

SdI30 W07: *ESTYMACJA PRZEDZIAŁOWA PARAMETRÓW POPULACJI*

- 1. Estymacja przedziałowa**
- 2. Przykładowa konstrukcja przedziału ufności**
 - Przykład 1**
 - Przykład 2**
 - Przykład 3**
- 3. Minimalna liczebność próby**
 - Przykład 4, 5, 6**
- 4. Próby z populacji skończonych**
 - Przykład 7**
- 5. Estymacja przedziałowa parametrów w dwóch populacjach**
- 6. Zestaw zadań**

1. Estymacja przedziałowa

Estymacja przedziałowa (interval estimation) to grupa metod statystycznych służących do oszacowania parametrów rozkładu cechy w populacji generalnej. Oceną nieznanego parametru θ nie jest konkretna wartość, ale pewien przedział, który z określonym prawd. pokrywa wartość tego parametru. Pojęcie przedziału ufności wprowadził do statystyki polski matematyk Jerzy Sława-Neyman¹ w 1933 r.



¹ **Jerzy Sława-Neyman** (ur. 16 kwietnia 1894 w Benderach w Besarabii, zm. 5 sierpnia 1981 w Berkeley). W 1863 jego rodzina została deportowana do Rosji. Studiował matematykę w Charkowie. W 1921 wrócił do Polski. Od 1938 przebywał w USA, gdzie został profesorem Uniwersytetu w Berkeley.

Niech cecha X ma rozkład w populacji z nieznanym parametrem θ . Z populacji pobierana jest próba losowa X_1, X_2, \dots, X_n .

Dwustronnym przedziałem ufności (CI *confidence interval*) parametru θ nazywamy przedział (θ_1, θ_2) , którego końce są statystykami wyznaczonymi na podstawie próby losowej, tj. $\theta_i = \theta_i(X_1, X_2, \dots, X_n)$, $i = 1, 2$ oraz

$$P(\theta_1 < \theta < \theta_2) = 1 - \alpha$$

Wielkość $1 - \alpha$ nazywamy **poziomem ufności**. Im mniejsza wartość $1 - \alpha$, tym większa dokładność estymacji, ale jednocześnie tym większe ryzyko popełnienia błędu. Wybór poziomu $1 - \alpha$ jest kompromisem pomiędzy dokładnością estymacji a ryzykiem błędu. W praktyce zwykle przyjmujemy $1 - \alpha = 0,99; 0,95$ lub $0,90$.

Różnica $L_n = \theta_2 - \theta_1$ jest losową długością przedziału ufności. Im bliższy 1 poziom ufności, tym dłuższy jest przedział ufności, a tym samym mniejsza dokładność estymacji parametru.

Wybór najlepszych statystyk sprowadza się do poszukiwania przedziałów najkrótszych.

Dla próby x_1, x_2, \dots, x_n obliczone końce (q_1, q_2) są realizacją przedziału ufności. Przedział ten z prawdopodobieństwem $1 - \alpha$ pokrywa nieznaną wartość parametru θ .

2. Przedziały ufności parametrów jednej populacji

Zajmiemy się zagadnieniem estymacji przedziałowej podstawowych parametrów zmiennej losowej X , będącej modelem badanej cechy w populacji ogólnej. Parametrami tymi są: wartość oczekiwana m , wariancja σ^2 lub odchylenie standardowe σ oraz wskaźnik struktury p , w przypadku rozkładu $B(p)$.

Konstrukcje dwustronnych przedziałów ufności dla wartości oczekiwanej, wariancji i wskaźnika struktury są zestawione w tabeli 2.

Tabela 2. Dwustronne przedziały ufności dla wartości oczekiwanej, wariancji i wskaźnika struktury

L.p.	Założenia	Parametr	Końce przedziału	Oznaczenia
1	$X \sim \mathcal{N}(m, \sigma)$, σ znane, n dowolne	m	$\bar{X}_n \mp z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$	X – zm. l. będąca modelem badanej cechy w populacji, X_1, X_2, \dots, X_n – n -elementowa prosta próba losowa (SRS) pobrana z populacji, $1 - \alpha$ – poziom ufności przedziału, n – liczebność próby, $m = \mathbb{E}X$ – wartość oczekiwana, \bar{X}_n – średnia arytmetyczna z próby, σ – odchylenie standardowe populacji, S_n – odchylenie standardowe z próby (statystyka nieobciążona), p – wskaźnik struktury populacji, $K_n = \sum_{i=1}^n X_i$ – liczba elementów wyróżnionych w próbie, $\bar{P}_n = \frac{1}{n} K_n$ frakcja elementów wyróżnionych w próbie, Z_α – kwantyl rzędu α rozkładu $\mathcal{N}(0; 1)$ $t_{\alpha;v}$ – kwantyl rzędu α rozkładu t -Studenta o v stopniach swobody, $\chi_{\alpha;v}^2$ – kwantyl rzędu α rozkładu chi-kwadrat o v stopniach swobody.
2	$X \sim \mathcal{N}(m, \sigma)$, σ nieznane, n dowolne	m	$\bar{X}_n \mp t_{1-\frac{\alpha}{2}, n-1} \cdot \frac{S_n}{\sqrt{n}}$	
3	$X \sim$ dowolny, σ nieznane, $n > 30$	m	$\bar{X}_n \mp z_{1-\frac{\alpha}{2}} \cdot \frac{S_n}{\sqrt{n}}$	
4	$X \sim \mathcal{N}(m, \sigma)$, m nieznane, n dowolne	σ^2	$\left(\frac{(n-1)S_n^2}{\chi_{1-\frac{\alpha}{2}, n-1}^2}; \frac{(n-1)S_n^2}{\chi_{\frac{\alpha}{2}, n-1}^2} \right)$	
5	$X \sim \mathcal{N}(m, \sigma)$, m nieznane, $n > 30$	σ	$\left(\frac{\sqrt{\frac{n-1}{n}} S_n}{1 + \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{2n}}}; \frac{\sqrt{\frac{n-1}{n}} S_n}{1 - \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{2n}}} \right)$	
6	$X \sim B(p)$, p nieznane, $0 < \bar{p}_n \mp 3 \cdot \sqrt{\frac{\bar{p}_n(1-\bar{p}_n)}{n}} < 1$	p	$\bar{P}_n \mp z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{P}_n(1-\bar{P}_n)}{n}}$	

3. Przykładowa konstrukcja przedziału ufności

<http://www.youtube.com/watch?v=Ohz-PZqaMtk&feature=related>

<http://www.youtube.com/watch?v=A0IYTQuFgNM&feature=related>

Problem. Skonstruować przedział ufności dla wartości oczekiwanej, gdy badana cecha X w populacji ma rozkład normalny z nieznanymi parametrami, tj. $X \sim \mathcal{N}(\mu = ?, \sigma = ?)$.

Konstrukcja 1. Niech X_1, \dots, X_n będzie SRS z populacji, w której badana cecha X ma rozkład normalny z nieznanymi parametrami. Z twierdzenia Gosseta wiemy, że

$$t = \frac{\bar{X}_n - \mu}{S_n} \sqrt{n} \sim t(n - 1)$$

Niech $t_{\alpha, n-1}$ oznacza kwantyl rzędu α rozkładu t -Studenta z $n - 1$ stopniami swobody, wówczas

$$P\left(t_{\alpha/2, n-1} < \frac{\bar{X}_n - \mu}{S_n} \sqrt{n} < t_{1-\alpha/2, n-1}\right) = 1 - \alpha$$

Ponieważ

$$t_{\frac{\alpha}{2}; n-1} = -t_{1-\frac{\alpha}{2}; n-1}$$

więc po przekształceniach nierówności, otrzymamy $100(1 - \alpha)\%$ końce dwustronnego przedziału ufności dla wartości oczekiwanej

$$\theta_1 = \bar{X}_n - t_{1-\frac{\alpha}{2}, n-1} \frac{S_n}{\sqrt{n}} \quad \theta_2 = \bar{X}_n + t_{1-\frac{\alpha}{2}, n-1} \frac{S_n}{\sqrt{n}}$$

Dla jednostronnych przedziałów ufności otrzymamy:

- prawostronny przedział: (θ_1, ∞) , $\theta_1 = \bar{X}_n - t_{1-\alpha, n-1} \frac{S_n}{\sqrt{n}}$,
- lewostronny przedział: $(-\infty, \theta_2)$, $\theta_2 = \bar{X}_n + t_{1-\alpha, n-1} \frac{S_n}{\sqrt{n}}$.

Konstrukcja 2. Zakładamy, że rozkład populacji ogólnej jest normalny $\mathcal{N}(m = ?, \sigma)$, tj. odchylenie standardowe jest znane. Jako estymatora nieznanej wartości oczekiwanej m z n -elementowej próby użyjemy mediany X_M , która ma rozkład asymptotycznie normalny

$$X_M \sim \mathcal{N}\left(m, \sqrt{\frac{\pi}{2n}} \sigma\right)$$

Dla dostatecznie dużej próby ($n \geq 30$) przedział ufności parametru m ma postać:

$$\left(X_M - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\pi}{2n}} \sigma, X_M + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\pi}{2n}} \sigma \right)$$



Przykład 1. . Zbadano przebiegi 17 opon samochodowych pewnego typu wycofanych z eksploatacji i otrzymano wyniki (w tys. km): $\bar{x}_{17} = 37,25$, $s_{17} = 5,80$. Wyznaczyć przedział ufności dla odchylenia standardowego przebiegu opon tego typu na poziomie ufności $1 - \alpha = 0,95$, przyjmując założenie, że wyniki pomiarów mają rozkład normalny.

Odp.: $?(4,33; 8,83)$ [tys. km].



Przykład 2. . Trwałość pewnego typu świetlówek (liczona w godzinach świecenia) ma rozkład normalny z odchyleniem standardowym 120 godzin.

Wylosowana niezależnie z tej partii próba dała następujące wyniki trwałości: 2630, 2820, 2900, 2810, 2770, 2840, 2700, 2950, 2690, 2720, 2800, 2970.

Przyjmując poziom ufności 0,9

- a) oszacować metodą przedziałową średnią trwałość świetlówek.
- b) wyznaczyć prawostronny przedział ufności przeciętnej trwałości świetlówek.
- c) wyznaczyć prawostronny przedział ufności przeciętnej trwałości świetlówek stosując medianę z próby jako estymator nieznanej wartości oczekiwanej.

Rozwiązanie. Niech X oznacza trwałość świetlówek badanego typu. Z treści problemu wiadomo, że

$$X \sim \mathcal{N}(\mu = ?, \sigma = 120) [h], n = 12, 1 - \alpha = 0,9,$$

a) model: $\bar{x}_n \pm z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$, gdzie $z_{0,95} \overset{\text{TABLICE}}{=} 1,6449$,

$\bar{x}_n = 2800$, stąd przedział $(2743, 2857)[h]$ z prawd. 0,9 pokrywa nieznaną średnią trwałość świetlówek.

b) Model prawostronnego przedziału ufności ma postać:

$$\left(\bar{x}_n - z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}, \infty \right)$$

$z_{0,90} \overset{\text{TABLICE}}{=} 1,2816$, stąd przedział $(2755,6; \infty)[h]$ z prawdopodobieństwem 0,9 pokrywa nieznaną średnią trwałość świetlówek.



Przykład 3. . Spośród partii żarówek wyprodukowanych przez pewną fabrykę wylosowano niezależnie próbę 100 sztuk i sprawdzono ich jakość. Okazało się, że 6 żarówek było wybrakowanych. Wyznaczyć realizację 98-procentowego lewostronnego przedziału ufności frakcji braków w wyprodukowanej partii żarówek.

Rozwiązanie. Oznaczenia:

Ω – partia badanych żarówek,

$\omega_1, \omega_2, \dots, \omega_{100}$ – wylosowane do badań żarówki,

$$X_i = X_i(\omega_i) = \begin{cases} 1, & \text{jeśli } i \text{ – ta żarówka jest brakiem,} \\ 0, & \text{w p. p.} \end{cases}$$

zmienne losowe X_1, X_2, \dots, X_{100} są wynikiem kontroli jakości wylosowanych żarówek. Badana cecha X ma rozkład Bernoulliego z nieznaną frakcją braków p , tj. $X \sim B(p = ?)$

Jeżeli spełnione jest założenie

$$0 < \bar{p}_n \mp 3 \sqrt{\frac{\bar{p}_n(1-\bar{p}_n)}{n}} < 1,$$

to lewostronny przedział ufności dla parametru p jest postaci $(0, \theta_2)$, gdzie θ_2 wyznaczane jest z modelu

$$\theta_2 = \bar{P}_n + z_{1-\alpha} \cdot \sqrt{\frac{\bar{P}_n(1-\bar{P}_n)}{n}}$$

Obliczenia pomocnicze i sprawdzenie założenia

$$\bar{p}_{100} = \frac{6}{100}, 1 - \alpha = 0,98, z_{0,98} \stackrel{\text{TABLICE}}{=} 2,0537$$

Ponieważ $3\sqrt{\frac{\bar{p}_n(1-\bar{p}_n)}{n}} = 3\sqrt{\frac{\frac{6}{100}\frac{94}{100}}{100}} \approx 0,071246$, więc założenie nie jest spełnione i $\theta_2 = \frac{6}{100} + 2,0537 \cdot \sqrt{\frac{\frac{6}{100}\frac{94}{100}}{100}} = 0,10877$.

Wniosek. 98-procentową lewostronną realizacją przedziału ufności frakcji braków żarówek jest przedział $(0; 0,10877)$.

<http://www.youtube.com/watch?v=bekNKJoxYbQ&feature=related>

<http://www.youtube.com/watch?v=iX0bKAeLbDo&feature=watch-vrec>

3. Minimalna liczebność próby

Maksymalny błąd estymacji to połowa przedziału ufności

$$\Delta = \frac{\theta_2 - \theta_1}{2}$$

Ustalamy minimalną liczebność próby zapewniającą, przy danym poziomie ufności $1 - \alpha$, nieprzekroczenie przez maksymalny błąd szacunku z góry założonej wielkości d :

- przy estymacji wartości oczekiwanej w populacji normalnej ze znaną oraz nieznaną wariancją

w przypadku znanej wariancji $n = \left\lceil \frac{z_{1-\frac{\alpha}{2}}^2 \sigma^2}{d^2} \right\rceil$,

w przypadku nieznanej wariancji $n = \left\lceil \frac{t_{1-\frac{\alpha}{2}; n_0-1}^2 S_{n_0}^2}{d^2} \right\rceil$,

gdzie n_0 jest liczebnością próby pilotującej, a $S_{n_0}^2$ jest wariancją z próby pilotującej,

- przy estymacji wskaźnika struktury p w rozkładzie $B(p)$:
 - a) jeśli znane jest p_0 , tj. spodziewany rząd wielkości p , to

$$n = \left\lceil \frac{z_{1-\frac{\alpha}{2}}^2 p_0(1-p_0)}{d^2} \right\rceil,$$

- b) jeśli nie znany jest rząd wielkości p , to $n = \left\lceil \frac{z_{1-\frac{\alpha}{2}}^2}{4d^2} \right\rceil$.



Przykład 4 . W celu oszacowania przeciętnej trwałości baterii (liczonej czasem jej użytkowania w godzinach) zmierzono trwałości 9-ciu losowo wybranych baterii z dużej ich partii. Otrzymano następujące dane: 212, 215, 205, 214, 216, 208, 210, 215, 220. Zakładając, że trwałość baterii ma rozkład normalny wyznaczyć minimalną wielkość próby potrzebną do oszacowania przeciętnej trwałości baterii z dopuszczalnym błędem maksymalnym $2[h]$. Przyjąć poziom ufności 0,95.

Rozwiązanie. Niech zmienna losowa X oznacza trwałość baterii w badanej ich partii. Ponieważ wariancja trwałości baterii nie jest znana, więc 9-elementowa próba jest badaniem pilotażowym.

Dane: $n_0 = 9$,

- maksymalny błąd szacunku $d = 2$,
- poziom ufności $1 - \alpha = 0,95$, stąd $\frac{\alpha}{2} = 0,025$.

Ponieważ wariancja trwałości w całej partii baterii nie jest znana, więc wielkość próby wyznaczana jest ze wzoru:

$$n = \left\lceil \frac{t_{1-\frac{\alpha}{2}; n_0-1}^2 s^2}{d^2} \right\rceil$$

$$t_{0,975; 8} \stackrel{\text{EXCEL}}{=} \text{ROZKŁ. T. ODWR}(0,975; 8) \approx 2,306,$$

$$s_9^2 \stackrel{\text{EXCEL}}{=} \text{WARIANCJA. PRÓBK}(A1:A9) = 20,69(4)$$

$$n = \left\lceil \frac{(2,306)^2 \cdot 20,69(4)}{4} \right\rceil = [27,5115] = 28.$$

Odp.: Do oszacowania trwałości baterii niezbędna jest próba o liczebności 28.



Przykład 5 . Waga netto oliwek pakowanych do pewnego typu słoików wynosi $600g \pm 10g$.

Wyznaczyć liczbę słoików, które należy pobrać do próby dla oszacowania rzeczywistej zawartości oliwek, z błędem nieprzekraczającym $5g$. Przyjąć poziom ufności 0,98.

Rozwiązanie. Niech X oznacza wagę netto oliwek. Przyjmujemy, że $X \sim \mathcal{N}(m = ?, \sigma = 10)[g]$, ponadto wiadomo, że $d = 5$ i $1 - \frac{\alpha}{2} = 0,99$, więc liczebność ustalimy ze wzoru:

$$n = \left\lceil \frac{z_{1-\frac{\alpha}{2}}^2 \sigma^2}{d^2} \right\rceil$$

$$z_{0,99} \stackrel{\text{EXCEL}}{=} \text{ROZKŁ. NORMALNY.S. ODWR}(0,99) = 2,3263$$

Stąd

$$n = \left\lceil \frac{(2,3263)^2 \cdot 100}{25} \right\rceil = [21,65] = 22$$

Odp.: Do oszacowania rzeczywistej zawartości oliwek należy sprawdzić wagę oliwek w 22 słoikach.



Przykład 6 . Agencja marketingowa zamierza oszacować procent firm w Polsce korzystających z pewnego oprogramowania wspomagającego zarządzanie. Pewne fragmentaryczne dane wskazują, że około 30% firm korzysta z takiego oprogramowania. Określić niezbędną wielkość próby przy maksymalnym dopuszczalnym błędzie szacunku 5% oraz poziomie ufności 0,9.

Rozwiązanie. Niech Ω oznacza populację firm w Polsce oraz

$$X(\omega) = \begin{cases} 1, & \text{jeśli firma } \omega \text{ korzysta z oprogramowania,} \\ 0, & \text{w przeciwnym przypadku.} \end{cases}$$

$X \sim B(p)$, gdzie nieznana frakcja p firm korzystających z oprogramowania jest przedmiotem badania.

Ponieważ z fragmentarycznych badań wiadomo, że $p_0 = 30\% = 0,3$, więc do oszacowania niezbędnej liczby firm należy zastosować wzór:

$$n = \left\lceil \frac{z_{1-\frac{\alpha}{2}}^2 \cdot p_0(1-p_0)}{d^2} \right\rceil$$

gdzie $1 - \frac{\alpha}{2} = 0,95$, $z_{0,95} \stackrel{\text{TABLICE}}{=} 1,6449$, $d = 5\% = 0,05$.

Stąd

$$n = \left\lceil \frac{(1,6449)^2 \cdot 0,3 \cdot 0,7}{(0,05)^2} \right\rceil = [227,278] = 228.$$

Odp.: W celu oszacowania wskaźnika firm w Polsce korzystających z pewnego oprogramowania wspomagającego zarządzanie należy z populacji Ω wylosować do badań 228 firm.

4. Próby z populacji skończonych

W CTG zakładamy, że próby pobierane są ze zwracaniem i/lub populacja jest nieskończona.

W praktyce często dysponujemy skończoną populacją wielkości N a próba pobierana jest bez zwracania.

W takiej sytuacji, jeśli tylko wielkość próby n nie jest zbyt mała w porównaniu z całkowitą liczebnością populacji ($n \geq 5\%N$), stosujemy poprawkę na skończoną populację (ang. *finite population correction fpc*), w celu obliczenia odchylenia standardowego i proporcji. Ma ona postać:

$$fpc = \sqrt{\frac{N-n}{N-1}}$$

i jest zawsze mniejsza od 1.

Zastosowanie tej poprawki prowadzi do zmiany postaci przedziałów ufności dla średniej oraz proporcji. Mają one w takim przypadku postać:

$$\bar{X}_n \pm z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

$$\bar{X}_n \pm t_{1-\frac{\alpha}{2}, n-1} \cdot \frac{S_n}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

$$\bar{P}_n \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{P}_n(1-\bar{P}_n)}{n}} \sqrt{\frac{N-n}{N-1}}$$

stosując *fpc*, w przypadku przedziałów ufności uzyskujemy przedziały krótsze o tym samym jednak poziomie ufności.

Wprowadzenie *fpc* pociąga za sobą również konieczność zmiany minimalnej wielkości próby, która wynosi teraz:

$$n = \frac{n_0 N}{n_0 + N - 1}$$

gdzie n_0 oznacza minimalną liczebność uzyskaną bez zastosowania poprawki.

Przykład 7 (TG s.205). Stany Zjednoczone są podzielone na 3078 dystryktów. Losowo wybrano bez powtórzeń 300 z nich i obliczono, ile akrów przeznaczono pod uprawę. Uzyskano średnią 297,9tys. akrów z odchyleniem standardowym 344,6.

- a) Ile wynosi błąd standardowy średniej?
- b) Wyznaczyć 95% przedział ufności dla średniej.

- c) Obliczyć prawdopodobieństwo wylosowania próby o średniej poniżej 300 tys. akrów.
- d) Jak dużą próbę należy pobrać, aby uzyskać błąd nie większy niż 20 tys. akrów?

Rozwiązanie. Ponieważ populacja jest skończona, losowanie odbywa się bez powtórzeń oraz

$$\frac{n}{N} = \frac{300}{3078} = 9,75\% > 5\%$$

więc należy zastosować poprawkę fpc .

Mamy

$$fpc = \sqrt{\frac{3078 - 300}{3078 - 1}} = 0,95$$

a) Błąd standardowy średniej

$$s_{\bar{x}} = \frac{344,6}{\sqrt{300}} fpc = 18,94$$

b) 95% przedział ufności dla średniej

$$\begin{aligned} \mu &\in (297,9 - t_{0,975;299} \cdot 18,94, 297,9 + t_{0,975;299} \cdot 18,94) \\ &= (260,64; 335,16) \end{aligned}$$

Dla porównania wyznaczmy przedział ufności bez poprawki fpc

$$\begin{aligned} \mu &\in \left(297,9 - t_{0,975;299} \cdot \frac{344,6}{\sqrt{299}}, 297,9 + t_{0,975;299} \cdot \frac{344,6}{\sqrt{299}} \right) \\ &= (258,68; 337,12) \end{aligned}$$

c) Na mocy CTG $\bar{X}_{300} \approx \sim \mathcal{N}(297900; 344,6)$

$$P(\bar{X}_{300} < 300000) =$$

d) Obliczymy jak dużą próbę należałoby pobrać bez stosowania poprawki fpc , aby błąd nie przekroczył 20 tys. akrów.

$$n_0 = \frac{t_{0,975;299}^2 (344,6)^2}{20^2} = 1149,71$$

Tak uzyskaną liczebność (1150) możemy wykorzystać w celu wyliczenia liczebności próby z zastosowaniem poprawki:

$$n = \frac{1150 \cdot 3078}{1150 + 3078 - 1} = 837,40$$

Jak widać użycie poprawki prowadzi do znacznej redukcji wielkości próby.

..... \mathcal{R}

fpc = **sqrt**((3078-300)/(3078-1))

sx = 344.6/**sqrt**(299)*fpc

c(297.9-**qt**(0.975, 299)*sx, 297.9+**qt**(0.975, 299)*sx)

c(297.9-**qt**(0.975, 299)*344.6/**sqrt**(299), 297.9+**qt**(0.975, 299)* 344.6/**sqrt**(299))

n0 = **qt**(0.975, 299)^2*344.6^2/20^2

(n0*3078)/(n0+3078-1)

.....

5. Estymacja przedziałowa parametrów w dwóch populacjach

Przykład 8. (KA przykład 5.12). Dokonano losowej obserwacji dochodu osób fizycznych z dwóch miast.

Na podstawie otrzymanych wyników

$$M1 = [1189, 840, 1020, 980],$$

$$M2 = [853, 900, 733, 785]$$

wyznaczyć 95-procentową realizację przedziału ufności dla różnicy wartości oczekiwanych dochodów osób fizycznych z badanych miast.

Odp.: Jeżeli wariancje są różne, to 95-procentową realizacją przedziału ufności dla różnicy wartości oczekiwanych dochodów w badanych miastach jest przedział

$$(-25,68; 404,7)$$

Jeżeli wariancje dochodów w obydwu miastach są równe, to można z 95-procentową ufnością twierdzić, że przedział

$$(-8,1; 387,1)$$

pokrywa nieznana różnicę wartości oczekiwanych dochodów w badanych miastach.

Ponieważ wyznaczone przedziały w obydwu przypadkach zawierają 0, więc dla obydwu przypadków można wnioskować, że nieznane wartości oczekiwane mogą być sobie równe.

W tablicy 3 zestawione są modele przedziałów ufności dla różnicy wartości oczekiwanych, ilorazu wariancji oraz różnicy wskaźników struktury.

Tablica 3. Estymacja przedziałowa parametrów w dwóch populacjach

L.p.	Parametr	Założenia	Końce przedziału ufności	Oznaczenia
1	$m_1 - m_2$	$X_i \sim \mathcal{N}(m_k, \sigma_k), k = 1, 2$ $\sigma_k = ?$ ale $\sigma_1 = \sigma_2$ n_k dowolne	$(\bar{X}_1 - \bar{X}_2) \mp t_{1-\frac{\alpha}{2}, n_1+n_2-2} \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$	X_1, X_2 – zm. l-owe będące modelami badanej cechy w dwóch populacjach, $\mathbf{X}_k = (X_{k1}, X_{k2}, \dots, X_{kn_k})$ – n_k -elementowe niezależne proste próby losowe, $k = 1, 2$; $1 - \alpha$ – poziom ufności przedziału, $k = 1, 2$ – numer populacji lub próby, n_k – liczebność k -tej próby, m_k – wartość oczekiwana k -tej populacji, \bar{X}_k – średnia arytmetyczna k -tej próby, $\overline{X_1 - X_2}$ – średnia arytmetyczna z różnic dla prób związanych w pary, σ_k – odch. std. dla k -tej populacji, S_k – odch. std. dla k -tej próby losowej (statystyka nieobciążona), $S_{\bar{X}_1 - \bar{X}_2}$ – odch. std. dla różnic z prób powiązanych, p_k – wskaźnik struktury dla k -tej populacji, K_k – liczba elementów wyróżnionych w k -
2	$m_1 - m_2$	$X_k \sim \mathcal{N}(m_k, \sigma_k), k = 1, 2$ $\sigma_k = ?$ ale $\sigma_1 \neq \sigma_2$ n_k dowolne	$(\bar{X}_1 - \bar{X}_2) \mp t_{1-\frac{\alpha}{2}, \nu} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$	
3	$m_1 - m_2$	$X_k \sim ?$ lub dowolny, $k = 1, 2$ $\sigma_k = ?$, $n_k > 30$	$(\bar{X}_1 - \bar{X}_2) \mp z_{1-\frac{\alpha}{2}} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$	
4	$m_1 - m_2$	populacja par, $(X_1 - X_2) \sim \mathcal{N}(m, \sigma)$, $\sigma = ?$, $n_1 = n_2 = n$	$\overline{X_1 - X_2} \mp t_{1-\frac{\alpha}{2}, n-1} \frac{S_{\bar{X}_1 - \bar{X}_2}}{\sqrt{n}}$	
5	$\frac{\sigma_1^2}{\sigma_2^2}$	$X_k \sim \mathcal{N}(m_k, \sigma_k), k = 1, 2$ n_k dowolne	$\frac{S_1^2}{S_2^2} F_{\frac{\alpha}{2}, n_1-1, n_2-1}; \frac{S_1^2}{S_2^2} F_{1-\frac{\alpha}{2}, n_1-1, n_2-1}$	

6	$p_1 - p_2$	$X_k \sim B(p_k), k = 1, 2$ $\bar{p}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} X_{ki}$ $\bar{p}_k \mp 3 \cdot \sqrt{\frac{\bar{p}_k(1 - \bar{p}_k)}{n_k}}$ $\subset (0, 1)$	$(\bar{p}_1 - \bar{p}_2) \mp z_{1-\frac{\alpha}{2}} \sqrt{\frac{\bar{p}_1(1 - \bar{p}_1)}{n_1} + \frac{\bar{p}_2(1 - \bar{p}_2)}{n_2}}$	tej próbie, $\bar{p}_k = \frac{K_k}{n_k}$ – frakcja wyróżnionych elementów w k -tej próbie, z_α – kwantyl rzędu α rozkładu $\mathcal{N}(0; 1)$, $t_{\alpha;v}$ – kwantyl rzędu α rozkładu t -Studenta z v stopniami swobody, $\chi_{\alpha;v}^2$ – kwantyl rzędu α rozkładu chi- kwadrat z v stopniami swobody. $F_{\alpha;n_1-1,n_2-1}$ – kwantyl rzędu α rozkładu F z $n_1 - 1$ i $n_2 - 1$ stopniami swobody
---	-------------	---	---	--

Na podstawie niezależnych prób prostych X_1, \dots, X_n i Y_1, \dots, Y_n pochodzących z populacji X oraz Y można zbudować przedział ufności dla nieznanego ilorazu wariancji $\frac{\sigma_1^2}{\sigma_2^2}$.

Populacje, z których pochodzą próby powinny mieć rozkłady zbliżone do normalnego, zwłaszcza dla małych prób.

Przykład 9 (KA przykład 5.14). Na podstawie danych z przykładu 8. dotyczących dochodu osób fizycznych, wyznaczyć 95-procentowe przedziały ufności dla ilorazów wariancji dochodów.

Odp.: Z przeprowadzonych obliczeń wynika, że 95-procentowe realizacje przedziałów ufności dla ilorazów wariancji $\frac{\sigma_1^2}{\sigma_2^2}$ i $\frac{\sigma_2^2}{\sigma_1^2}$ wynoszą:

(0,246574; 58,7755) i (0,0170139; 4,05557).

Porównanie wskaźników struktury

Niech p_1 oraz p_2 będą nieznanymi wskaźnikami wyróżnionych elementów populacji o rozkładach $X \sim B(p_1)$, $Y \sim B(p_2)$, odpowiednio.

Na podstawie niezależnych prób prostych X_1, \dots, X_n i Y_1, \dots, Y_n pochodzących z populacji X oraz Y można zbudować przedział ufności dla różnicy wskaźników struktury.

Przykład 10 (KA 5.16).

Na 1200 kobiet oraz 900 mężczyzn, wylosowanych niezależnie spośród dorosłych mieszkańców dużego miasta, 400 kobiet i 220 mężczyzn zna co najmniej jeden język obcy. Wyznaczyć 95-procentowy przedział ufności dla różnicy wskaźników dla kobiet i mężczyzn znających co najmniej jeden język obcy w badanym mieście. Jaki wynika stąd wniosek?

Odp.: 95-procentową realizacją przedziału ufności dla różnicy wskaźników dla kobiet i mężczyzn znających co najmniej jeden język obcy jest przedział $(-0,239; 0,417)$. Ponieważ $0 \in (-0,239; 0,417)$, więc w badanym mieście wskaźniki osób znających co najmniej jeden język obcy dla kobiet i mężczyzn mogą być sobie równe.

6. Zestaw zadań W07

Niezbędne tablice statystyczne

1. Modele przedziałów ufności dla wartości oczekiwanej, wariancji i wskaźnika struktury dla jednej populacji.
2. Modele przedziałów ufności dla parametrów w dwóch populacjach.

Tablice rozkładów podstawowych statystyk:

<http://www.statsoft.com/textbook/sttable.html#chi>

ZADANIA

1. Wyprowadzić wzory na przedział ufności dla
 - a) wartości oczekiwanej,
 - b) wariancji,cechy o rozkładzie normalnym z nieznanymi parametrami.

2. Korzystając z dostępnego oprogramowania wybrać rozkład i wygenerować małą oraz dużą próbę i na ich podstawie dokonać estymacji punktowej przedziałowej parametrów.
3. Rozkład wyników pomiarów głębokości morza w pewnym rejonie jest normalny. Dokonano 5 niezależnych pomiarów głębokości morza w tym rejonie i otrzymano następujące wyniki (w $[m]$): 871, 862, 870, 876, 866. Na poziomie ufności 0,90 wyznaczyć CI dla wartości oczekiwanej oraz dla wariancji głębokości morza w badanym rejonie.
4. Pośrednik w handlu nieruchomościami chce oszacować przeciętną wartość kawalerki w pewnej dzielnicy. W losowej próbie 16 kawalerek średnia wyniosła 120 000 PLN. Odchylenie standardowe wartości kawalerek

- a) jest znane pośrednikowi i wynosi 5500PLN;
- b) nie jest znane pośrednikowi i obliczone z próby odchylenie standardowe wynosi 5500PLN,
- c) wygenerować 16 elementową próbę według rozkładu $\mathcal{N}(120000; 5500)$.

Wyznaczyć oraz porównać 95% i 99% przedziały ufności dla przeciętnej wartości kawalerki w rozważanej dzielnicy.

5. Linia lotnicza chce oszacować frakcję Polaków, którzy będą korzystać z nowo otwartego połączenia między Poznaniem a Londynem. Wybrano losową próbę 347 pasażerów korzystających z tego połączenia, z których 201 okazało się Polakami.

a) Wyznaczyć 90% przedział ufności dla frakcji Polaków wśród pasażerów korzystających z nowo otwartego połączenia. **Odp.:** (0,536; 0,623).

b) Wygenerować 347 elementową próbę według rozkładu $B(0,58)$ identyfikującą polskich pasażerów i na tej podstawie wyznaczyć 90% przedział ufności.

6. Frekwencja widzów na seansie filmowym w jednym z kin ma rozkład $\mathcal{N}(\mu = ? ; \sigma = 30)$. Na podstawie rejestru liczby widzów na 25 losowo wybranych seansach filmowych oszacowano przedział liczbowy (184; 216) dla nieznannej przeciętnej frekwencji na wszystkich seansach.

- a) Obliczyć średnią liczbę widzów w badanej próbie.
- b) Jaki poziom ufności przyjęto przy estymacji?

7. Wzrost losowo wybranej osoby z pewnej populacji ma rozkład normalny o nieznanym parametrach. Pobrano próbę losową o liczności $n = 26$ i po obliczeniu przedziału ufności na poziomie 0,9 otrzymano następujący wynik: (162; 178). Obliczyć średni wzrost i wariancję wzrostu w pobranej próbie.


Odp.: $\bar{X} = 170, S_n^2 \approx 570,4$.

8. Ustalić tak liczebność próby, aby na poziomie ufności 0,99 można było oszacować oczekiwany czas zdatności akumulatorów z dokładnością do i) 20[h]; ii) 10[h], jeśli odchylenie standardowe w populacji jest

- a) znane i wynosi $\sigma = 40$ [h];
- b) nieznanie i wyznaczone z n_0 -elementowej próby wstępnej wynosi $s = 40$ [h].

9. Wykonujemy pomiary głębokości morza w pewnym określonym miejscu. Ile niezależnych pomiarów głębokości należy wykonać w tym miejscu, aby przyjmując poziom ufności $1 - \alpha = 0,95$ wyznaczyć głębokość z błędem mniejszym niż 5[m] zakładając, że rozkład błędów jest rozkładem normalnym $\mathcal{N}(0, \sqrt{180})$ [m].

Odp.: 28.

10  Na ilu potencjalnych klientach należy przeprowadzić ankietę, aby oszacować odsetek osób mających zamiar zakupić nowy samochód w ciągu najbliższych 2 lat? Przyjmując poziom ufności 0,95 oraz maksymalny dopuszczalny błąd szacunku 6%.

Odp.: 267

11. Odtworzyć przykład 5.9 (KA s. 176)

12. Rozwiązać KA zadanie 16 s.189.

13. Rozwiązać KA zadanie 17 s.189.

14. Czas obsługi w okienku bankowym nie powinien mieć dużej wariancji, gdyż w przeciwnym przypadku kolejki mają tendencję do rozrastania się. Bank regularnie sprawdza czas obsługi w okienkach, by oceniać jego wariancję. Obserwacja 22 czasów obsługi losowo wybranych klientów dała wariancję równą 8 minut². Wyznacz 95% i 99% przedział ufności dla wariancji czasu obsługi w okienku bankowym.

Odp.: 95% (4,74; 16,34).

15. (Studium przypadku). Z partii kondensatorów wybrano losowo 12 kondensatorów i zmierzono ich pojemności,

otrzymując wyniki (w pF): 4,45, 4,40, 4,42, 4,38, 4,44, 4,36, 4,40, 4,39, 4,45, 4,35, 4,40, 4,35.

- a) Znaleźć ocenę wartości oczekiwanej \bar{x}_{12} i wariancji s_{12}^2 pojemności kondensatora pochodzącego z danej partii.
- b) Wygenerować 100 elementową próbę według rozkładu $\mathcal{N}(\bar{x}_{12}, s_{12})$.
- c) Znaleźć ocenę wskaźnika kondensatorów, które nie spełniają wymagań technicznych, przyjmując, że kondensator nie spełnia tych wymagań, gdy jego pojemność jest mniejsza od 4,39 pF .
- d) Znaleźć ocenę wariancji pojemności kondensatorów.
- e) Wyznaczyć 90-procentową ocenę przedziału ufności dla wartości oczekiwanej pojemności kondensatora pochodzącego z danej partii.

f) Wyznaczyć 90-procentową realizację przedziału ufności dla wskaźnika kondensatorów, które nie spełniają wymagań technicznych w badanej partii.

16. Dla wylosowanej próby studentów otrzymano następujący rozkład tygodniowego czasu nauki (w godzinach $[h]$):

Czas nauki	$[0, 2)$	$[2, 4)$	$[4, 6)$	$[6, 8)$	$[8, 10)$	$[10, 12)$
Liczba studentów	10	28	42	30	15	7

- a) Oszacować metodą punktową średni czas poświęcony tygodniowo na naukę oraz wariancję tego czasu.
- b) Przyjmując poziom ufności 0,90 oszacować metodą przedziałową średni tygodniowy czas nauki oraz wariancję tego czasu. **Odp.:** $\bar{x} = 5,5 [h]$, $s = 2,54[h]$, $m \in (5,14; 5,86)[h]$.

17. A random sample of 64 observation from a population produced the following summary statistics:

$$\sum x_i = 500, \sum (x_i - \bar{x})^2 = 3,566.$$

- a) Find 95% confidence interval for m .
- b) Interpret the confidence interval you found in part (a).

18. A random sample of size $n = 400$ yielded $\bar{p}_n = 0,42$.

- (a) Is the sample size large enough to use the methods of this section to construct a confidence interval for p ? Explain.

- (b) Construct a 95% confidence interval for p .

Answer: (a) Yes, $p_0 \pm 3\mathbb{D}\bar{P}_n$ lies in $(0, 1)$; (b) $(0,372; 0,468)$.

19. Jak liczna powinna być próba, aby na jej podstawie można było z prawd. 0,99 oszacować średni wzrost noworodków

przy maksymalnym błędzie szacunku 1cm? Zakładamy, że rozkład wzrostu noworodków jest rozkładem normalnym z odchyleniem standardowym 2,5cm.