii na taśmie, itd.
- nieprzeliczalne – czas świecenia żarówki, temperatura, prędkość wiatru, poziom cukru we krwi, ciśnienie, ceny akcji na giełdzie, itd.

Pierwsze wielkości opisują zmienne losowe dyskretne, drugie – zmienne losowe ciągłe. Zmienne losowe oznacza się zwykle dużymi literami X, Y czy Z (w przeszłości częściej literami ξ, η, ζ).

Precyzyjna definicja zmiennej losowej – 01RozkladyDyskretne.pdf

Rozkład zmiennej losowej dyskretnej

Podając wartości jakie przyjmuje zmienna losowa dyskretna

$$X \in \{x_1, x_2, x_3, \ldots\}$$

oraz prawdopodobieństwa ich wystąpienia

$$p_1 = P(X = x_1), \; p_2 = P(X = x_2), \;\; p_3 = P(X = x_3), \ldots$$

określamy rozkład prawdopodobieństwa (krótko rozkład) tej zmiennej. Przy czym zachodzi

$$p_1 + p_2 + p_3 + \ldots = 1, \quad p_i \geqslant 0.$$

Zmienna losowa dyskretna – przykład

Przykład

Rzucamy raz kostką symetryczną. W zależności od wypadniętej liczby oczek wygrywamy lub przegrywamy pewną kwotę. Opisuje to zmienna losowa X, w tabeli mamy jej rozkład.

$$X(\omega) = \begin{cases} -40 & \textit{jeśli wypadnie } 1, 2, 3 \\ 30 & \textit{jeśli wypadnie } 4, 5 \\ 120 & \textit{jeśli wypadnie } 6 \end{cases}$$

x_i	-40	30	120
p_i	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$

Dla każdego rozkładu definiujemy parametry, które niosą o nim pewne informacje – są to np. wartość oczekiwana, wariancja, odchylenie standardowe czy skośność.

- Wartość oczekiwana

$$m = E(X) = x_1 p_1 + x_2 p_2 + \cdots + x_n p_n + \cdots = \sum_n x_n p_n$$

- Wariancja

$$\mathrm{Var}(X) = E(X-m)^2 = (x_1-m)^2 p_1 + (x_2-m)^2 p_2 + \ldots + (x_n-m)^2 p_n \ldots =$$

$$= \sum_n (x_n - m)^2 p_n$$

- Odchylenie standardowe: $\sigma = \sqrt{\mathrm{Var}(X)}$

Definiujemy też parametry

- k-ty moment zwykły: $E(X^k) = \sum_n x_n^k p_n$
- k-ty moment centralny: $E(X-m)^k = \sum_n (x_n - m)^k p_n$.

Wariancja jest drugim momentem centralnym.

Parametry rozkładu – przykład

Przykład cd.

Wróćmy do rozkładu zmiennej losowej X opisującej wygraną

x_i	-40	30	120
p_i	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$

Wartość oczekiwana

$$E(X) = -40 \cdot \frac{1}{2} + 30 \cdot \frac{1}{3} + 120 \cdot \frac{1}{6} = 10$$

Wariancja

$$Var(X) = (-40 - 10)^2 \cdot \frac{1}{2} + (30 - 10)^2 \cdot \frac{1}{3} + (120 - 10)^2 \cdot \frac{1}{6} = 3400$$

Odchylenie standardowe

$$\sigma = \sqrt{3400} \approx 58$$

Własności wartości oczekiwanej i wariancji

Dla dowolnych liczb $a, b \in \mathbb{R}$ i dowolnych zmiennych losowych X i Y, zachodzi

- $E(aX + b) = aEX + b, \quad E(X + Y) = EX + EY$
- $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X), \quad \mathrm{Var}(X) = E(X^2) - E^2(X).$

Przykład cd.

Dla zmiennej losowej X opisującej wygraną

x_i	-40	30	120
p_i	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$

mamy

$$E(X^2) = (-40)^2 \cdot \frac{1}{2} + 30^2 \cdot \frac{1}{3} + 120^2 \cdot \frac{1}{6} = 3500.$$

Skąd wariancja jako różnica drugiego momenu i kwadratu wartości oczekiwanej wynosi

$$Var(X) = E(X^2) - E^2X = 3500 - 10^2 = 3400.$$

Ważne rozkłady dyskretne

- rozkład jednopunktowy – zmienna losowa X przyjmuje jedną wartość a z prawdopodobieństwem $P(X = a) = 1$.

  Momenty: $EX = a$ i $\text{Var}X = 0$

Ważne rozkłady dyskretne

- rozkład jednopunktowy – zmienna losowa X przyjmuje jedną wartość a z prawdopodobieństwem $P(X = a) = 1$.

  Momenty: $EX = a$ i $\mathrm{Var}X = 0$

- rozkład dwupunktowy – zmienna losowa X przyjmuje dwie różne wartości a i b, z prawdopodobieństwami

  $$P(X = a) = p \qquad i \qquad P(X = b) = q,$$

  gdzie $p, q > 0$ i $p + q = 1$. Szczególnym przypadkiem tego rozkładu jest rozkład zero-jedynkowy, gdzie $a = 1$ i $b = 0$.

  Momenty: $EX = pa + qb$ i $\mathrm{Var}X = pq(a - b)^2$

Ważne rozkłady dyskretne

- rozkład jednopunktowy – zmienna losowa X przyjmuje jedną wartość a z prawdopodobieństwem $P(X = a) = 1$.

  Momenty: $EX = a$ i $\mathrm{Var}X = 0$
- rozkład dwupunktowy – zmienna losowa X przyjmuje dwie różne wartości a i b, z prawdopodobieństwami

  $$P(X = a) = p \qquad i \qquad P(X = b) = q,$$

  gdzie $p, q > 0$ i $p + q = 1$. Szczególnym przypadkiem tego rozkładu jest rozkład zero-jedynkowy, gdzie $a = 1$ i $b = 0$.

  Momenty: $EX = pa + qb$ i $\mathrm{Var}X = pq(a - b)^2$
- rozkład Bernoulliego (dwumianowy) $B(n, p)$ – zmienna losowa X przyjmuje wartośsci $0, 1, 2, \ldots, n$ z prawdopodobieństwem

  $$P(X = k) = \binom{n}{k} p^k q^{n-k},$$

  gdzie $p \in (0, 1)$ i $p + q = 1$. Jest to rozkład liczby sukcesów w n doświadczeniach Bernoulli'ego, gdzie szansa sukcesu w jednym doświadczeniu wynosi p.

  Momenty: $EX = np$ i $\mathrm{Var}X = npq$

Ważne rozkłady dyskretne

- rozkład Poissona $P(\lambda)$ – zmienna losowa X przyjmuje wartości $0, 1, 2, 3, \ldots,$ z prawdopodobieństwem

  $$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda},$$

  gdzie $\lambda > 0$ jest parametrem.

  Momenty: $EX = \lambda$ i $\mathrm{Var}X = \lambda$

  Uwaga. Czasem rozkład Poissona nazywa się „rozkładem zdarzeń rzadkich" albo „prawem małych liczb"'. Mogą to być pożary, wypadki czy też główne nagrody w grach losowych. Występowanie rozkładu Poissona badał Bortkiewicz w pracy z roku 1895. Była tam mowa o zgonach na skutek kopnięcia przez konia w armii pruskiej. (*Wstęp do teorii prawdopodobieństwa* - Jacek Jakubowski, Rafał Sztencel.)

Zobacz Przykład 1 i Przykład 2 w skrypcie 01RozkladyDyskretne-wyklad.R

Rozkład Bernoullie'go a rozkład Poissona

Rozkład Bernoullie'go przy spełnieniu pewnych warunków można aproksymować rozkładem Poissona.

Twierdzenie Poissona
Niech (X_n) będzie ciągiem zmiennych losowych o rozkładzie dwumianowym $B(n, p_n)$. Jeżeli

$$\lim_{n\to\infty} n \cdot p_n = \lambda > 0,$$

to

$$\lim_{n\to\infty} P(X_n = k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

dla $k = 0, 1, 2, \ldots,$.

Rozkład Bernoullie'go a rozkład Poissona

Przykład. Urna zawiera 1 kulę białą i 49 kul czarnych. Losujemy z niej 50 razy po jednej kuli zwracając zawsze wylosowaną kulę z powrotem do urny. Jakie jest prawdopodobieństwo wylosowania co najmniej dwa razy kuli białej?

Mamy tutaj do czynienia ze schematem Bernoullie'go o prawdopodobieństwie sukcesu (wylosowania kuli białej) w jednym doświadczeniu $p = \frac{1}{50}$. Niech X będzie liczbą sukcesów w pięćdziesięciu doświadczeniach. Prawdopodobieństwo, że bedziemy mieli ich dokładnie k, wynosi

$$P(X = k) = \binom{50}{k} \left(\frac{1}{50}\right)^k \left(\frac{49}{50}\right)^{50-k}.$$

Szukane prawdopodobieństwo wynosi zatem

$$P(X \geqslant 2) = 1 - P(X < 2) = 1 - P(X = 0) - P(X = 1)$$

$$= 1 - \binom{50}{0} \left(\frac{1}{50}\right)^0 \left(\frac{49}{50}\right)^{50} - \binom{50}{1} \left(\frac{1}{50}\right) \left(\frac{49}{50}\right)^{49} = 0{,}2642286.$$

Rozkład Bernoullie'go a rozkład Poissona

Przykład. Urna zawiera 1 kulę białą i 49 kul czarnych. Losujemy z niej 50 razy po jednej kuli zwracając zawsze wylosowaną kulę z powrotem do urny. Jakie jest prawdopodobieństwo wylosowania co najmniej dwa razy kuli białej?

Mamy tutaj do czynienia ze schematem Bernoullie'go o prawdopodobieństwie sukcesu (wylosowania kuli białej) w jednym doświadczeniu $p = \frac{1}{50}$. Niech X będzie liczbą sukcesów w pięćdziesięciu doświadczeniach. Prawdopodobieństwo, że bedziemy mieli ich dokładnie k, wynosi

$$P(X = k) = \binom{50}{k} \left(\frac{1}{50}\right)^k \left(\frac{49}{50}\right)^{50-k}.$$

Szukane prawdopodobieństwo wynosi zatem

$$P(X \geqslant 2) = 1 - P(X < 2) = 1 - P(X = 0) - P(X = 1)$$

$$= 1 - \binom{50}{0} \left(\frac{1}{50}\right)^0 \left(\frac{49}{50}\right)^{50} - \binom{50}{1} \left(\frac{1}{50}\right) \left(\frac{49}{50}\right)^{49} = 0{,}2642286.$$

Korzystając z twierdzenia Poissona mamy ($\lambda = np = 1$)

$$P(X \geqslant 2) = 1 - P(X = 0) - P(X = 1)$$

$$\approx 1 - e^{-1}\frac{1^0}{1!} - e^{-1}\frac{1^1}{1!} = 0{,}2642411.$$

Rozkład Bernoullie'go a rozkład Poissona

Przykład. Prawdopodobieństwo p trafienia „szóstki" w Toto-Lotku jest równe $1/\binom{49}{6} = 1/13983816 \approx 7 \cdot 10^{-8}$. Ilu „szóstek" można się spodziewać w każdym tygodniu, jeśli grający wypełniają kupony całkowicie losowo i niezależnie od siebie, a kuponów jest $n = 10^7$?

Zgodnie z twierdzeniem Poissona liczba „szóstek" ma rozkład zbliżony do rozkładu Poissona z parametrem $\lambda = np = 0,7151$. Szanse pojawienia się 0, 1 i 2 „szóstek" wynoszą zatem

$$P(X = 0) \approx \frac{\lambda^0}{0!} \cdot e^{-\lambda} = 0{,}4891, \qquad P(X = 1) \approx \frac{\lambda^1}{1!} \cdot e^{-\lambda} = 0{,}3498,$$

$$P(X = 2) \approx \frac{\lambda^2}{2!} \cdot e^{-\lambda} = 0{,}1251.$$

Zobacz Przykład 3 w skrypcie 01RozkladyDyskretne.R

Skośność i kurtoza

W analizie rozkładów korzysta się też ze skośności i kurtozy. Po transformacji liniowej zmiennej X

$$Y = \frac{X - m}{\sigma},$$

gdzie $m = EX$, $\sigma^2 = \mathrm{Var}(X)$, otrzymujemy zmienną standaryzowaną czyli taką, dla której

$$EY = 0, \quad \mathrm{Var}(Y) = 1.$$

Skośność i kurtoza

W analizie rozkładów korzysta się też ze skośności i kurtozy. Po transformacji liniowej zmiennej X

$$Y = \frac{X - m}{\sigma},$$

gdzie $m = EX$, $\sigma^2 = \mathrm{Var}(X)$, otrzymujemy zmienną standaryzowaną czyli taką, dla której

$$EY = 0, \quad \mathrm{Var}(Y) = 1.$$

Skośność i kurtozę zmiennej X definiujemy jako odpowiednio trzeci i czwarty moment tej zmiennej po standaryzacji

$$sk = EY^3 = \frac{E(X - m)^3}{\sigma^3}, \qquad \kappa = EY^4 = \frac{E(X - m)^4}{\sigma^4}.$$

Jeśli skośność jest dodatnia to mówimy, że rozkład jest prawostronnie skośny, jeśli ujemna – lewostronnie skośny. Gdy równa zeru rozkład jest symetryczny. (Ładne wykresy, zobacz: https://en.wikipedia.org/wiki/Skewness)
Kurtoza informuje o spłaszczeniu rozkładu. Więcej na ten temat przy rozkładach ciągłych.

Zobacz Przykład 4 w skrypcie 01RozkladyDyskretne.R

Dystrybuanta

Rozkład zmiennej losowej opisujemy też dystrybuantą.

Definicja

Dystrybuantą zmiennej losowej X nazywamy funkcję $F : \mathbb{R} \to \mathbb{R}$, która każdemu elementowi $x \in \mathbb{R}$ przyporządkowuje prawdopodobieństwo

$$F(x) = P(X < x).$$

Przykład

Dystrybuanta zmiennej losowej X opisującej wygraną dana jest wzorem

$$F(x) = \begin{cases} 0 & dla\ x \leqslant -40 \\ \frac{1}{2} & dla\ x \in (-40, 30\rangle \\ \frac{5}{6} & dla\ x \in (30, 120\rangle \\ 1 & dla x > 120. \end{cases}$$

x_i	-40	30	120
p_i	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$

Dystrybuanta – przykład

Przykład

Dystrybuanta zmiennej losowej X dana jest wzorem

$$F(x) = \begin{cases} 0 & dla\ x \leqslant 1 \\ \frac{1}{8} & dla\ x \in (1,3\rangle \\ \frac{3}{8} & dla\ x \in (3,5\rangle \\ \frac{7}{8} & dla\ x \in (5,7\rangle \\ 1 & dla\ x > 7. \end{cases}$$

Podaj, jakie wartości przyjmuje zmienna losowa X z niezerowym prawdopodobieństwem.

Wartości z odpowiednimi prawdopodobieństwami podane są w tabeli.

x_i	1	3	5	7
p_i	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{8}$

Dystrybuanta – własności

Twierdzenie

Dystrybuanta spełnia następujące własności

- $\lim_{x \to -\infty} F(x) = 0, \ \lim_{x \to \infty} F(x) = 1,$
- *jest funkcją niemalejącą,*
- *jest funkcją lewostronnie ciągłą.*

Uwaga. W niektórych zastosowaniach wygodniej jest zdefiniować dystrybuantę jako funkcję $F(x) = P(X \leqslant x)$.