

Elementarna statystyka

Przedziały ufności

Alexander Bendikov

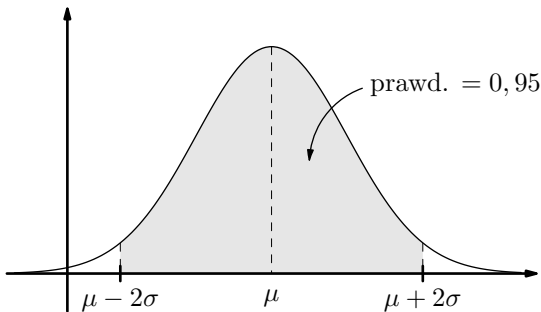
Uniwersytet Wrocławski

26 kwietnia 2017

1. Estymatory punktowe

- $X, x_1, x_2, \dots$ zmienna losowa z obserwacjami $x_1, x_2, \dots$
- Rozkład prawdopodobieństwa zmiennej X zależy od parametrów $a, b, c, \dots$
- **Podstawowe zadanie:** Mając obserwacje $x_1, x_2, \dots$ oszacować parametry $a, b, c, \dots$

Przykład: Zmienne o rozkładzie normalnym $X \sim N(\mu, \sigma)$. Gęstość prawdopodobieństwa $p(x) = p(x; \mu, \sigma)$ zależy od parametrów μ, σ



Korzystając z Prawa Wielkich Liczb mamy:

- 1) $\bar{x} \approx \mu = E(X),$
- 2) $s^2 \approx \sigma^2 = \text{Var}(X) = E(X - \mu)^2,$
- 3) $r^2 \approx R_{X,Y} = E\left(\frac{X - \mu_X}{\sigma_X}\right)\left(\frac{Y - \mu_Y}{\sigma_Y}\right).$

Uwagi: 1. Estymatory $\bar{x}, s^2, r^2$ są zmiennymi losowymi, są to tak zwane estymatory punktowe: dla danej próby $x_1, x_2, \dots$ podają jedną liczbę (punkt).

2. Te estymatory są *nieobciążone*:

$$E \bar{x} = \mu, \quad E s^2 = \sigma^2, \quad E r^2 = R_{X,Y}.$$

2. Estymatory przedziałowe

Problem: Dla danej liczby $0 < C < 1$ (poziom ufności) znaleźć przedział (a_1, a_2) z losowymi końcami a_1, a_2 , zależny od próby $x_1, x_2, \dots, x_n$ taki, że

$$P(a \in (a_1, a_2)) = C.$$

Najczęściej wybieranymi poziomami ufności są $C = 0,95, 0,99, 0,999$. Przedział (a_1, a_2) nazywa się przedziałem ufności parametru a na poziomie ufności C

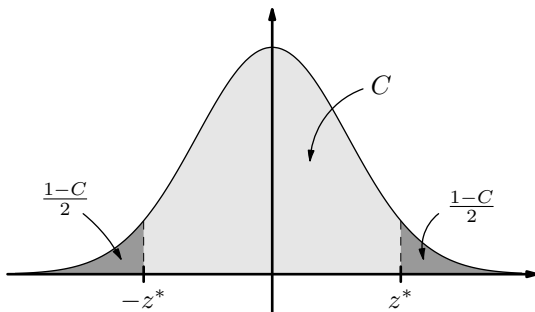
Definicja

Przedział ufności parametru a na poziomie ufności C obliczony na podstawie danych próby to przedział zawierający prawdziwą wartość parametru a z prawdopodobieństwem C .

Przykład: Przedział ufności dla średniej μ rozkładu normalnego $N(\mu, \sigma)$ gdy σ jest znane. Mamy

$$\bar{x} - z^* \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z^* \frac{\sigma}{\sqrt{n}}$$

z prawdopodobieństwem C , gdzie z^* jest tak zwaną wartością C -krytyczną



Rysunek: Wartość z^* dla danego C

Tablica: Charakterystyczne wartości z^*

C	0,7	0,8	0,9	0,95	0,99	0,999
$p = \frac{1-C}{2}$	0,15	0,10	0,05	0,025	0,005	0,0005
z^*	1,036	1,282	1,645	1,960	2,576	3,291

Przykład: Zmienna losowa X oznacza zawartość aktywnego składnika w pewnym leku. Wiemy, że ma rozkład normalny, i $\sigma = 3$. Znajdź przedział ufności μ na poziomie $C = 0,95$, na podstawie próby $x_1, \dots, x_n$ o wielkości $n = 36$ i średniej $\bar{x} = 4,1$.

Rozwiązanie: 1. W tabeli znajdujemy wartość krytyczną z^* dla $C = 0,95$, $z^* = 1,96$.

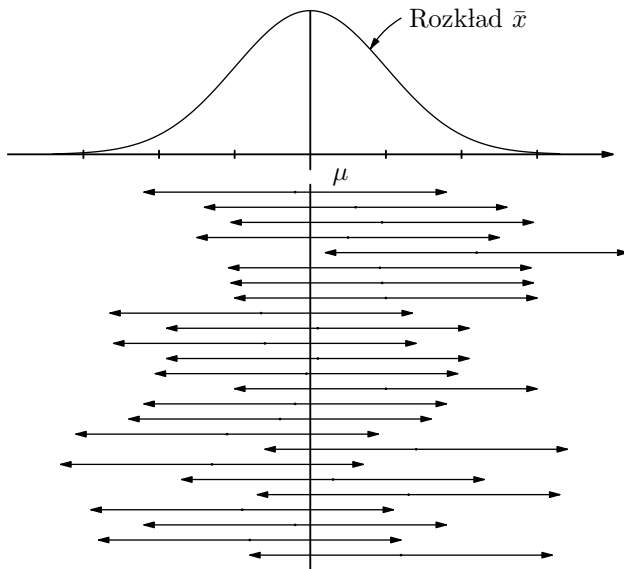
2. Znajdujemy margines błędu $z^* \frac{\sigma}{\sqrt{n}} = \frac{1,96 \cdot 3}{\sqrt{36}} = 0,98$.

3. Przedział ufności na poziomie $C = 0,95$ to

$$(\bar{x} - 0,98, \bar{x} + 0,98) = (3,12, 5,08),$$

a więc, na poziomie ufności $C = 0,95$ mamy $3,12 \leq \mu \leq 5,08$.

Uwaga: Poziom ufności oznacza prawdopodobieństwo, że dana metoda da prawidłową odpowiedź. Innymi słowy, jeżeli użyjemy poziomu ufności 0,95, to na dłuższą metę 95% wyznaczonych przedziałów ufności będzie zawierało prawdziwą wartość parametru (zauważmy, że pozostałe 5% nie będzie zawierało). W konkretnym przypadku nie ma całkowitej pewności, że prawdziwa wartość parametru rzeczywiście wpadnie do przedziału ufności.



Rysunek: 25 prób losowych z tej samej populacji dało powyższe 95% przedziały ufności. Średnio 95% prób losowych da przedział, który zawiera prawdziwą wartość μ populacji

Wyjaśnijmy

- $\bar{x} = \frac{x_1 + \dots + x_n}{n}$ jest zmienną losową normalną, z parametrami

$$E(\bar{x}) = \mu, \quad SD(\bar{x}) = \frac{\sigma}{\sqrt{n}}.$$

- Rozważmy zmienną losową unormowaną Z

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \in N(0, 1).$$

- Z nazywa się Z -statystyką dla $\bar{x}$.

$$\begin{aligned}
 P(-z^* \leq Z \leq z^*) &= C, \\
 -z^* &\leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq z^*, \\
 \bar{x} - z^* \frac{\sigma}{\sqrt{n}} &\leq \mu \leq \bar{x} + z^* \frac{\sigma}{\sqrt{n}}.
 \end{aligned}$$

- $\sigma/\sqrt{n}$ nazywa się *błędem standardowym* (SE), a $m = z^* \sigma/\sqrt{n}$ *marginiesem błędu*.

Przykład (kontynuacja): X to zawartość substancji aktywnej, $n = 36$, $\bar{x} = 4,1$, $\sigma = 3$. W takim razie błąd standardowy to $3/6 = 0,5$.

1)

$$\begin{aligned}
 C &= 0.95, \quad p = (1 - C)/2 = 0,025, \\
 z^* &= 1.96, \quad m = z^* \sigma/\sqrt{n} = 0,98, \\
 3,12 &\leq \mu \leq 5,08.
 \end{aligned}$$

2)

$$\begin{aligned}C &= 0.99, & p &= 0,005, \\z^* &= 2,576, & m &= 1,288, \\2,812 &\leq \mu \leq 5,388.\end{aligned}$$

3)

$$\begin{aligned}C &= 0.999, & p &= 0,0005, \\z^* &= 3,291, & m &= 1,645, \\2,455 &\leq \mu \leq 5,745.\end{aligned}$$

Wniosek: $m \nearrow \infty$ gdy $C \nearrow 1$.

Analiza przedziału ufności

1. Przy ustalonym poziomie ufności C niech $n \nearrow$. Wtedy margines błędu $m = z^* \sigma / \sqrt{n} \searrow$.
2. Przy ustalonym rozmiarze próby n niech $C \nearrow 1$. Wtedy margines błędu $m = z^* \sigma / \sqrt{n} \nearrow$.
3. Ustalmy poziom ufności C i margines błędu $m = z^* \sigma / \sqrt{n}$. Rozmiar próby potrzebny w tym przypadku to $n = \left(\frac{z^* \sigma}{m}\right)^2$.

Przykład: Niech X będzie ilością substancji powstałej w jakiejś reakcji chemicznej. Wiemy, że $X \in N(\mu, \sigma)$, gdzie $\sigma = 3$ a μ jest nieznane. Ile pomiarów należy wykonać i uśrednić, żeby rezultat był dokładny z $m = \pm 1$ dla poziomu ufności C : (a) 0,95, (b) 0,99, (c) 0,999?

Rozwiązanie: (a) $C = 0,95$, $z^* = 1,96$, $m = 1$.

$$n = \left(\frac{z^* \sigma}{m}\right)^2 = (1,96 \cdot 3)^2 = 34,57 \approx 35.$$

(b) $C = 0,99$, $z^* = 2,576$, $m = 1$.

$$n = \left(\frac{z^* \sigma}{m} \right)^2 = (2,576 \cdot 3)^2 = 59,75 \approx 60.$$

(c) $C = 0,999$, $z^* = 3,291$, $m = 1$.

$$n = \left(\frac{z^* \sigma}{m} \right)^2 = (3,291 \cdot 3)^2 = 97,42 \approx 98.$$

Przedział ufności dla średniej zmiennej normalnej (σ nieznane)

- $x_1, x_2, \dots, x_n$ obserwacje zmiennej $X \in N(\mu\sigma)$
- Przedział ufności dla μ na poziomie ufności C jest dany przez

$$\bar{x} - t_* \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_* \frac{s}{\sqrt{n}},$$

gdzie $\bar{x}$ jest średnią próby, s odchyleniem standardowym próby, a t_* jest wartością krytyczną (analogicznie do z^*) dla poziomu ufności C rozkładu t (zwanego też rozkładem Studenta).

Przykład: Zmienna X ma rozkład normalny. Znajdźmy przedział ufności dla średniej, na poziomie $C = 0,95$, na podstawie próby $x_1, x_2, \dots, x_n$ o liczności $n = 16$, $\bar{x} = 20,2$ i $s = 0,8$.

Rozwiązanie:

1. Korzystamy z tabeli wartości t_* (lub programu takiego jak Minitab).
Dla $C = 0,95$, $df = 16 - 1 = 15$ otrzymujemy $t_* = 2,13$
2. $m = t_* s / \sqrt{n} = 2,13 \cdot 0,8 / \sqrt{16} = 2,13 \cdot 0,2 \approx 0,426$
3. $\bar{x} - m = 20,2 - 0,426 = 19,774$, $\bar{x} + m = 20,2 + 0,426 = 20,626$
4. $19,774 \leq \mu \leq 20,626$.

Uwaga: Jeżeli próba jest mała $n < 30$ użycie tabeli dla zmiennej normalnej i przyjęcie $\sigma = s$ zamiast tabeli dla zmiennej t prowadzi do błędu – przedział ufności jest za mały.

Przykład: Rozważmy poprzedni przykład, i przyjmijmy, że σ jest znane i równe s .

1. Z tabeli dostajemy $z^* = 1,96$ (porównajmy z $t_* = 2,13$)
2. $m = z^* \sigma / \sqrt{n} = 1,96 \cdot 0,8 / 4 = 0,392$ (porównajmy do $m = 0,426$)
3. $19,808 \leq \mu \leq 20,592$.

Wyjašnjmy:

- $X, x_1, x_2, \dots, x_n$ zmienne losowe normalne $N(\mu, \sigma)$,
- $\bar{x} = \frac{1}{n}(x_1 + x_2 + \dots + x_n)$,
- $s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$,
- T -statystyka: $T = \frac{\bar{x} - \mu}{s/\sqrt{n}}$,
- Rozkład zmiennej T nie zależy od μ ani σ . Jest to tak zwany rozkład t (lub Studenta), odkryty przez Gosseta (który posługiwał się pseudonimem Student):

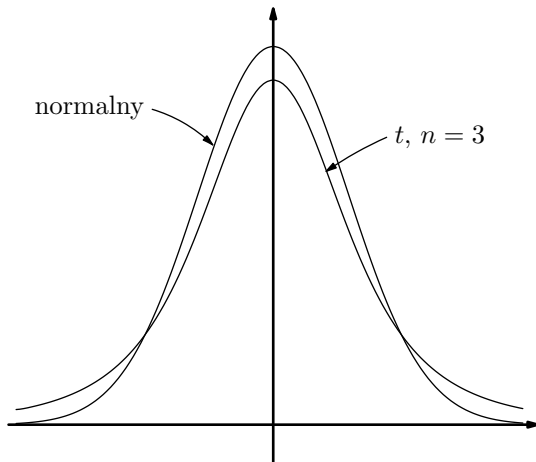
$$t(x, n) = B_n \left(1 + \frac{x^2}{n-1}\right)^{-\frac{n}{2}}, \quad \text{gdzie } B_n = \frac{\Gamma(n/2)}{\sqrt{\pi} \Gamma((n-1)/2)}.$$

- Przedział ufności dla rozkładów T

$$P(-t_* \leq T \leq t_*) = C$$
$$P(\bar{x} - t_* s / \sqrt{n} \leq \mu \leq \bar{x} + t_* s / \sqrt{n}) = C$$

- Przedział ufności dla średniej na poziomie C :

$$\bar{x} - t_* s / \sqrt{n} \leq \mu \leq \bar{x} + t_* s / \sqrt{n}.$$



Rysunek: Rozkład normalny i t z $n = 3$

Przedział ufności dla odchylenia standardowego σ

- $X, x_1, x_2, \dots, x_n$ zmienne losowe normalne $N(\mu, \sigma)$,
- Przedział ufności dla σ na poziomie ufności C jest dany przez:

$$s(1 - q_*) \leq \sigma \leq s(1 + q_*),$$

gdzie s jest odchyleniem standardowym próby, a q_* jest wartością krytyczną, dla poziomu ufności C pewnego rozkładu o którym za chwilę.

Przykład: Zmienna X ma rozkład normalny, otrzymaliśmy próbę $x_1, x_2, \dots, x_{25}$ o liczności $n = 25$, dla której $s = 0,8$. Znajdziemy przedział ufności dla σ na poziomie ufności $C = 0,95$.

Rozwiązanie: Wartość q_* odczytujemy z tabeli. Dla $C = 0,95$, $n = 25$ otrzymujemy $q_* = 0,32$.

1) $s(1 - q_*) = 0,8 \cdot (1 - 0,32) = 0,8 \cdot 0,68 = 0,544,$

$$s(1 + q_*) = 0,8 \cdot (1 + 0,32) = 0,8 \cdot 1,32 = 1,056,$$

2) $0,544 \leq \sigma \leq 1,056.$

Uwaga: Gdyby w naszych danych $n = 6$, to otrzymalibyśmy:

1) $q_* = 1,09$

2) $s(1 - q_*) = 0,8 \cdot (1 - 1,09) = 0,8 \cdot (-0,09) = -0,072,$
 $s(1 + q_*) = 0,8 \cdot (1 + 1,09) = 0,8 \cdot 2,09 = 1,672,$

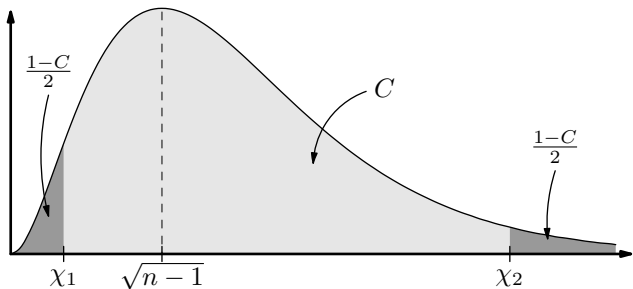
3) $-0,072 \leq \sigma \leq 1,672.$

Wiemy, że $\sigma > 0$, więc przedział możemy poprawić:

$$0 \leq \sigma \leq 1,672.$$

Wyjaśnijmy: rozkład χ^2

- Rozważmy χ^2 -statystykę, zmienną $\chi^2 = \frac{s^2}{\sigma^2}(n-1)$, $n > 2$



Rysunek: Rozkład zmiennej χ

- Rozkład zmiennej losowej $\chi^2 = \frac{s^2}{\sigma^2}(n-1)$ znany jest jako *rozkład χ^2* z $n-1$ stopniami swobody

- Znajdźmy wartości krytyczne χ_1 i χ_2 jak na rysunku, czyli

$$P(\chi^2 \leq \chi_1^2) = P(\chi^2 \leq \chi^2) = \frac{1 - C}{2}$$

- Zastąpmy nierówności dla $\chi^2 = \frac{s^2}{\sigma^2}(n - 1)$ nierównościami dla σ

$$\chi_1 \leq \frac{s}{\sigma} \sqrt{n - 1} \leq \chi_2$$
$$\frac{s}{\chi_2} \sqrt{n - 1} \leq \sigma \leq \frac{s}{\chi_1} \sqrt{n - 1}.$$

- Wtedy można wziąć

$$\frac{\sqrt{n - 1}}{\chi_2} = 1 - q_*, \quad \frac{\sqrt{n - 1}}{\chi_1} = 1 + q_*$$

Uwaga: Można używać bezpośrednio χ_1, χ_2 zamiast q_*

Przedziały ufności dla parametrów rozkładów innych niż normalne

Niech X będzie zmienną losową o rozkładzie dwumianowym $X \in B(N, p)$, gdzie parametr p jest nieznany. Wiemy:

$$P(X = k) = c_N^k p^k (1 - p)^{N-k}.$$

Niech $x_1, x_2, \dots, x_n$ będzie niezależną próbą losową z populacji o tym rozkładzie. Ustalamy

$$\hat{p} = \frac{1}{nN} \sum_{i=1}^n x_i.$$

Wtedy:

- $E(\hat{p}) = p$
- $Var(\hat{p}) = \frac{p(1-p)}{nN}$
- $\frac{\hat{p}-p}{\sqrt{p(1-p)/nN}} \longrightarrow Z \in N(0, 1)$ gdy $n \rightarrow \infty$

W szczególności,

$$P\left(-z_* < \frac{\hat{p} - p}{\sqrt{p(1-p)/nN}} < z_*\right) \approx P(-z_* < Z < z_*).$$

Dla danego poziomu ufności C znajdujemy z_* takie, że

$$P\left(-z_* < Z < z_*\right) = C.$$

Wtedy przedział ufności dla nieznanego parametru p na poziomie C ma postać $x_1 < p < x_2$, gdzie x_1, x_2 są rozwiązaniami równania kwadratowego

$$x^2 - \frac{2\hat{p} + \frac{z_*^2}{Nn}}{1 + \frac{z_*^2}{Nn}} x + \frac{\hat{p}^2}{1 + \frac{z_*^2}{Nn}} = 0.$$

Jeżeli $\frac{z_*}{\sqrt{Nn}}$ jest małe (czyli jeżeli Nn jest duże), to równanie możemy uprościć, i otrzymujemy przybliżony przedział ufności na poziomie C :

$$\hat{p} - \frac{z_* \sqrt{\hat{p}(1 - \hat{p})}}{\sqrt{Nn}} < p < \hat{p} + \frac{z_* \sqrt{\hat{p}(1 - \hat{p})}}{\sqrt{Nn}}.$$

Rozważmy obecnie zmienną X o rozkładzie Poissona $X \in Poi(\lambda)$, z nieznanym parametrem λ . Mamy

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

Niech $x_1, x_2, \dots, x_n$ będzie niezależną próbą losową z populacji o tym rozkładzie. Ustalamy

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^n x_i,$$

wtedy:

- $E(\hat{\lambda}) = \lambda$
- $Var(\hat{\lambda}) = \frac{\lambda}{n}$
- $\frac{\hat{\lambda} - \lambda}{\sqrt{\lambda/n}} \longrightarrow Z \in N(0, 1)$ gdy $n \rightarrow \infty$

W szczególności

$$P\left(-z_* < \frac{\hat{\lambda} - \lambda}{\sqrt{\lambda/n}} < z_*\right) \approx P(-z_* < Z < z_*).$$

Dla danego poziomu ufności C znajdujemy z_* takie, że

$$P\left(-z_* < Z < z_*\right) = C.$$

Wtedy przedział ufności dla nieznanego parametru λ na poziomie C ma postać $x_1 < \lambda < x_2$, gdzie x_1, x_2 są rozwiązaniami równania kwadratowego

$$x^2 - \left(2\hat{\lambda} + \frac{z_*^2}{n}\right)x + \hat{\lambda}^2 = 0.$$

Jeżeli $\frac{z_*}{N}$ jest małe (czyli n duże) otrzymujemy przybliżony przedział ufności na poziomie C :

$$\hat{\lambda} - \frac{z_* \sqrt{\hat{\lambda}}}{n} < \lambda < \hat{\lambda} + \frac{z_* \sqrt{\hat{\lambda}}}{n}.$$

Uwaga: Otrzymane powyżej przedziały ufności są przybliżone, ale niedokładność powstała z ich stosowania dąży do 0 gdy $n \rightarrow \infty$.

Rozważmy zmienną losową X o rozkładzie jednostajnym $U[0, \theta]$, gdzie parametr θ jest nieznany. Mamy

$$P(X \in I) = \frac{|I|}{\theta}.$$

Niech $x_1, x_2, \dots, x_n$ będzie niezależną próbą losową pobraną z populacji o tym rozkładzie, i niech

$$x_n^* = \max_{1 \leq k \leq n} x_k, \quad \theta^* = \frac{n+1}{n} x_n^*.$$

Wtedy:

- $E\theta^* = \theta$
- $Var(\theta^*) = \frac{\theta^2}{n(n+2)}$
- $P(x_n^* < \theta < \frac{\theta_n^*}{(1-C)^{1/n}}) = C$

Mamy więc przedział ufności na poziomie C

$$x_n^* < \theta < \frac{\theta_n^*}{(1-C)^{1/n}}).$$