Spis treści

1 Rachunek prawdopodobieństwa

1.1 Podstawy

1.1.1 Zdarzenia losowe, prawdopodobieństwo

Definicja 1. *Rodzinę $\mathcal{F}$ spełniającą warunki*

1. $\mathcal{F} \neq \emptyset$
2. *Jeśli* $A \in \mathcal{F}$, *to* $\Omega \setminus A \in \mathcal{F}$
3. *Jeśli* $A_i \in \mathcal{F}$ *dla* $i = 1, 2, \ldots,$ *to* $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

nazywamy σ – **ciałem podzbiorów zbioru** Ω.

Uwaga 1. Zdarzenie losowe *jest elementem rodziny* $\mathcal{F}$

Definicja 2. Prawdopodobieństwem *nazywamy dowolną funkcję* P, *określoną na* $\sigma-$*ciele zdarzeń* $\mathcal{F} \subseteq 2^{\Omega}$, *spełniającą warunki*

1. $P : \mathcal{F} \to \mathcal{R}_+$;
2. $P(\Omega) = 1$
3. *Jeśli* $A_i \in \mathcal{F},\ i = 1, 2, \ldots$ *oraz* $A_i \cap A_j = \emptyset$ *dla* $i \neq j$, *to*

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

Uwaga 2. *Mówimy, że matematyczny model doświadczenia losowego to trójka* $(\Omega, \mathcal{F}, P)$, *którą nazywamy* **przestrzenią probabilistyczną**

Twierdzenie 1. *Jeśli* $(\Omega, \mathcal{F}, P)$ *jest przestrzenią probabilistyczną i* $A,\ B,\ A_1,\ A_2, \ldots,\ A_n \in \mathcal{F}$, *to:*

1. $P(\emptyset) = 0$
2. *Jeśli* $A_1,\ A_2, \ldots,\ A_n$ *wykluczają się wzajemnie, tj.* $A_i \cap A_j = \emptyset$ *dla* $i \neq j$, *to*

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$$

3. $P(A') = 1 - P(A)$, *gdzie* $A' = \Omega \setminus A$

4. *Jeśli* $A \subseteq B$, *to* $P(B \setminus A) = P(B) - P(A)$

5. *Jeśli* $A \subseteq B$, *to* $P(A) \leqslant P(B)$

6. $P(A) \leqslant 1$

7. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

8. **Wzór włączeń i wyłączeń**

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = \sum_{1 \leqslant i \leqslant n} P(A_i) - \sum_{1 \leqslant i_1 \leqslant i_2 \leqslant n} P(A_{i_1} \cap A_{i_2}) + \\ \ldots + (-1)^{n+1} P(A_1 \cap A_2 \cap \ldots \cap A_n)$$

Twierdzenie 2 (O ciągłości)**.** *Niech*$(\Omega, \mathcal{F}, P)$ *będzie przestrzenią probabilistyczną.*

1. *Jeśli* $(A_n)_{n=1}^{\infty}$ *jest wstępującą rodziną zdarzeń oraz* $\bigcup_{n=1}^{\infty} A_n = A$, *to* $P(A) = \lim_{n \to \infty} P(A_n)$.

2. *Jeśli* $(A_n)_{n=1}^{\infty}$ *jest zstępującą rodziną zdarzeń oraz* $\bigcap_{n=1}^{\infty} A_n = A$, *to* $P(A) = \lim_{n \to \infty} P(A_n)$.

Rodzinę zdarzeń A_i *nazywamy wstępującą, jeśli*

$$A_1 \subseteq A_2 \subseteq \ldots \subset A_n \subseteq A_{n+1} \cdots$$

i zstępującą, jeśli

$$A_1 \supseteq A_2 \supseteq \ldots \supset A_n \supseteq A_{n+1} \cdots$$

1.1.2 Prawdopodobieństwo warunkowe

Definicja 3. **Prawdopodobieństwem warunkowym** *zajścia zdarzenia A pod warunkiem zajścia zdarzenia B, gdzie $P(B) > 0$, nazywamy liczbę*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Uwaga 3. *Przy ustalonym B prawdopodobieństwo warunkowe $P(A|B)$ jest zwykłym prawdopodobieństwem na $(\Omega, \mathcal{F})$, a także na $(B, \mathcal{F}_B)$, gdzie*

$$\mathcal{F}_B = \{A \cap B : \ A \in \mathcal{F}\}$$

Twierdzenie 3 (Wzór łańcuchowy)**.** *Jeśli $P(A_1 \cap \ldots \cap A_{n-1}) > 0$, to*

$$\begin{aligned} &P(A_1 \cap \ldots \cap A_n) = \\ &\quad = P(A_1)P(A_2|A_1) \cdot P(A_3|A_1 \cap A_2) \cdot \ldots \cdot P(A_n|A_1 \cap \ldots \cap A_{n-1}) \end{aligned}$$

Definicja 4. **Rozbiciem przestrzeni Ω** *nazywamy rodzinę zdarzeń $\{H_i\}_{i \in I}$, które wzajemnie wykluczają się, zaś ich suma jest równa Ω.*

Twierdzenie 4. *Jeżeli $\{H_1, \ H_2, \ldots, \ H_n\}$ jest rozbiciem Ω na zdarzenia o dodatnim prawdopodobieństwie, to dla dowolnego zdarzenia A*

$$P(A) = \sum_{i=1}^{n} P(A|H_i)P(H_i)$$

Uwaga 4. *Twierdzenie jest prawdziwe dla $n = \infty$.*

Uwaga 5. *Niech $\{H_i\}_{i \in I}$ będzie rozbiciem Ω na zdarzenia o dodatnim prawdopodobieństwie. Dla $P(B) > 0$ zachodzi*

$$P(A|B) = \sum_{i \in I} P(A|B \cap H_i)P(H_i|B),$$

gdzie zbiór indeksów I jest skończony lub przeliczalny.

Twierdzenie 5 (Wzór Bayes'a)**.** *Niech $\{H_i\}_{i \in I}$ będzie rozbiciem Ω na zdarzenia o dodatnim prawdopodobieństwie i $P(A) > 0$, to dla dowolnego $j \in I$ mamy*

$$P(H_j|A) = \frac{P(A|H_j)P(H_j)}{\sum_{i \in I} P(A|H_i)P(H_i)}$$

1.1.3 Niezależność zdarzeń

Zdarzenia niezależne

Zdarzenie B nie zależy od zdarzenia A, gdy wiedza o tym, że zaszło A nie wpływa na prawdopodobieństwo zajścia B.

$$\begin{gathered} P(B|A) = P(B), \quad P(A) > 0 \\ \Downarrow \\ P(A \cap B) = P(A)P(B) \end{gathered}$$

Definicja 5. *Zdarzenia A oraz B nazywamy* **niezależnymi**, *gdy*

$$P(A \cap B) = P(A)P(B)$$

Definicja 6. *Zdarzenia $A_1,\ A_2, \ldots,\ A_n$ nazywamy niezależnymi, gdy*

$$P(A_{i1} \cap A_{i2} \cap \ldots \cap A_{ik}) = P(A_{i1}) \ldots P(A_{ik})$$

dla $1 \leq i_i < i_2, \ldots < i_k \leq n,\ k = 2, 3, \ldots, n$

Przyjmijmy konwencję: $A^0 = A,\ A^1 = A'$

Twierdzenie 6. *Następujące warunki są równoważne:*

1. *Zdarzenia $A_1,\ A_2, \ldots,\ A_n$ są niezależne;*
2. *Dla każdego ciągu $\varepsilon_1,\ \varepsilon_2, \ldots,\ \varepsilon_n$, gdzie $\varepsilon_i \in \{0, 1\},\ i = 1, 2, \ldots, n$, zdarzenia $A_1^{\varepsilon_1}, \ldots, A_n^{\varepsilon_n}$ są niezależne;*
3. *Dla każdego ciągu $\varepsilon_1,\ \varepsilon_2, \ldots,\ \varepsilon_n$, gdzie $\varepsilon_i \in \{0, 1\},\ i = 1, 2, \ldots, n$, zachodzi równość*

$$P(A_1^{\varepsilon_1} \cap \ldots \cap A_n^{\varepsilon_n}) = P(A_1^{\varepsilon_1}) \ldots P(A_n^{\varepsilon_n})$$

Definicja 7. *Zdarzenia $A_1, A_2, \ldots$ nazywamy niezależnymi, gdy dla każdego n zdarzenia $A_1,\ A_2, \ldots,\ A_n$ są niezależne.*

1.2 Zmienna losowa

Zmienna losowa

Definicja 8. **Zmienna losowa** *jest to funkcja rzeczywista*

$$X : \Omega \to \mathcal{R}$$

o własności:

$$\bigwedge_{x \in \mathcal{R}} \{\omega \in \Omega : \ X(\omega) \leq x\} \in \mathcal{F}$$

Definicja 9. **Rozkładem prawdopodobieństwa zmiennej losowej** X *nazywamy rozkład prawdopodobieństwa* P_X *określony wzorem*

$$P_X(A) = P(\{\omega \in \Omega : \ X(\omega) \in A\}) \quad \textit{dla dowolnego } A \in \mathcal{B}(\mathcal{R})$$

Uwaga 6. *Oznaczamy* $X^{-1}(A) = \{\omega \in \Omega : \ X(\omega) \in A\}$

1.2.1 zmienne typu skokowego

Zmienne losowe typu skokowego

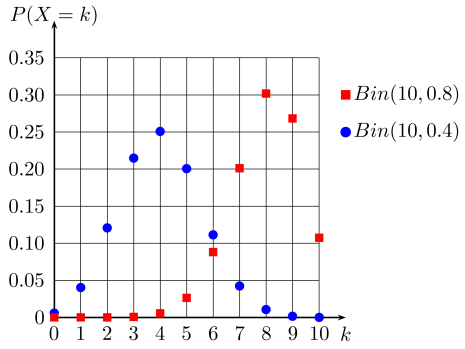
Definicja 10. *Mówimy, że zmienna losowa* $X : \Omega \to \mathcal{R}$ *jest* **typu skokowego** *(dyskretnego), jeżeli istnieje zbiór skończony lub przeliczalny* $\mathcal{X} \subset \mathcal{R}$ *taki, że*

$$P_X(\mathcal{X}) = 1$$

Definicja 11. *Zmienna losowa* X *ma* **rozkład dwumianowy** $Bin(n, p)$ *z parametrami* n *oraz* p*, jeżeli*

$$P_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n.$$

Schemat Bernoulliego
Wykonujemy doświadczenie Bernoulliego. Wyniki nazywane są umownie sukcesem oraz porażką. Prawdopodobieństwo sukcesu wynosi p. Doświadczenie wykonujemy w sposób niezależny n krotnie. Niech zmienną losową X będzie liczba sukcesów. Zmienna X ma rozkład $X \sim Bin(n, p)$.



Definicja 12. *Zmienna losowa X ma* **rozkład Poissona** *$Po(\lambda)$ z parametrem $\lambda > 0$, jeżeli*

$$P_X(k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \ldots.$$

Rozkład Poissona a rozkład dwumianowy
Jeżeli $X_n \sim Bin(n, p_n)$ oraz $\lim\limits_{n\to\infty} np_n = \lambda$, to

$$\lim_{n\to\infty} P_{X_n}(k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \ldots.$$

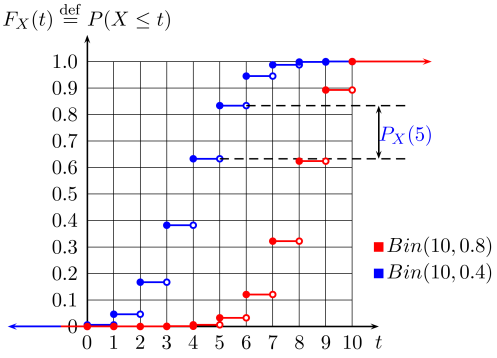
Definicja 13. *Zmienna losowa X ma* **rozkład ujemny dwumianowy** *$NB(r, p)$ z parametrami r oraz p, jeżeli*

$$P_X(k) = \binom{r+k-1}{k} p^r (1-p)^k, \quad k = 0, 1, 2, \ldots$$

Liczba porażek do r–tego sukcesu
Wykonujemy **doświadczenie Bernoulliego**. Wyniki nazywane są umownie *sukcesem* oraz *porażką*. Prawdopodobieństwo sukcesu wynosi p. Doświadczenie wykonujemy w sposób niezależny aż do uzyskania r-tego sukcesu. Niech zmienną losową X będzie liczba porażek. Zmienna X ma rozkład $NB(r, p)$.

Definicja 14. *Zmienna losowa X ma* **rozkład hipergeometryczny** *$H(n, N, M)$ z parametrami n, N oraz M, jeżeli*

$$P_X(k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}, \quad \max\{0, n-(N-M)\} \le k \le \min\{n, M\}.$$



Schemat urnowy
Z urny zawierającej N kul, w tym M białych, losujemy bez zwracania n kul. Zmienną losową jest liczba wylosowanych kul białych. Ta zmienna losowa ma rozkład hipergeometryczny $H(n, N, M)$

Definicja 15. **Dystrybuanta** *zmiennej losowej X, jest to funkcja $F : \mathcal{R} \to [0, 1]$ określona wzorem*

$$F_X(x) = P(X \le x) \text{ dla } x \in \mathcal{R}$$

Uwaga 7. *Zapis $\{X \le x\}$ oznacza zbiór $\{\omega \in \Omega : \ X(\omega) \le x\}$*

Dystrybuanta $F : \mathcal{R} \to [0, 1]$ ma następujące własności

1. **dystrybuanta jest funkcją niemalejącą**
2. $\lim\limits_{x \to -\infty} F(x) = 0, \quad \lim\limits_{x \to \infty} F(x) = 1$
3. **dystrybuanta jest funkcją prawostronnie ciągłą**
4. $P(a < X \le b) = F(b) - F(a)$
5. $P(X = a) = F(a) - F(a-)$
6. $P(a \le X \le b) = F(b) - F(a-)$
7. $P(a < X < b) = F(b-) - F(a)$

Uwaga 8. *$F(a-)$ oznacza $\lim\limits_{x \to a^-} F(x)$*

1.2.2 zmienne typu ciągłego

Definicja 16. *Mówimy, że zmienna losowa o dystrybuancie F jest* **typu ciągłego**, *jeżeli istnieje taka funkcja $f \geq 0$, że dla każdego $x \in \mathcal{R}$ zachodzi równość*

$$F(x) = \int_{-\infty}^{x} f(u)\, du$$

Funkcję f nazywamy **gęstością prawdopodobieństwa** *zmiennej losowej X lub w skrócie* **gęstością**

Własności

1. W punktach, w których f jest ciągła zachodzi $\frac{d}{dx}F(x) = f(x)$
2. $\int_{-\infty}^{\infty} f(x)\, dx = 1$
3. Każda nieujemna funkcja f spełniająca: $\int_{-\infty}^{\infty} f(x)\, dx = 1$, wyznacza dystrybuantę F za pomocą wzoru

$$F(x) = \int_{-\infty}^{x} f(u)\, du$$

Definicja 17. *Zmienna losowa X ma* **rozkład jednostajny** *$U(a, b)$ na przedziale (a, b), jeżeli jej funkcja gęstości rozkładu prawdopodobieństwa określona jest wzorem*

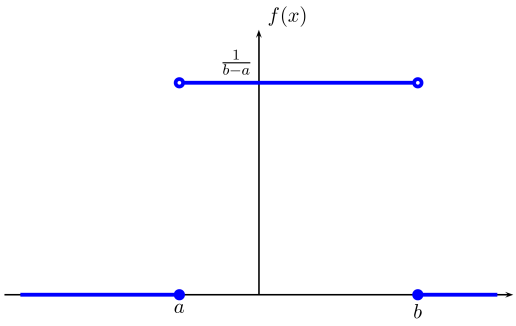
$$f(x) = \frac{1}{b-a}\mathbf{1}_{(a,b)}(x).$$

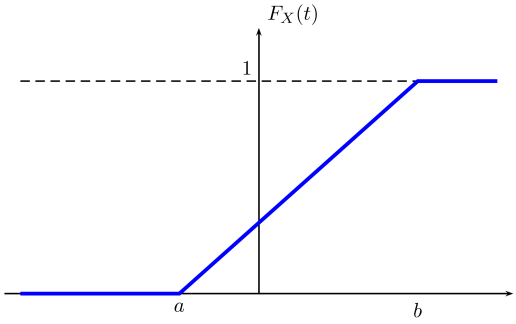
Definicja 18. *Zmienna losowa X ma* **rozkład wykładniczy** *$E(\lambda)$ z parametrem $\lambda > 0$, jeżeli jej funkcja gęstości rozkładu prawdopodobieństwa określona jest wzorem*

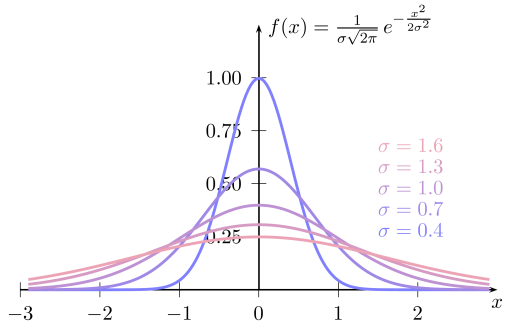
$$f(x) = \frac{1}{\lambda} \exp\left\{-\frac{x}{\lambda}\right\} \mathbf{1}_{(0,\infty)}(x)$$

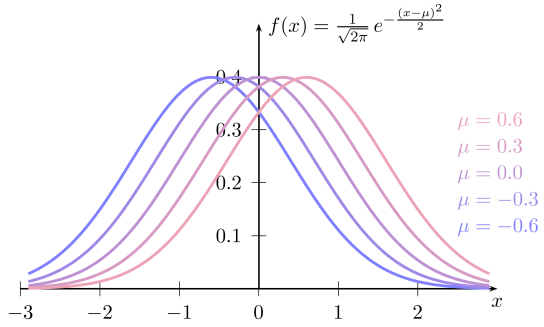
Definicja 19. *Zmienna losowa X ma* **rozkład normalny** *$N(\mu, \sigma^2)$, jeżeli funkcja gęstości jej rozkładu prawdopodobieństwa wyraża się wzorem*

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}.$$


$f(x)$
$\frac{1}{b-a}$
a
b


$F_X(t)$
1
a
b


$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{x^2}{2\sigma^2}}$
1.00
0.75
0.50
0.25
$\sigma = 1.6$
$\sigma = 1.3$
$\sigma = 1.0$
$\sigma = 0.7$
$\sigma = 0.4$
−3
−2
−1
0
1
2
x



1.2.3 Funkcje zmiennej losowej

Funkcje zmiennej losowej

Przykład
Niech $Y = aX + b$, gdzie $a \neq 0$ oraz X jest zmienną losową o rozkładzie

$$P_X(0) = 1/4, \;\; P_X(1) = 3/4.$$

Chcemy znaleźć rozkład zmiennej losowej Y.

$$P_X(0) = P_Y(b) = 1/4$$
$$P_X(1) = P_Y(a + b) = 3/4$$

Przykład
Niech X będzie zmienną losową typu ciągłego o gęstości f_X, dystrybuancie F_X oraz niech $Y = aX + b, \;\; a < 0$. Chcemy znaleźć rozkład Y

$$F_Y(y) = P(Y \leq y) = P(X \geq \frac{y-b}{a}) =$$
$$= 1 - P\left(X < \frac{y-b}{a}\right) = 1 - F_X\left(\frac{y-b}{a}\right)$$

Zatem $f_Y(y) = \frac{d}{dy} F_Y(y) = -\frac{1}{a} f_X\left(\frac{y-b}{a}\right)$

Przykład
Niech X oznacza zmienną losową ciągłą o dystrybuancie F_X oraz gęstości f_X. Niech f_X jest funkcją ciągłą, a g funkcją ściśle monotoniczną oraz niech $h = g^{-1}$. Wtedy dystrybuantą zmiennej losowej $Y = g(X)$ jest:

(dla g - rosnącej)

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(g(X) \leq y) \\ &= P(X \leq h(y)) = F_X(h(y)) \end{aligned}$$

Jeżeli h jest funkcją różniczkowalną, to

$$\frac{d}{dy} F_Y(y) = f_X(h(y))h'(y)$$

jest gęstością zmiennej losowej $Y = g(X)$ (dla g - malejącej)

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(g(X) \leq y) \\ &= P(X \geq h(y)) = 1 - F_X(h(y)) \end{aligned}$$

Jeżeli h jest funkcją różniczkowalną, to

$$\frac{d}{dy} F_Y(y) = f_X(h(y))(-h'(y))$$

jest gęstością zmiennej losowej $Y = g(X)$

Zatem w obu przypadkach $f_Y(y) = f_X(h(y))|h'(y)|$

Twierdzenie 7. *Niech X będzie zmienną losową typu ciągłego. Niech g będzie funkcją określoną na zbiorze $\bigcup_{k=1}^{n}[a_k, b_k]$, która na każdym przedziale otwartym (a_k, b_k) jest funkcją ściśle monotoniczną oraz ma ciągłą pochodną. Niech $h_k(y)$ będzie funkcją odwrotną do funkcji $g(x)$ na przedziale $I_k = g((a_k, b_k))$. Wówczas funkcja gęstości zmiennej losowej $Y = g(X)$ ma następującą postać*

$$f_Y(y) = \sum_{k=1}^{n} f_X(h_k(y)) \cdot |h_k'(y)| \cdot I_{I_k}(y)$$

1.3 Parametry zmiennej losowej

Wartość oczekiwana i wariancja zmiennej losowej

Definicja 20. *Wartość oczekiwaną (wartość przeciętna, nadzieję matematyczną) zmiennej losowej X oznaczamy symbolem $E(X)$ i określamy w następujący sposób:*

- *Jeżeli X jest zmienną losową typu skokowego, $\mathbb{X} = \{x_1, x_2, \ldots\}$, przy czym szereg*
$$\sum_k |x_k| P(X = x_k)$$
jest zbieżny, to
$$E(X) = \sum_k x_k P(X = x_k)$$

- *Jeżeli X jest zmienną losową typu ciągłego o gęstości f i zbieżna jest całka*
$$\int_{\mathcal{R}} |x| f(x)\, dx,$$
to
$$E(X) = \int_{\mathcal{R}} x\, f(x)\, dx$$

- **Ogólnie:** $E(X) = \int_\Omega X(\omega) dP(\omega)$

Własności wartości oczekiwanej

Jeżeli $E(X) < \infty,\ E(Y) < \infty$, to

- $E(X + Y) = E(X) + E(Y)$
- $E(aX + b) = aE(X) + b, \quad$ dla $a, b \in \mathcal{R}$
- Jeżeli $X \geq 0$, to $E(X) = \int_0^\infty P(X > t)\, dt$
- Jeżeli X oraz Y są niezależne, to
$$E(XY) = E(X)E(Y)$$

Twierdzenie 8. *Jeżeli funkcja φ jest borelowska, to*

- *Dla X z rozkładu skokowego*

$$E(\varphi(X)) = \sum_k \varphi(x_k) P(X = x_k)$$

- *Dla X z rozkładu ciągłego o gęstości $f(x)$*

$$E(\varphi(X)) = \int_{\mathcal{R}} \varphi(x) f(x)\, dx$$

Definicja 21. *Jeżeli $E(X - EX)^2 < \infty$, to tę liczbę nazywamy wariancją zmiennej losowej X i oznaczamy:*

$$D^2X = E(X - EX)^2$$

Definicja 22. *Pierwiastek z wariancji nazywamy odchyleniem standardowym i oznaczamy przez DX*

Uwaga 9.

$$\begin{aligned} D^2X = E(X - EX)^2 &= E(X^2 - 2X \cdot EX + (EX)^2) \\ &= EX^2 - (EX)^2 \end{aligned}$$

Własności wariancji
Jeżeli X jest zmienną losową, dla której $EX^2 < \infty$, to istnieje D^2X oraz:

- $D^2X \geq 0$
- $D^2(cX) = c^2 D^2X$
- $D^2(X + a) = D^2X$
- $D^2X = 0$ *wtedy i tylko wtedy, gdy zmienna losowa X jest z prawdopodobieństwem 1 stała*

Uwaga 10.

$$\begin{aligned} E(X-t)^2 =& E(X-EX+EX-t)^2 \\ =& E(X-EX)^2 + E(X-t)^2 - \\ & - 2E((X-EX)(EX-t)) \\ =& E(X-EX)^2 + E(X-t)^2 - \\ & - 2E(X-EX)\cdot E(EX-t) \\ \geq& E(X-EX)^2 \end{aligned}$$

Zatem funkcja $f(t) = E(X-t)^2$ *przyjmuje minimum – równe wariancji – dla* $t = EX$

1.4 Wektory losowe

Wektory losowe

Definicja 23. *Wektor losowy* $\mathbb{X} = (X_1, \ldots, X_n)$ *to odwzorowanie*

$$\mathbb{X} : \Omega \to \mathcal{R}^n$$

o własności:

$$\{\omega \in \Omega : \ X_1(\omega) \leq x_1, \ldots, X_n(\omega) \leq x_n\} \in \mathcal{F}$$

dla dowolnego $(x_1, x_2, \ldots, x_n) \in \mathcal{R}^n$

Definicja 24. *Rozkładem prawdopodobieństwa wektora losowego* $\mathbb{X}$ *nazywamy rozkład prawdopodobieństwa* $P_{\mathbb{X}}$ *określony wzorem*

$$P_{\mathbb{X}}(A) = P(\{\omega \in \Omega : \ \mathbb{X}(\omega) \in A\}) \quad dla\ A \in \mathcal{B}(\mathcal{R}^n)$$

Definicja 25. *Funkcja* $F_{\mathbb{X}} : \mathcal{R}^n \to [0, 1]$ *postaci*

$$F_{\mathbb{X}}(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n)$$

nazywamy dystrybuantą wektora losowego $\mathbb{X}$

Definicja 26. *Wektor losowy jest typu skokowego, jeżeli istnieje zbiór przeliczalny* $\mathcal{X} \subset \mathcal{R}^n$*, taki że* $P_{\mathbb{X}}(\mathcal{X}) = 1$

Definicja 27. *Wektor losowy jest typu ciągłego, jeżeli istnieje nieujemna funkcja* $f_{\mathbb{X}}(x_1, x_2, \ldots, x_n)$*, zwana gęstością, taka że dla każdego* $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathcal{R}^n$

$$F_{\mathbb{X}}(\mathbf{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_2} f_{\mathbb{X}}(u_1, \ldots, u_n)\, du_1 \ldots du_n$$

Uwaga 11. *Prawie wszędzie ma miejsce równość*

$$\frac{\partial F_{\mathbb{X}}(x_1, \ldots, x_n)}{\partial x_1, \ldots, \partial x_n} = f_{\mathbb{X}}(x_1, \ldots, x_n)$$

Dla dowolnego $A \in \mathcal{B}(\mathcal{R}^n)$ *zachodzi*

$$\int_A f_{\mathbb{X}}(\mathbf{x})\, d\mathbf{x}$$

Zauważmy, że

$$\begin{aligned} P(X_1 \in A) &= P(X_1 \in A, X_2 \in \mathcal{R}, \ldots, X_n \in \mathcal{R}) \\ &= \int_A \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbb{X}}(x_1, \ldots, x_n)\, dx_1 \ldots dx_n \\ &= \int_A \left(\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbb{X}}(x_1, \ldots, x_n)\, dx_2 \ldots dx_n \right) dx_1 \end{aligned}$$

Zatem

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbb{X}}(x_1, \ldots, x_n)\, dx_2 \ldots dx_n$$

Jest to tzw. **brzegowa gęstość prawdopodobieństwa.**

Rozkłady brzegowe, przypadek zmiennych typu ciągłego

$$\begin{aligned} f_{(X_1,X_2)}(x_1, x_2) &= \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbb{X}}(x_1, \ldots, x_n)\, dx_3 \ldots dx_n \end{aligned}$$

$$\begin{aligned} f_{(X_1,X_2,X_3)}(x_1, x_2, x_3) &= \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbb{X}}(x_1, \ldots, x_n)\, dx_4 \ldots dx_n \end{aligned}$$

Rozkłady brzegowe, przypadek zmiennych dyskretnych
Niech wektor losowy (X, Y) ma rozkład określony liczbami

$$p_{ik} = P(X = x_i, Y = y_k), \text{ gdzie } i \in I,\ k \in K.$$

Wówczas rozkład zmiennej losowej X określają liczby

$$p_i = P(X = x_i) = \sum_{k \in K} p_{ik}, \text{ gdzie } i \in I$$

Definicja 28. *Niech $(\Omega, \mathcal{F}, P)$ będzie przestrzenią probabilistyczną, a $X_1,\ X_2, \ldots,\ X_n$ będą zmiennymi losowymi określonymi na tej przestrzeni. Mówimy, że te zmienne losowe są niezależne, jeżeli dla dowolnych zbiorów borelowskich $A_1,\ A_2, \ldots,\ A_n$ zachodzi:*

$$P(X_1 \in A_1, \ldots, X_n \in A_n) = P(X_1 \in A_1) \ldots P(X_n \in A_n)$$

Definicja 29. *Mówimy, że zmienne losowe $X_1,\ X_2, \ldots$ są niezależne, jeżeli każdy skończony podciąg ciągu $X_1,\ X_2, \ldots$ składa się z niezależnych zmiennych losowych*

Twierdzenie 9. *Dla zmiennych losowych $X_1,\ X_2, \ldots,\ X_n$ następujące warunki są równoważne*

- *zmienne losowe są niezależne*
- *dla $\mathbf{x} = (x_1,\ x_2, \ldots,\ x_n) \in R^n$*

$$F_{\mathbb{X}}(\mathbf{x}) = F_{X_1}(x_1) \ldots F_{X_n}(x_n)$$

Twierdzenie 10. *Jeżeli $\mathbb{X} = (X_1,\ X_2, \ldots,\ X_n)$ jest wektorem losowym typu skokowego to warunkiem koniecznym i wystarczającym niezależności zmiennych losowych $X_1,\ X_2, \ldots,\ X_n$ jest:*

$$P(X_1 = x_1, \ldots, X_n = x_n) = P_1(X_1 = x_1) \ldots P_n(X_n = x_n)$$

dla każdego $(x_1, \ldots, x_n) \in \mathcal{R}^n$, gdzie P_k oznacza brzegowy rozkład prawdopodobieństwa zmiennej losowej X_k $(k = 1, 2, \ldots, n)$.

Twierdzenie 11. *Jeżeli $\mathbb{X} = (X_1,\ X_2, \ldots,\ X_n)$ jest wektorem losowym typu ciągłego o gęstości $f_{\mathbb{X}}$, to warunkiem koniecznym i wystarczającym niezależności zmiennych losowych $X_1,\ X_2, \ldots,\ X_n$ jest:*

$$f_{\mathbb{X}}(\mathbf{x}) = f_{X_1}(x_1) \ldots f_{X_n}(x_n),$$

dla każdego $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{R}^n$, gdzie f_{X_k} jest gęstością rozkładu brzegowego zmiennej losowej X_k $(k = 1, \ldots, n\ )$

1.5 Parametry rozkładów - wektory losowe

Wartość oczekiwana oraz macierz kowariancji wektora losowego

Definicja 30 (Wartość oczekiwana wektora losowego)**.**

$$E(\mathbb{X}) = E\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} EX_1 \\ EX_2 \\ \vdots \\ EX_n \end{bmatrix}$$

Definicja 31 (Macierz kowariancji wektora losowego)**.**

$$D^2(\mathbb{X}) = \begin{bmatrix} Cov(X_1, X_1) & Cov(X_1, X_2) & \cdots & Cov(X_1, X_n) \\ Cov(X_2, X_1) & Cov(X_2, X_2) & \cdots & Cov(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_n, X_1) & Cov(X_n, X_2) & \cdots & Cov(X_n, X_n) \end{bmatrix},$$

gdzie

$$Cov(X_i, X_j) = E[(X_i - EX_i)(X_j - EX_j)] = E(X_iX_j) - EX_iEX_j$$

Podstawowe własności
Jeżeli $A,\ B$ są macierzami odpowiedniego wymiaru, to

- $E(A\mathbb{X}) = AE(\mathbb{X})$
- $E(A\mathbb{X}B) = AE(\mathbb{X})B$
- $D^2(A\mathbb{X}) = AD^2(\mathbb{X})A'$

Wielowymiarowy rozkład normalny $N(\mu, \Sigma)$

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)\right],\ \mathbf{x} \in \mathcal{R}^n$$

- $E(\mathbb{X}) = \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}$

- $D^2(\mathbb{X}) = \Sigma$
- $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$

1.6 Rodzaje zbieżności

Definicja 32. *Ciąg zmiennych losowych $(X_n)_{n=1}^{\infty}$ jest zbieżny do zmiennej losowej X* **według prawdopodobieństwa**, *jeżeli*

$$\text{dla każdego } \varepsilon > 0 \quad \lim_n P(|X_n - X| > \varepsilon) = 0$$

oznaczenie: $X_n \xrightarrow{P} X$

Definicja 33. *Ciąg zmiennych losowych $(X_n)_{n=1}^{\infty}$ jest zbieżny do zmiennej losowej X* **prawie na pewno**, *jeżeli*

$$P\left(\left\{\omega : \lim_n X_n(\omega) = X(\omega)\right\}\right) = 1$$

oznaczenie $X_n \xrightarrow{p.n.} X$

Definicja 34. *Ciąg zmiennych losowych $(X_n)_{n=1}^{\infty}$ jest zbieżny do zmiennej losowej X* **według rozkładu**,*jeżeli ciąg dystrybuant $(F_{X_n})_{n=1}^{\infty}$ jest zbieżny do dystrybuanty F_X w każdym punkcie jej ciągłości.*

oznaczenie $X_n \xrightarrow{D} X$

$$(X_n \xrightarrow{p.n.} X) \Rightarrow (X_n \xrightarrow{P} X) \Rightarrow (X_n \xrightarrow{D} X)$$

Prawa wielkich liczb

Oznaczmy

$$S_n = X_1 + X_2 + \cdots + X_n, \quad \bar{X}_n = \frac{X_1 + X_2 + \ldots +, X_n}{n}$$

Niech $X_1,\ X_2, \ldots$ będzie ciągiem niezależnych zmiennych losowych o tym samym rozkładzie, o wartości średniej μ i wariancji $0 < \sigma^2 < \infty$. Wtedy zachodzi

Słabe prawo wielkich liczb

$$(\forall \varepsilon > 0) \quad \lim_{n\to\infty} P\left(\left|\bar{X}_n - \mu\right| < \varepsilon\right) = 1$$

$$\boxed{\bar{X}_n \xrightarrow{P} \mu}$$

Mocne prawo wielkich liczb

$$P\left(\lim_{n\to\infty} \bar{X}_n = \mu\right) = 1$$

$$\boxed{\bar{X}_n \overset{p.n.}{\to} \mu}$$

Centralne twierdzenie graniczne

$$\sup_{x\in\mathcal{R}} \left| P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right) - \Phi(x) \right| \xrightarrow[n\to\infty]{} 0$$

$$\boxed{\frac{\bar{X}_n - \mu}{\sigma}\sqrt{n} \overset{D}{\to} N(0,1)}$$

2 Wnioskowanie statystyczne

2.1 Podstawowe pojęcia

Wnioskowanie statystyczne

Statystyka Nauka poświęcona metodom badania (analizowania) zjawisk masowych; polega na systematyzowaniu obserwowanych cech ilościowych i jakościowych; posługuje się rachunkiem prawdopodobieństwa.

Pojęcia podstawowe

Populacja Zbiór obiektów z wyróżnioną cechą (cechami).Obiektami mogą być przedmioty lub wartości cechy

Próba Wybrana część populacji podlegająca badaniu. Próba powinna stanowić reprezentację populacji w tym sensie, że częstości występowania w próbie każdej z badanych cech nie powinny się znacznie różnić od częstości występowania tych cech w populacji

Cecha losowa Wielkość losowa charakteryzująca obiekty danej populacji

Rodzaje cech

niemierzalna – zwana też jakościową – przyjmuje wartości nie będące liczbami (np. *kolor*, *płeć*,*smak*)

mierzalna – zwana też ilościową – przyjmuje pewne wartości liczbowe (np. *długość, wytrzymałość, ciężar*)

Rodzaje cech mierzalnych

skokowa – zwana też dyskretną – nie przyjmuje wartości pośrednich (np. *ilość bakterii*, *ilość pracowników*, *ilość pasażerów*,).

ciągła – przyjmuje wartości z pewnego przedziału liczbowego (np. *wzrost*, *waga*, *ciśnienie*, *czas obsługi*)

Przykłady parametrów charakteryzujących populację

Mediana badanej cechy to wartość, która dzieli populację na dwie części. Połowa obiektów w populacji ma cechę o wartości poniżej mediany, a połowa powyżej.

Kwartyl dolny badanej cechy to wartość, która dzieli populację w stosunku 1:3. Jedna czwarta obiektów w populacji ma cechę o wartości poniżej kwartyla dolnego, a pozostałe trzy czwarte powyżej.

Kwartyl górny badanej cechy to wartość, która dzieli populację w stosunku 3:1. Trzy czwarte obiektów w populacji ma cechę o wartości poniżej kwartyla dolnego, a pozostała jedna trzecia powyżej.

Średnia ...

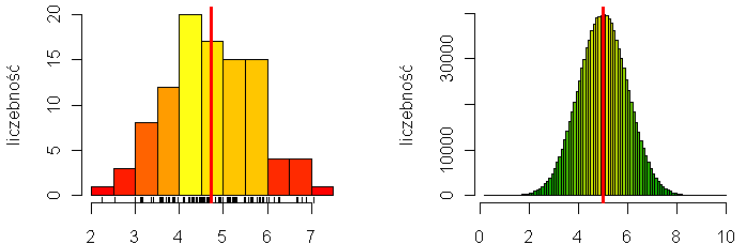
Wariancja ...

Mierniki położenia i rozproszenia próby; przykłady

Niech X oznacza cechę losową. Niech wartości $x_1, x_2, \ldots, x_n$ oznaczają n realizacji tej cechy. Przez $x_{1:n}, x_{2:n}, \ldots, x_{n:n}$ będą oznaczane realizacje tej cechy w kolejności od najmniejszej do największej.

Mierniki położenia	Oznaczenia	Wzór
średnia	$\bar{x}$	$\frac{1}{n}\sum_{i=1}^{n} x_i$
mediana	Me	$x_{(n+1)/2:n}$ (gdy n nieparzyste) $(x_{n/2:n} + x_{n/2+1:n})/2$ (gdy n parzyste)
dolny kwartyl	Q_1	$x_{[n/4]:n}$
górny kwartyl	Q_3	$x_{[3n/4]:n}$
dominanta (moda)	D	najczęściej występująca wartość
minimium	Min	$x_{1:n}$
maksimum	Max	$x_{n:n}$

$[a]$ – część całkowita liczby a



Mierniki rozproszenia	Oznaczenia	Wzór
rozstęp	R	$Max - Min$
wariancja	S^2	$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
odchylenie standardowe	S	$\sqrt{S^2}$
odchylenie przeciętne	d	$\frac{1}{n}\sum_{i=1}^{n}\lvert x_i - \bar{x}\rvert$
odchylenie ćwiartkowe	Q	$\frac{Q_3 - Q_1}{2}$
współczynnik zmienności	V	$\frac{S}{\bar{x}}100\%$

Wnioskowanie statystyczne polega na wnioskowaniu o populacji na podstawie próby

Przykład
Badamy rozkład wartości cechy w populacji na podstawie próby 100 elementowej. Średnia z próby $\bar{x} = 4.717641$. W tym przykładzie wiadomo, że średnia populacyjna $\mu = 5$

2.2 Estymacja punktowa

Estymacja punktowa

Niech $X_1, X_2, \ldots, X_n$ oznacza próbę z populacji oraz θ parametr charakteryzujący tę populację. Na podstawie próby chcemy oszacować (przybliżyć) wartość parametru θ.

Estymator punktowy jest funkcją próby. Przybliża wartość parametru θ:

$$\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$$

Estymacja punktowa parametrów cechy $X \sim N(\mu, \sigma^2)$

Estymator średniej μ — średnia arytmetyczna

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + \cdots + X_n}{n}$$

Estymator wariancji σ^2 — wariancja próbkowa

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Estymator odchylenia standardowego σ

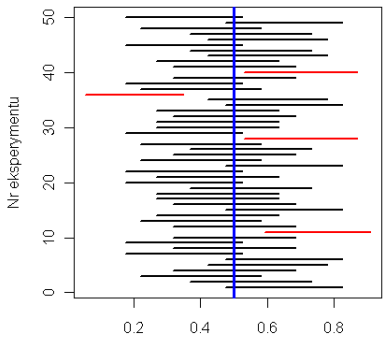
$$S = \sqrt{S^2}$$

Estymacja punktowa parametru p **cechy** $X \sim D(p)$

Niech n oznacza liczbę obiektów wylosowanych z populacji, wśród których znalazło się k obiektów, które posiadają wyróżnioną właściwość. Przyjmując, że p oznacza prawdopodobieństwo wylosowania z populacji obiektu o wyróżnionej właściwości mamy:

$$\hat{p} = \frac{k}{n}$$

Uwaga. Przyjmując dla $i = 1, 2, \ldots, n$, że $P(X_i = 1) = p = 1 - P(X_i = 0)$, mamy $\hat{p} = \bar{X}$.



2.3 Estymacja przedziałowa

Estymacja przedziałowa

Przedział ufności (estymator przedziałowy) jest przedziałem o końcach zależnych od próby, który z pewnym z góry zadanym prawdopodobieństwem $1-\alpha$ pokrywa nieznaną wartość parametru θ:

$$P\{\theta \in (\underline{\theta}(X_1,\ldots,X_n), \overline{\theta}(X_1,\ldots,X_n))\} \geq 1-\alpha \quad (\forall\theta).$$

Poziom ufności jest to ustalone prawdopodobieństwo $1-\alpha$.

Ilustracja estymacji przedziałowej parametru $\theta = 0.5$ (oznaczonego pionową niebieską linią) na poziomie ufności $1-\alpha = 0.9$ Populacja z wyróżnioną cechą X

Przedział ufności dla średniej μ w rozkładzie normalnym $N(\mu, \sigma^2)$

Wariancja σ^2 jest nieznana

Poziom ufności: $1-\alpha$

$$(\bar{X} - t(1-\alpha/2; n-1)\frac{S}{\sqrt{n}}, \bar{X} + t(1-\alpha/2; n-1)\frac{S}{\sqrt{n}})$$

$t(\gamma;\nu)$ jest stablicowanym kwantylem rzędu γ rozkładu t (t–Studenta) z ν stopniami swobody.

	Kwantyl rozkładu $t-$ Studenta			
	γ			
ν	0.9500	0.9750	0.9875	0.9950
8	1.8595	2.3060	2.7515	3.3554
9	1.8331	2.2622	2.6850	3.2498
10	1.8125	2.2281	2.6338	3.1690

Przykład
Na podstawie próby 1.1, 1.2, 0.8, 0.9, 1.2, 1.3, 1.0, 0.7, 0.8, 1.0 oszacować wartość średnią μ rozkładu obserwowanej cechy $X \sim N(\mu, \sigma^2)$, na poziomie ufności $1 - \alpha = 0.95$.

$$\bar{x} = \frac{1.1 + 1.2 + \cdots + 1.0}{10} = 1.0$$

$$\sum (x_i - \bar{x})^2 = (1.1 - 1.0)^2 + \cdots + (1.0 - 1.0)^2 = 0.36$$

$$s^2 = \frac{0.36}{10 - 1} = 0.04, \ s = \sqrt{s^2} = 0.2$$

$$t(0.975; 9) = 2.2622$$

$$t(0.975; 9)\frac{s}{\sqrt{n}} = 2.2622\frac{0.2}{\sqrt{10}} = 0.14$$

$$(1 - 0.14, 1 + 0.14) = (0.86, 1.14)$$

Wniosek. Średnia wartość cechy jest jakąś liczbą z przedziału $(0.86, 1.14)$. Zaufanie do tego wniosku wynosi 95%.

Przedział ufności dla wariancji w rozkładzie normalnym

Średnia μ jest nieznana

Poziom ufności: $1 - \alpha$

$$\left(\frac{\sum\limits_i (X_i - \bar{X})^2}{\chi^2(\frac{\alpha}{2}; n - 1)}, \frac{\sum\limits_i (X_i - \bar{X})^2}{\chi^2(1 - \frac{\alpha}{2}; n - 1)} \right)$$

$\chi^2(\alpha; \nu)$ jest stablicowaną wartością krytyczną rozkładu chi–kwadrat z ν stopniami swobody.

ν	Wartości krytyczne $\chi^2(\alpha; r)$ α			
	0.975	0.950	0.050	0.025
8	2.1797	2.7326	15.5073	17.5345
9	2.7004	3.3251	16.9190	19.0228
10	3.2470	3.9403	18.3070	20.4832

Przykład
Na podstawie próby 1.1, 1.2, 0.8, 0.9, 1.2, 1.3, 1.0, 0.7, 0.8, 1.0 oszacować zróżnicowanie rozkładu obserwowanej cechy.

$$\bar{x} = \frac{1.1 + 1.2 + \cdots + 1.0}{10} = 1.0$$

$$\sum_i (x_i - \bar{x})^2 = (1.1 - 1.0)^2 + \cdots + (1.0 - 1.0)^2 = 0.36$$

$$s^2 = \frac{0.36}{10 - 1} = 0.04,\ s = \sqrt{s^2} = 0.2$$

Poziom ufności $1 - \alpha = 0.95$, czyli $\alpha = 0.05$.

$$\chi^2(\frac{\alpha}{2}; n - 1) = \chi^2(0.025; 9) = 19.0228$$

$$\chi^2(1 - \frac{\alpha}{2}; n - 1) = \chi^2(0.975; 9) = 2.7004$$

$$\left(\frac{0.36}{19.0228}, \frac{0.36}{2.7004}\right) = (0.019, 0.133)$$

Wniosek. Wariancja cechy jest liczbą z przedziału $(0.019, 0.133)$. Zaufanie do tego wniosku wynosi 95%.

Estymacja prawdopodobieństwa sukcesu
Przedział przybliżony

$$\left(\hat{p} - u_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + u_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

	Kwantyle u_α rozkładu normalnego $N(0,1)$				
α	0.002	0.003	0.004	0.005	0.006
0.96	1.7744	1.7866	1.7991	1.8119	1.8250
0.97	1.9110	1.9268	1.9431	1.9600	1.9774
0.98	2.0969	2.1201	2.1444	2.1701	2.1973
0.99	2.4089	2.4573	2.5121	2.5758	2.6521

Na przykład $u_{0.975} = 1.96$
Populacja 1, cecha X_1

Populacja 2, cecha X_2

Oznaczenia

Próby: $X_{11}, \ldots, X_{1n_1}$; $X_{21}, \ldots, X_{2n_2}$

$$\bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}, \quad s_i^2 = \frac{\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2}{n_i - 1}$$

$$s_e^2 = \frac{\sum_{j=1}^{n_1}(X_{1j} - \bar{X}_1)^2 + \sum_{j=1}^{n_2}(X_{2j} - \bar{X}_2)^2}{n_1 + n_2 - 2}, \quad s_r^2 = s_e^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

Ocena różnicy między średnimi $\mu_1 - \mu_2$

Ocena punktowa: $\bar{X}_1 - \bar{X}_2$

Założenia:

1. $X_1 \sim N(\mu_1, \sigma_1^2), X_2 \sim N(\mu_2, \sigma_2^2)$
2. X_1, X_2 są niezależne
3. $\sigma_1^2 = \sigma_2^2$

Przedział ufności (poziom ufności $1 - \alpha$)

$$(\bar{X}_1 - \bar{X}_2 - t(1-\alpha/2; n_1+n_2-2)s_r, \bar{X}_1 - \bar{X}_2 + t(1-\alpha/2; n_1+n_2-2)s_r)$$

Przykład. Z dwóch populacji pobrano próby: 60, 62, 65, 63, 60 oraz 58, 53, 57, 56, 61. Ocenić różnicę średnich.

$$\bar{x}_1 = 62, \sum_{i=1}^{5}(x_{1i} - \bar{x}_1)^2 = 18, \bar{x}_2 = 57, \sum_{i=1}^{5}(x_{2i} - \bar{x}_2)^2 = 34$$

$$s_r^2 = \frac{18+34}{5+5-2}\left(\frac{1}{5}+\frac{1}{5}\right) = 2.6$$

$$t(0.975; 8) = 2.3060; t(0.975; 8)s_r = 3.72$$

$$(62 - 57 - 3.72, 62 - 57 + 3.72) = (1.28, 8.72)$$

Wniosek. Różnica średnich jest liczbą z przedziału $(1.28, 8.72)$ **Ocena różnicy frakcji** $p_1 - p_2$

Założenia: $X_1 \sim D(p_1), \quad X_2 \sim D(p_2)$
Cechy X_1, X_2 są niezależne

Próba 1: $X_{11}, X_{12}, \ldots, X_{1n_1}$ ($X_{1i} = 0$ lub 1)
Próba 2: $X_{21}, X_{22}, \ldots, X_{2n_2}$ ($X_{2i} = 0$ lub 1)

$k_1 = \sum_{i=1}^{n_1} X_{1i}$
$k_2 = \sum_{i=1}^{n_2} X_{2i}$

Ocena punktowa: $\hat{p}_1 - \hat{p}_2 = \frac{k_1}{n_1} - \frac{k_2}{n_2}$

Przybliżony przedział ufności (poziom ufności $1 - \alpha$)

$$\hat{p}_1 - \hat{p}_2 \pm u_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Iloraz frakcji: $\frac{p_1}{p_2}$ **(ryzyko względne)**

$$\ln(\frac{\hat{p}_1}{\hat{p}_2}) \pm u_{1-\frac{\alpha}{2}}\sqrt{\frac{1-\hat{p}_1}{n_1\hat{p}_1} + \frac{1-\hat{p}_2}{n_2\hat{p}_2}}$$

Przykład: Porównanie lekarstw ze względu na odsetek osób, które nie reagują na podany lek

p_1	p_2	$p_1 - p_2$	p_1/p_2
0.01	0.001	0.009	10
0.410	0.401	0.009	1.02

Rozkład prawdopodobieństwa oraz dane

	Y	
X	p_{11}	p_{12}
	p_{21}	p_{22}

	Y	
X	n_{11}	n_{12}
	n_{21}	n_{22}

Iloraz szans $\theta = \frac{p_{11}/p_{12}}{p_{21}/p_{22}}$

Estymator ilorazu szans $\hat{\theta} = \frac{\hat{p}_{11}/\hat{p}_{12}}{\hat{p}_{21}/\hat{p}_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$

Przedział ufności dla $\ln(\theta)$

$$\ln(\hat{\theta}) \pm u_{1-\frac{\alpha}{2}} \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

2.4 Weryfikacja hipotez statystycznych

Weryfikacja hipotez statystycznych

Błąd I rodzaju Błąd wnioskowania polegający na odrzuceniu hipotezy, gdy w rzeczywistości jest ona prawdziwa.

Błąd II rodzaju Błąd wnioskowania polegający na nieodrzuceniu hipotezy, gdy w rzeczywistości jest ona fałszywa.

	Decyzja o hipotezie	
Hipoteza	nie odrzucić	odrzucić
prawdziwa	prawidłowa	błędna
fałszywa	błędna	prawidłowa

Błąd I rodzaju kontroluje się przez zadanie małej wartości dla poziomu istotności. **Poziom istotności** jest to górne ograniczenie prawdopodobieństwa popełnienia błędu I rodzaju.

Błędu II rodzaju nie można kontrolować w taki sposób, jak błąd I rodzaju. W praktyce nie wiadomo, ile dokładnie wynosi prawdopodobieństwo popełnienia tego błędu.

Porównanie średniej z normą

Cecha X ma rozkład normalny $N(\mu, \sigma^2)$

Średnia μ oraz wariancja σ^2 są nieznane

$$H_0 : \ \mu = \mu_0$$

Test Studenta (poziom istotności α)

Próba: $X_1, \ldots, X_n$

Statystyka testowa

$$t_{\text{emp}} = \frac{\bar{X} - \mu_0}{S}\sqrt{n} \ .$$

Jeżeli $|t_{\text{emp}}| > t(1 - \alpha/2; n - 1)$, to hipotezę odrzucamy.

Przykład. W biochemicznym doświadczeniu badano czas życia komórek w pewnym środowisku. Dokonano ośmiu pomiarów uzyskując wyniki (w godzinach): 4.7, 5.3, 4.0, 3.8, 6.2, 5.5, 4.5, 6.0. Czy można uznać, że średni czas życia komórek w badanym środowisku wynosi 4 godziny?

Cecha X — czas życia komórki ($X \sim N(\mu, \sigma^2)$)

$$H_0 : \mu = 4$$

Test Studenta; poziom istotności $\alpha = 0.05$

$$\bar{x} = 5, \ s = 0.891227, \ t_{\text{emp}} = 3.1736, \ t(0.975; 7) = 2.3646$$

Weryfikacja: Ponieważ $t_{\text{emp}} > t(0.975; 7)$, odrzucamy hipotezę

Wniosek: średni czas życia komórek w badanym środowisku nie wynosi 4 godziny.

Porównanie zróżnicowania z normą

Cecha X ma rozkład normalny $N(\mu, \sigma^2)$
Średnia μ oraz wariancja σ^2 są nieznane

$$H_0 : \sigma^2 = \sigma_0^2$$

Statystyka chi–kwadrat (poziom istotności α)
Próba: $X_1, \ldots, X_n$

Statystyka testowa $\chi^2_{\text{emp}} = \frac{\sum_i (X_i - \bar{X})^2}{\sigma_0^2}$

Wartości krytyczne $\chi^2(1 - \frac{\alpha}{2}; n-1)$, $\chi^2(\frac{\alpha}{2}; n-1)$

Jeżeli $\chi^2_{\text{emp}} < \chi^2(1 - \frac{\alpha}{2}; n-1)$ lub $\chi^2_{\text{emp}} > \chi^2(\frac{\alpha}{2}; n-1)$ to hipotezę $H_0 : \sigma^2 = \sigma_0^2$ odrzucamy.

Porównanie frakcji z normą

Cecha $X \sim D(p)$

p nie jest znane

$$H_0 : \ p = p_0$$

Statystyka testowa

$$u_{\text{emp}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Wartość krytyczna: $u_{1-\frac{\alpha}{2}}$

Jeżeli $|u_{\text{emp}}| > u_{1-\frac{\alpha}{2}}$, to hipotezę odrzucamy. **Przykład.** Dziesięć lat temu odsetek dzieci chorych na astmę wynosił 4%. Czy odsetek ten uległ zmianie, jeżeli w próbie dwustu dzieci rozpoznano osiemnaście przypadków astmy?

Niech X oznacza liczbę przypadków astmy wśród wylosowanych dzieci.

Możemy założyć, że $X \sim B(200, p)$, gdzie p oznacza prawdopodobieństwo wylosowania dziecka chorego na astmę.

Cel: Zweryfikować hipotezę $H_0 : \ p = 0.04$

Zadaję poziom istotności $\alpha = 0.05$.

Wyznaczam $\hat{p} = 0.09, \ u_{\text{emp}} = \frac{0.09-0.04}{\sqrt{\frac{0.04(1-0.04)}{200}}} = 2.887, \ u_{0.975} = 1.96$

Ponieważ $|u_{\text{emp}}| > u_{0.975}$, hipotezę odrzucamy.

Wniosek: Odsetek dzieci chorych na astmę uległ zmianie.

Porównanie średnich

Cecha X_1 ma rozkład normalny $N(\mu_1, \sigma_1^2)$
Cecha X_2 ma rozkład normalny $N(\mu_2, \sigma_2^2)$
Średnia μ_1 oraz wariancja σ_1^2 są nieznane
Średnia μ_2 oraz wariancja σ_2^2 są nieznane
$\sigma_1^2 = \sigma_2^2$

$$H_0 : \mu_1 = \mu_2$$

test t–Studenta

$$t_{\text{emp}} = \frac{\bar{X}_1 - \bar{X}_2}{S_r}$$

Wartość krytyczna $t(1 - \alpha/2; n_1 + n_2 - 2)$
Jeżeli $|t_{\text{emp}}| > t(1-\alpha/2; n_1+n_2-2)$, to hipotezę $H_0 : \mu_1 = \mu_2$ odrzucamy

Porównanie frakcji

Cecha X_1 ma rozkład dwupunktowy $D(p_1)$
Cecha X_2 ma rozkład dwupunktowy $D(p_2)$

$$H_0 : \ p_1 = p_2$$

Statystyka testowa

$$u_{\text{emp}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

gdzie

$$\hat{p}_1 = \frac{k_1}{n_1}, \ \hat{p}_2 = \frac{k_2}{n_2}, \ \hat{p} = \frac{(k_1 + k_2)}{(n_1 + n_2)}$$

Jeżeli $|u_{\text{emp}}| \geq u_{1-\alpha/2}$, to hipotezę $H_0 : p_1 = p_2$ odrzucamy

2.5 Regresja

MODEL REGRESJI LINIOWEJ

Przyjmujemy następujący model:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

gdzie $\varepsilon_i, \quad i = 1, 2, \ldots, n$, są niezależnymi zmiennymi losowymi o tym samym rozkładzie $N(0, \sigma^2)$.

Uwagi

- $Y_1, Y_2, \ldots, Y_n$, są zmiennymi losowymi, a $y_1, y_2, \ldots, y_n$ są ich realizacjami.
- Model dotyczy rozkładu warunkowego $Y|X = x$.

Metoda najmniejszych kwadratów

Parametry β_0 oraz β_1 dobieramy tak, aby średniokwadratowy błąd dopasowania, mianowicie $\sum_i e_i^2 = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$, był minimalny. W ten sposób dobrane parametry oznaczamy przez $\hat{\beta}_0$ oraz $\hat{\beta}_1$. Wyrażają się one wzorami

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x}_i)^2}, \ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

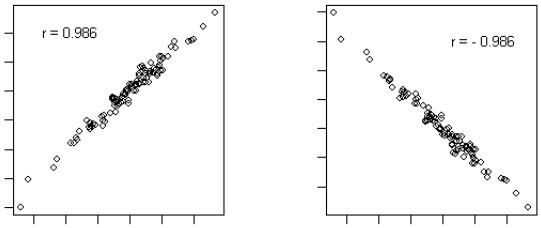
Zachodzi

$$\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = (1 - r^2) \sum_i (y_i - \bar{y})^2,$$

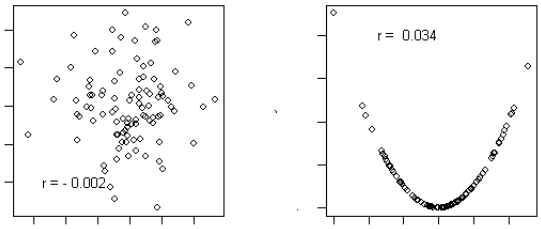
gdzie

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (y_i - \bar{y}_i)^2 \sum_i (x_i - \bar{x})^2}}$$

Współczynnik r jest miernikiem zależności liniowej.



Wartość r jest zawsze z przedziału <-1,1>



Estymacja

- $\hat{\beta}_0,\ \hat{\beta}_1$ są oszacowaniami punktowymi parametrów β_0 oraz β_1.
- Oszacowania przedziałowe dla β_0 oraz β_1 są postaci

$$\beta_1 \in (\hat{\beta}_1 - t(\alpha; n-2)S_{\beta_1}, \hat{\beta}_1 + t(\alpha; n-2)S_{\beta_1})$$

$$\beta_0 \in (\hat{\beta}_0 - t(\alpha; n-2)S_{\beta_0}, \hat{\beta}_0 + t(\alpha; n-2)S_{\beta_0})$$

gdzie

$$S^2_{\beta_1} = \frac{S^2}{\mathrm{var}x}, \quad S^2_{\beta_0} = \frac{S^2}{\mathrm{var}x}\left(\frac{\mathrm{var}x}{n} + \bar{x}^2\right)$$

$$S^2 = \frac{\mathrm{var}y - \hat{\beta}_1 \mathrm{cov}(x,y)}{n-2} = \frac{\mathrm{var}y(1-r^2)}{n-2}$$

Weryfikacja hipotez

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

Statystyka testowa

$$F_{\text{emp}} = \frac{\hat{\beta}_1^2}{S_{\beta_1}^2} = \frac{\hat{\beta}_1 \text{cov}(x,y)}{S^2}$$

Hipotezę odrzucamy, jeżeli $F_{\text{emp}} > F(\alpha; 1, n-2)$.

$F(\alpha; 1, n-2)$ jest wartością krytyczną rozkładu F.

Weryfikacja hipotez

$$H_0 : \beta_1 = a$$
$$H_1 : \beta_1 \neq a$$

Statystyka testowa

$$t_{\text{emp}} = \frac{\hat{\beta}_1 - a}{S_{\beta_1}}$$

Hipotezę odrzucamy, jeżeli $|t_{\text{emp}}| > t(\alpha; n-2)$.

$t(\alpha; n-2)$ jest wartością krytyczną rozkładu $t-$Studenta.

Obszar ufności dla prostej regresji oraz obszar predykcji

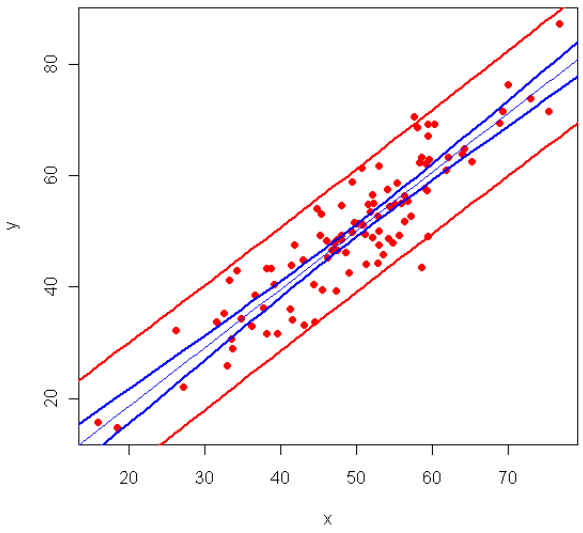
Obszar ufności dla prostej regresji

Obszar ufności dla prostej regresji umożliwia nam wnioskowanie o wartościach średnich zmiennej Y jednocześnie dla wielu wybranych wartości zmiennej X.

$$f(x) \in (\hat{f}(x) - t(\alpha; n-2)S_Y; \hat{f}(x) + t(\alpha; n-2)S_Y)$$

gdzie

$$\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$
$$S_Y^2 = S^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\text{var}x} \right)$$



Obszar predykcji

Obszar predykcji umożliwia nam wnioskowanie o wartościach zmiennej Y jednocześnie dla wielu wybranych wartości zmiennej X.

$$Y(x) \in (\hat{f}(x) - t(\alpha; n-2)S_{Y(x)}; \hat{f}(x) + t(\alpha; n-2)S_{Y(x)})$$

gdzie $Y(x)$ oznacza wartość zmiennej Y dla wybranej wartości x zmiennej X oraz

$$S^2_{Y(x)} = S^2 \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\text{var}x}\right)$$