

ELEMENTY TEORII ESTYMACJI

Próba statystyczna prosta (losowa)

X – zmienna losowa (cecha), która w populacji ma określony rozkład. Na przykład: X – czas dojazdu pracowników DINO.

Chcemy pobrać próbę n -elementową z populacji.

Rezerwujemy n „szufladek”, których zawartość będzie losowa. Stąd dla każdej „szufladki” mamy odrębną zmienną losową X_i o takim samym rozkładzie jaki ma badana zmienna losowa (cecha) X .

„szufladki”

„szufladka” nr 1	„szufladka” nr 2	...	„szufladka” nr n
X_1	X_2	...	X_n

Zawartość „szufladek”
po wylosowaniu z populacji

x_1	x_2	...	x_n
-------	-------	-----	-------

Def. Ciąg $\{x_1, x_2, \dots, x_n\}$ (zawartość „szufladek”) nazywamy **próbą statystyczną prostą**

dokonaną na zmiennych losowych X_1, X_2, \dots, X_n .

Statystyka

Def. Statystyką nazywamy zmienną losową Z_n , która jest funkcją zmiennych losowych X_1, X_2, \dots, X_n

$$Z_n = g(X_1, X_2, \dots, X_n)$$

Przykłady statystyk

Średnia z próby

$$(7.1) \quad \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

Wariancja z próby

$$(7.2) \quad S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

$$(7.3) \quad S_1^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Częstość (frakcja, odsetek) z próby

$$w = X/n$$

X – liczba zdarzeń sprzyjających

n – liczebność próby

Estymacja parametrów w populacji na podstawie próby

Estymacja – szacowanie wartości nieznanych parametrów w populacji na podstawie próby losowej.

Q – wartość nieznanego parametru w populacji

\hat{Q} – estymator nieznanego parametru w populacji (np. jeden ze wzorów [(7.1), (7.2), (7.3)] lub wzór na częstość)

\hat{q} – wartość liczbowa estymatora nieznanego parametru w populacji (liczba) – ocena nieznanego parametru Q

Pożądane cechy estymatora \hat{Q}

1. Nieobciążoność - $E(\hat{Q}) = Q$

2. Zgodność - $\lim_{n \rightarrow \infty} P\{|\hat{Q} - Q| < \varepsilon\} = 1 \quad \varepsilon \rightarrow 0$

3. Najwyższa efektywność - wariancja $V(\hat{Q})$ jest najmniejsza spośród wariancji dla wszystkich innych estymatorów parametru Q

4. Dostateczność - estymator \hat{Q} wykorzystuje wszystkie informacje o parametrze Q zawarte w próbie

Estymacja punktowa

Estymacja punktowa polega na szacowaniu wartości nieznanego parametru Q w populacji za pomocą estymatora \hat{Q} (wzoru).

Liczba \hat{q} uzyskana na podstawie próby za pomocą estymatora (wzoru) \hat{Q} jest oceną nieznanego parametru Q w populacji

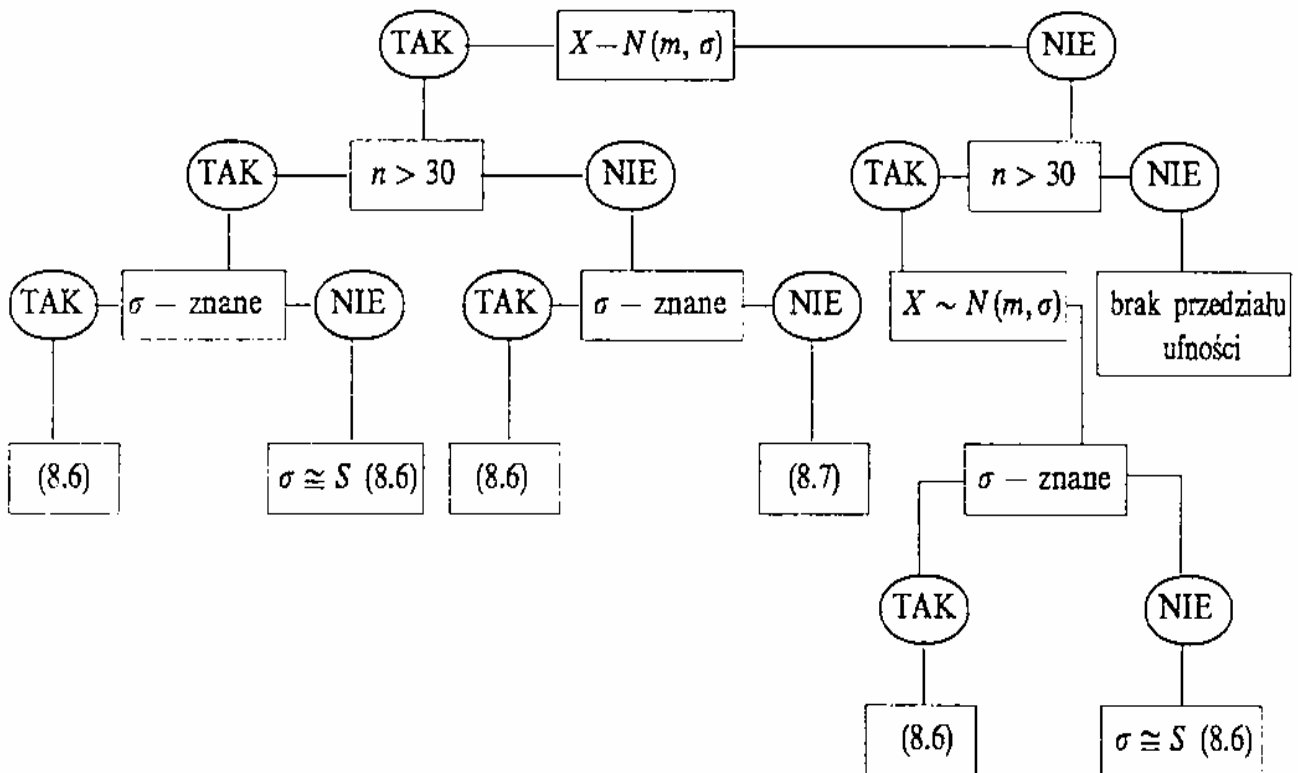
Estymacja przedziałowa

Estymacja przedziałowa polega na konstruowaniu tzw. przedziału ufności, w celu szacowania nieznanego wartości parametru Q w populacji.

Przedziałem ufności nazywamy taki przedział liczbowy, który z zadany z góry prawdopodobieństwem $(1-\alpha)$, zwanym poziomem ufności, pokrywa nieznaną wartość parametru w populacji generalnej.

Typowe wartości poziomu ufności: 0,95; rzadziej 0,90 lub 0,98; 0,99

Przedział ufności dla wartości przeciętnej m



Rys. 8.1. Schemat wyboru postaci przedziału ufności dla parametru m

$$(8.6) \quad \bar{X} - t_{\alpha} \frac{\sigma}{\sqrt{n}} < m < \bar{X} + t_{\alpha} \frac{\sigma}{\sqrt{n}}$$

Z tablic **dystrybuanty rozkładu normalnego $N(0 ; 1)$** odczytujemy

taką wartość $-t_{\alpha}$, dla której $\Phi(-t_{\alpha}) = \alpha/2$

$$(8.7) \quad \bar{X} - t_{\alpha, n-1} \frac{S}{\sqrt{n-1}} < m < \bar{X} + t_{\alpha, n-1} \frac{S}{\sqrt{n-1}}$$

Z tablic **rozkładu Studenta** odczytujemy dla $(n-1)$ stopni swobody

taką wartość $t_{\alpha, n-1}$, dla której $P\{|T_{n-1}| > t_{\alpha, n-1}\} > \alpha$.

$$\bar{X} - t_{\alpha, n-1} \frac{S_1}{\sqrt{n}} < m < \bar{X} + t_{\alpha, n-1} \frac{S_1}{\sqrt{n}}$$

(8.7a)

Wzór (8.7a) wykorzystujemy, gdy wariancję z próby S_1^2 liczymy wg wzoru (7.3).

PRZYKŁAD (8.9 – z puli do samodzielnego rozwiązania)

W 100 losowo wybranych gospodarstwach domowych średnia miesięczna opłata za energię elektryczną wyniosła 68 złotych, a odchylenie standardowe 14 złotych. Oszacuj za pomocą przedziału ufności średnie miesięczne wydatki na energię elektryczną w całej populacji (m) przyjmując poziom ufności 0,96.

Dane: $n = 100$ $\bar{x} = 68$ $S = 14$ $1 - \alpha = 0,96$

Założenie: Cecha ma w populacji rozkład normalny $N(m; \sigma)$.

Wg schematu na rys. 8.1 stosujemy wzór (8.6) przyjmując $\sigma \approx S$

Odczyt $-t_\alpha$: $\alpha = 0,04$ skąd $\alpha/2 = 0,02$

Z tablic dystrybuanty rozkładu normalnego odczytujemy wartość $-t_{0,02} = -2,05$, dla której $\Phi(-2,05) = 0,02$.

Przedział ufności wyliczymy następująco:

$$\bar{X} - t_\alpha \frac{\sigma}{\sqrt{n}} < m < \bar{X} + t_\alpha \frac{\sigma}{\sqrt{n}}$$

$$68 - 2,05 \frac{14}{\sqrt{100}} < m < 68 + 2,05 \frac{14}{\sqrt{100}}$$

$$65,1 < m < 70,9$$

INTERPRETACJA: Przedział (65,1 zł ; 70,9 zł)

z prawdopodobieństwem 0,96 (z ufnością 96%) pokrywa nieznane przeciętne wydatki na energię elektryczną w całej populacji.

PRZYKŁAD (czas dojazdu pracowników firmy DINO)

Dla 17 losowo wybranych pracowników firmy DINO otrzymano średni czas dojazdu 26 minut, a odchylenie standardowe 6 minut. Oszacuj za pomocą przedziału ufności przeciętny czas dojazdu w całej populacji pracowników DINO (m) przyjmując poziom ufności 0,95.

Dane: $n = 17$ $\bar{x} = 26$ $S = 6$ $1 - \alpha = 0,95$

Założenie: Cecha ma w populacji rozkład normalny $N(m; \sigma)$.

Wg schematu na rys. 8.1 stosujemy wzór (8.7)

Odczyt t_α : $\alpha = 0,05$. Z tablic rozkładu Studenta odczytujemy, przy $n-1=17-1=16$ stopniach swobody, wartość $t_{0,05;16} = 2,1199$.

Przedział ufności wyliczymy następująco:

$$\bar{X} - t_{\alpha, n-1} \frac{S}{\sqrt{n-1}} < m < \bar{X} + t_{\alpha, n-1} \frac{S}{\sqrt{n-1}}$$

$$26 - 2,1199 \frac{6}{\sqrt{17-1}} < m < 26 + 2,1199 \frac{6}{\sqrt{17-1}}$$

$$22,8 < m < 29,2$$

INTERPRETACJA: Przedział (22,8 minuty ; 29,2 minuty) z prawdopodobieństwem 0,95 (z ufnością 95%) pokrywa nieznaną przeciętną czas dojazdu w całej populacji pracowników DINO.

Przedział ufności dla wskaźnika struktury p (dla procentu, odsetka, frakcji)

Przedział taki konstruujemy tylko dla dużych prób ($n > 100$)

$$(8.12) \quad \frac{X}{n} - t_{\alpha} \sqrt{\frac{\frac{X}{n} \left(1 - \frac{X}{n}\right)}{n}} < p < \frac{X}{n} + t_{\alpha} \sqrt{\frac{\frac{X}{n} \left(1 - \frac{X}{n}\right)}{n}}$$

Z tablic dystribuanty rozkładu normalnego $N(0 ; 1)$ odczytujemy taką wartość $-t_{\alpha}$, dla której $\Phi(-t_{\alpha}) = \alpha/2$

PRZYKŁAD (8.7 – z puli do samodzielnego rozwiązania)

Zapytano 200 losowo wybranych przedstawicieli rodzin:
 „Kto podejmuje poważniejsze decyzje finansowe w domu?”
 W 72 przypadkach otrzymano odpowiedź, że podejmuje je małżonek.

Zbuduj przedział ufności dla odsetka rodzin (p), w których decyzje finansowe podejmuje małżonek przyjmując poziom ufności 0,99.

Dane: $n = 200$ $X = 72$ $1 - \alpha = 0,99$

Założenie: Cecha ma w populacji rozkład normalny $N(m; \sigma)$.

Odczyt $-t_\alpha$: $\alpha = 0,01$ skąd $\alpha/2 = 0,005$

Z tablic dystrybuanty rozkładu normalnego odczytujemy wartość
 $-t_{0,005} = -2,58$, dla której $\Phi(-2,58) = 0,005$.

Przedział ufności wyliczymy następująco:

$$\frac{X}{n} - t_\alpha \sqrt{\frac{\frac{X}{n} \left(1 - \frac{X}{n}\right)}{n}} < p < \frac{X}{n} + t_\alpha \sqrt{\frac{\frac{X}{n} \left(1 - \frac{X}{n}\right)}{n}}$$

$$\frac{72}{200} - 2,58 \sqrt{\frac{\frac{72}{200} \left(1 - \frac{72}{200}\right)}{200}} < p < \frac{72}{200} + 2,58 \sqrt{\frac{\frac{72}{200} \left(1 - \frac{72}{200}\right)}{200}}$$

$$0,272 < p < 0,448$$

INTERPRETACJA: Przedział (27,2% ; 44,8%)

z prawdopodobieństwem 0,99 (z ufnością 99%) pokrywa nieznaną (dla całej populacji) odsetek rodzin, w których decyzje finansowe podejmuje małżonek.