Statystyka i Analiza Danych

W2: Zmienna losowa

dr hab. inż. Katarzyna Filipiak, prof. PP

Instytut Matematyki
Politechnika Poznańska

2023/2024

Zmienna losowa

Zmienną losową X nazywamy funkcję $X = X(\omega)$ określoną na przestrzeni zdarzeń elementarnych Ω o wartościach w zbiorze liczb rzeczywistych $\mathbb{R}$

Niech zmienna losowa dyskretna (skokowa) X przyjmuje wartości $x_1, x_2, \ldots$ odpowiednio z prawdopodobieństwami $p_1, p_2, \ldots;$ $\sum_{i=1}^{\infty} p_i = 1$.

Zmienna losowa ciągła – zmienna przyjmująca wszystkie wartości z pewnego przedziału liczbowego.

Zmienna losowa dyskretna

Rozkład prawdopodobieństwa

Rozkładem prawdopodobieństwa zmiennej losowej dyskretnej X, nazywamy funkcję przyporządkowującą wartościom zmiennej x_i $(i = 1, 2, \ldots)$ prawdopodobieństwa ich przyjęcia:

$$P(X = x_i) = P(\{\omega : X(\omega) = x_i\}) = p_i.$$

Przykład

W pewnym eksperymencie wykorzystano trzy automatyczne aparaty fotograficzne w celu dokumentowania jego przebiegu. W danych warunkach prawdopodobieństwo wykonania poprawnej fotografii dla każdego aparatu jest takie samo i wynosi $p = 0,6$. Oblicz prawdopodobieństwo:
(a) nieudokumentowania eksperymentu;
(b) zarejestrowania eksperymentu przez co najmniej dwa aparaty.

rozkład prawdopodobieństwa (otrzymany za pomocą drzewa probabilistycznego):

x_i	0	1	2	3	
p_i					

Przykład - rozkład dwumianowy $\mathrm{bin}(n,\ p)$

- dla pojedynczego aparatu możliwe są tylko dwa zdarzenia: zrobienie zdjęcia ('sukces') lub awaria ('porażka') – na drzewie probabilistycznym na każdym poziomie rysujemy tylko dwie gałęzie
- prawdopodobieństwo, że aparat zadziała (nastąpi awaria), nie zależy od zadziałania (awarii) pozostałych aparatów – na drzewie probabilistycznym na każdym poziomie prawdopodobieństwa, które wpisujemy na gałęziach, są takie same dla każdego sukcesu (0,6) i dla każdej porażki (0,4), odpowiednio)
- najmniejszą wartością, jaką przyjmuje zmienna losowa X zliczająca aparaty, które zadziałały, jest 0, natomiast największą jest 3 - liczba wszystkich aparatów ($X = 0, 1, 2, 3$)

Wniosek: $X \sim \mathrm{bin}(n, p) \quad \Rightarrow \quad X \sim \mathrm{bin}(3,\ 0{,}6)$

$$
\begin{aligned}
P(X = 0) &= \texttt{dbinom(0, 3, 0.6)} = \\
P(X \geqslant 2) &= P(X = 2) + P(X = 3) \\
&=
\end{aligned}
$$

Dystrybuanta

Dystrybuantą zmiennej losowej X nazywamy funkcję $F(X)$ określoną na zbiorze liczb rzeczywistych taką, że

$$F(x) = P(X \leqslant x).$$

Dystrybuanta zmiennej losowej dyskretnej: $F(x) = \sum_{x_i \leqslant x} p_i$ pname(x , param)

Dla dowolnych liczb rzeczywistych a i b zachodzi:

$$P(X \leqslant a) = F(a)$$
$$P(X > a) = 1 - F(a)$$
$$P(a < X \leqslant b) = F(b) - F(a)$$

Przykład

W pewnym eksperymencie wykorzystano trzy automatyczne aparaty fotograficzne w celu dokumentowania jego przebiegu. W danych warunkach prawdopodobieństwo wykonania poprawnej fotografii dla każdego aparatu jest takie samo i wynosi $p = 0,6$. Oblicz prawdopodobieństwo:
(a) nieudokumentowania eksperymentu;
(b) zarejestrowania eksperymentu przez co najmniej dwa aparaty.

$$X \sim \mathrm{bin}(3,\ 0{,}6)$$

$$P(X = 0) = \texttt{dbinom(0, 3, 0.6)} =$$

$$P(X \geqslant 2) = P(X > 1) = 1 - F(1) = 1 - \texttt{pbinom(1, 3, 0.6)} =$$

Wartość oczekiwana

Niech zmienna losowa dyskretna X przyjmuje wartości $x_1, x_2, \ldots$ odpowiednio z prawdopodobieństwami $p_1, p_2, \ldots$.

Definicja

Wartością oczekiwaną zmiennej losowej X nazywamy liczbę oznaczoną symbolem $E(X)$ i określoną wzorem

$$E(X) = \sum_{i=1}^{\infty} x_i p_i$$

o ile nieskończony szereg $\sum_{i=1}^{\infty} x_i p_i$ jest zbieżny.

Własności $E(X)$

Niech $a, b \in \mathbb{R}$.

(1) $E(a) = a$

(2) $E(bX) = b \cdot E(X)$

(3) $E(X + a) = E(X) + a$

(4) Niech $g : \mathcal{D} \to \mathbb{R}$, gdzie $x_i \in \mathcal{D}$, $i = 1, 2, \ldots$.

Wówczas: $$E[g(X)] = \sum_{i=1}^{\infty} g(x_i) \cdot p_i$$

Przykład

W pewnym eksperymencie wykorzystano trzy automatyczne aparaty fotograficzne w celu dokumentowania jego przebiegu. W danych warunkach prawdopodobieństwo wykonania poprawnej fotografii dla każdego aparatu jest takie samo i wynosi $p = 0{,}6$. Zdjęć zrobionych przez ile aparatów można się spodziewać (ile średnio aparatów udokumentuje przebieg eksperymentu)?

x_i	0	1	2	3	
p_i	$0{,}064$	$0{,}288$	$0{,}432$	$0{,}216$	1

$$E(X) =$$

Wariancja

Niech $\mu = E(X)$.

Definicja

Wariancją zmiennej losowej X nazywamy liczbę oznaczoną symbolem $D^2(X)$ i określoną wzorem

$$D^2(X) = E[(X-\mu)^2] = E(X^2) - \mu^2.$$

Jeżeli X – zmienna losowa dyskretna:

$$D^2(X) = \sum_{i=1}^{\infty} x_i^2 \cdot p_i - \mu^2$$

Własności $D^2(X)$

Niech $a, b \in \mathbb{R}$.

$$D^2(X) \geqslant 0 \qquad \text{(!ZAWSZE!)}$$
$$D^2(X + a) = D^2(X)$$
$$D^2(bX) = b^2 \cdot D^2(X)$$
$$D^2(a + bX) = b^2 \cdot D^2(X)$$

Definicja

Odchyleniem standardowym zmiennej losowej X nazywamy liczbę oznaczoną symbolem $D(X)$ i określoną wzorem

$$D(X) = \sqrt{D^2(X)}.$$

Przykład

W pewnym eksperymencie wykorzystano trzy automatyczne aparaty fotograficzne w celu dokumentowania jego przebiegu. W danych warunkach prawdopodobieństwo wykonania poprawnej fotografii dla każdego aparatu jest takie samo i wynosi $p = 0{,}6$. Oblicz odchylenie standardowe liczby aparatów dokumentujących przebieg eksperymentu.

x_i	0	1	2	3	
p_i	0,064	0,288	0,432	0,216	1

$$D^2(X) =$$

Inne rozkłady dyskretne

- rozkład równomierny
- rozkład zero-jedynkowy (Bernoulliego)
- rozkład Poissona
- rozkład geometryczny
- rozkład Pascala
- rozkład hipergeometryczny
- ...

Rozkłady dyskretne w R

name = nazwa rozkładu

param = parametry rozkładu

Gęstość:	d (density)	+	*name*	=	d*name*(x, *param*)
Dystrybuanta:	p (probability)	+	*name*	=	p*name*(x, *param*)
Kwantyl:	q (quantile)	+	*name*	=	q*name*(α, *param*)
Losowa obserwacja:	r (random)	+	*name*	=	r*name*(N, *param*)

Przykładowe nazwy rozkładów dyskretnych:

- dwumianowy: `binom`
- Poissona: `pois`

Histogram rozkładu dyskretnego (wykres liniowy):

```
plot(x, dname(x, param), type = "h")
```

Zmienna losowa ciągła

Funkcja gęstości prawdopodobieństwa

Funkcja gęstości prawdopodobieństwa $f(x)$ – funkcja opisująca rozkład prawdopodobieństwa zmiennej losowej ciągłej posiadająca następujące cechy:

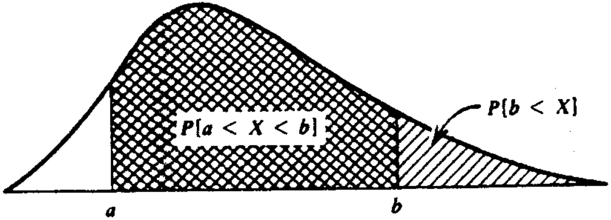
a) $f : \mathbb{R} \longrightarrow \mathbb{R}^+ \cup \{0\}$,

b) $P(a \leqslant X \leqslant b) =$ pole pod krzywą $f(x)$ między a i b.

Własności:

- $f(x) \geqslant 0$
- $P(a \leqslant X \leqslant b) = \int_a^b f(x)\,\mathrm{dx}$
- $\int_{-\infty}^{\infty} f(x)\,\mathrm{dx} = 1$
- $P(X > b) = 1 - P(X \leqslant b)$

Funkcja gęstości



$$
\begin{aligned}
P(a < X \leqslant b) &= P(a < X < b) \\
&= P(a \leqslant X \leqslant b) \\
&= P(a \leqslant X < b)
\end{aligned}
$$

$$P(X = a) = 0$$

Przykład

Czujnik śledzący stację wymaga dużej liczby wysokiej jakości taśm magnetycznych. Na taśmie magnetycznej mogą pojawić się rysy. Niech zmienna losowa X oznacza odległość (w cm) między kolejnymi rysami na powierzchni taśmy, a jej rozkład opisany jest funkcją gęstości

$$f(x) = \begin{cases} 0{,}01e^{-0{,}01x} & \text{dla} \quad x \geqslant 0, \\ 0 & \text{dla} \quad x < 0. \end{cases}$$

Załóżmy, że została znaleziona pierwsza rysa na taśmie. Oblicz prawdopodobieństwo, że kolejna zostanie znaleziona na kolejnych 50 cm taśmy.

X - zmienna losowa oznaczająca dystans między rysami

$$P(X \leqslant 50) = \int_0^{50} 0{,}01\mathrm{e}^{-0{,}01x}\,\mathrm{d}x = \texttt{integrate}(f,\, 0,\, 50)$$

$$f = \texttt{function}(x)\{0.01 * \texttt{exp}(-0.01 * x)\}$$

$$= 0{,}3934693$$

Dystrybuanta

Dystrybuantą zmiennej losowej X nazywamy funkcję $F(X)$ określoną na zbiorze liczb rzeczywistych taką, że

$$F(x) = P(X \leqslant x).$$

Dystrybuanta zmiennej losowej ciągłej: $F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$ pname(x, param)

Dla dowolnych liczb rzeczywistych a i b zachodzi:

$$P(X \leqslant a) = F(a)$$

$$P(X > a) = 1 - F(a)$$

$$P(a < X \leqslant b) = F(b) - F(a)$$

Przykład - rozkład wykładniczy $\mathrm{Exp}(\lambda)$

Czujnik śledzący stację wymaga dużej liczby wysokiej jakości taśm magnetycznych. Na taśmie magnetycznej mogą pojawić się rysy. Niech zmienna losowa X oznacza odległość (w cm) między kolejnymi rysami na powierzchni taśmy, a jej rozkład opisany jest funkcją gęstości

$$f(x) = \begin{cases} 0{,}01e^{-0{,}01x} & \text{dla} \quad x \geqslant 0, \\ 0 & \text{dla} \quad x < 0. \end{cases}$$

Załóżmy, że została znaleziona pierwsza rysa na taśmie. Oblicz prawdopodobieństwo, że kolejna zostanie znaleziona na kolejnych 50 cm taśmy.

X - zmienna losowa oznaczająca dystans między rysami

$$X \sim \mathrm{Exp}(0{,}01)$$

$$P(X \leqslant 50) = F(50) = \texttt{pexp(50, 0.01)} =$$

Wartość oczekiwana

Niech $f(x)$ – funkcja gęstości zmiennej losowej ciągłej X.

Definicja

Wartością oczekiwaną ciągłej zmiennej losowej X nazywamy

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x) \,\mathrm{dx}$$

(o ile całka istnieje).

Własności: (1) – (3) jak dla zmiennej losowej dyskretnej

(4) Niech $g : \mathcal{D} \longrightarrow \mathbb{R}$, gdzie $x \in \mathcal{D}$. Wówczas:

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f(x) \,\mathrm{dx}.$$

Przykład

Czujnik śledzący stację wymaga dużej liczby wysokiej jakości taśm magnetycznych. Na taśmie magnetycznej mogą pojawić się rysy. Niech zmienna losowa X oznacza odległość (w cm) między kolejnymi rysami na powierzchni taśmy, a jej rozkład opisany jest funkcją gęstości

$$f(x) = \begin{cases} 0{,}01e^{-0{,}01x} & \text{dla} \quad x \geqslant 0, \\ 0 & \text{dla} \quad x < 0. \end{cases}$$

Jaka przeciętnie odległość dzieli kolejne rysy na taśmie?

$$\mathrm{E}(X) = \int_0^\infty x \cdot 0{,}01\mathrm{e}^{-0{,}01x}\,\mathrm{d}x = \texttt{integrate}(f,\, 0,\, \texttt{Inf})$$

$$f = \texttt{function}(x)\{x * 0.01 * \texttt{exp}(-0.01 * x)\}$$

$$= 100$$

$$X \sim \mathrm{Exp}(0{,}01) \quad \Rightarrow \quad \mathrm{E}(X) = \frac{1}{\lambda} = \frac{1}{0{,}01} = 100$$

Wariancja

Niech $\mu = E(X)$.

Definicja

Wariancją zmiennej losowej X nazywamy liczbę oznaczoną symbolem $D^2(X)$ i określoną wzorem

$$D^2(X) = E[(X-\mu)^2] = E(X^2) - \mu^2.$$

Jeżeli X – zmienna losowa ciągła:

$$D^2(X) = \int_{-\infty}^{\infty} x^2 \cdot f(x)\,\mathrm{dx} - \mu^2$$

Własności: jak dla zmiennej losowej dyskretnej

Rozkład normalny

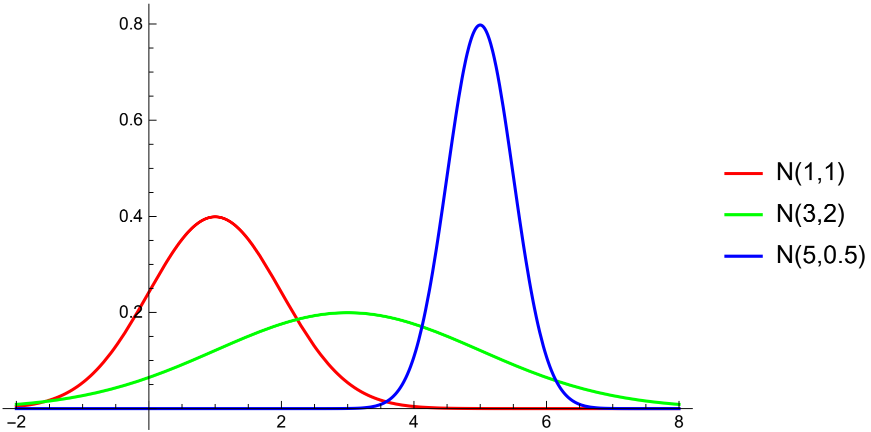
Funkcja gęstości ($\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}^+$):

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

Charakterystyki liczbowe:

$$E(X) = \mu, \qquad D^2(X) = \sigma^2$$

Rozkład normalny



Standardowy rozkład normalny $N(0,1)$

$\phi(x)$ – dystrybuanta $N(0,1)$ $\quad\diamondsuit$

$$X \sim N(\mu,\sigma) \quad \Rightarrow \quad Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$

$$P(X \leqslant b) = P\left(\tfrac{X-\mu}{\sigma} \leqslant \tfrac{b-\mu}{\sigma}\right) = P\left(Z \leqslant \tfrac{b-\mu}{\sigma}\right) = \phi\left(\tfrac{b-\mu}{\sigma}\right)$$

$$P(X \leqslant b) = F(b) = \texttt{pnorm}(b\,,\,\mu\,,\,\sigma)$$

$$P(a \leqslant X \leqslant b) = P\left(\tfrac{a-\mu}{\sigma} \leqslant Z \leqslant \tfrac{b-\mu}{\sigma}\right) = \phi\left(\tfrac{b-\mu}{\sigma}\right) - \phi\left(\tfrac{a-\mu}{\sigma}\right)$$

$$P(a \leqslant X \leqslant b) = F(b) - F(a) = \texttt{pnorm}(b\,,\,\mu\,,\,\sigma)-\texttt{pnorm}(a\,,\,\mu\,,\,\sigma)$$

Przykład

Niech X (w calach) będzie średnicą łożysk kulkowych produkowanych w pewnym zakładzie. Wedząc, że X podlega rozkładowi normalnemu z wartością oczekiwaną 1 cal oraz odchyleniem standardowym $0{,}001$ cala, tzn. $X \sim N(1,\ 0{,}001)$, oblicz prawdopodobieńswto, że średnica łożyska

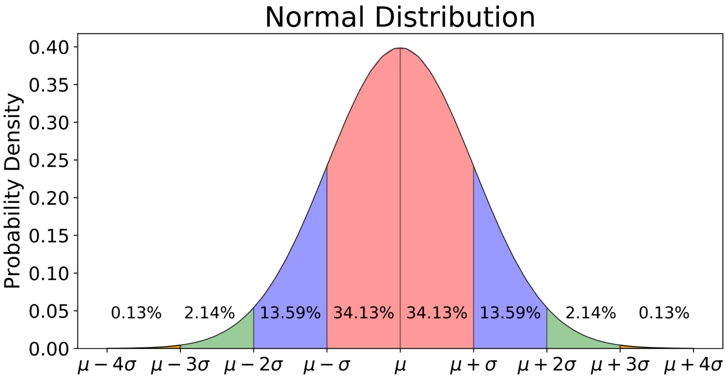
(a) nie przekracza $1{,}0015$ cala;
(b) przekracza $0{,}9995$ cala;
(c) znajduje się w przedziale od $0{,}9998$ do $1{,}0004$ cala.

$$P(X < 1{,}0015) = F(1{,}0015) = \texttt{pnorm(1.0015, 1, 0.001)} =$$

$$P(X > 0{,}9995) =$$

$$P(0{,}9998 < X < 1{,}0004) =$$

Reguła trzech sigm



Reguła trzech sigm

Niech $X \sim N(\mu, \sigma)$. Wówczas

$$P(\mu - \sigma < X < \mu + \sigma) = 0.683$$
$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.954$$
$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.997$$

Przybliżenie rozkładu dwumianowego

Niech $X \sim \mathrm{bin}(n, p)$, gdzie n jest duże i $p \in (0, 1)$. Wiadomo, że

$$E(X) = n \cdot p, \quad D^2(X) = n \cdot p \cdot q.$$

Wówczas

$$X \underset{\text{app}}{\sim} N(n \cdot p, \sqrt{n \cdot p \cdot q})$$

Przykład

Długoterminowe obszerne badania przeprowadzone w USA kilka lat temu wykazały, że 30% populacji dorosłych osób regularnie pije alkohol. Jeśli wyniki te są prawdziwe, to jakie jest prawdopodobieństwo, że w losowej próbie 1000 dorosłych liczba osób regularnie pijących alkohol jest nie większa niż 280?

A - zmienna losowa zliczająca osoby regularnie pijące alkohol

$$A \sim \mathrm{bin}(1000,\ 0{,}3)$$

$$A \underset{\text{app}}{\sim} N(1000 \cdot 0{,}3,\ \sqrt{1000 \cdot 0{,}3 \cdot 0{,}7})$$

$$P(A \leqslant 280) = F(280) =$$

$$P(A \leqslant 280) = F(280) \approx$$

Inne rozkłady ciągłe

- rozkład jednostajny (prostokątny)
- rozkład trójkątny
- rozkład t-Studenta
- rozkład Beta
- rozkład Gamma
- rozkład chi-kwadrat
- ...

Rozkłady ciągłe w R

name = nazwa rozkładu

param = parametry rozkładu

Gęstość:	d (density)	+	*name*	=	`d`*name*`(`x, *param*`)`
Dystrybuanta:	p (probability)	+	*name*	=	`p`*name*`(`x, *param*`)`
Kwantyl:	q (quantile)	+	*name*	=	`q`*name*`(`α, *param*`)`
Losowa obserwacja:	r (random)	+	*name*	=	`r`*name*`(`N, *param*`)`

Przykładowe nazwy rozkładów ciągłych:

wykładniczy:	`exp`	t-Studenta:	`t`
normalny:	`norm`	chi-kwadrat:	`chisq`
		F-Snedecora:	`f`

Wykres rozkładu ciągłego: `curve(d`*name*`(x,` *param*`))`

Najważniejsze informacje

Rozkłady dyskretne:

- dwumianowy bin(n,p):
  - prawdopodobieństwo: `dbinom(`*punkt*`,`*n*`,`*p*`)`
  - wykres: `plot(x, dbinom(x, `*n*`, `*p*`),type = "h")`
  - wartość oczekiwana i odch. std.: $\mathrm{E}(X) = n \cdot p$, $\mathrm{SD}(X) = \sqrt{n \cdot p \cdot (1-p)}$

Rozkłady ciągłe:

- wykładniczy EXP(λ):
  - prawdopodobieństwo (czerwony wzór): `pexp(`*punkt*`,`λ`)`
  - wykres: `curve(dexp(x, `λ`))`
  - wartość oczekiwana i odch. std.: $\mathrm{E}(X) = 1/\lambda$, $\mathrm{SD}(X) = 1/\lambda$
- normalny N(μ,σ):
  - prawdopodobieństwo (czerwony wzór): `pnorm(`*punkt*`,`μ`,`σ`)`
  - wykres: `curve(dnorm(x, `μ`, `σ`))`
  - wartość oczekiwana i odch. std.: $\mathrm{E}(X) = \mu$, $\mathrm{SD}(X) = \sigma$