

Curs 12

Intervale de încredere

Am văzut cum poate fi estimat un parametru folosind datele furnizate de un eșantion. Parametrul din populație nu este, în general, egal cu statistica calculată cu ajutorul eșantionului. Ne punem problema cât este de bună această estimare, adică vom calcula așa numita marjă de eroare.

Presupunem că studiem vâscozitatea unei anumite substanțe. Prin studierea unui eșantion s-a constatat că media acestei caracteristici este $\hat{\mu} = \bar{x} = 1000$. Dacă considerăm un alt eșantion este aproape imposibil să obținem aceeași estimare numerică pentru media vâscozității. Nu putem spune nimic despre relația dintre cele două medii. Problema pe care o punem este următoarea: valoarea reală a vâscozității este cuprinsă între 900 și 1100 sau între 990 și 1100? Răspunsul la această întrebare afectează deciziile ulterioare legate de acest proces. Marginile unui interval plauzibil pentru valorile mediei constituie un interval estimat.

Acest interval unde bănuim că este situată valoarea reală a parametrului populației studiate se numește *interval de încredere*. Intervalul de încredere constă din:

- *un interval*, obținut cu ajutorul datelor furnizate de o selecție,
- *un nivel de încredere*, care reprezintă probabilitatea ca intervalul să acopere valoarea reală a parametrului.

Nivelul de încredere se precizează. De regulă se consideră 0.90 sau mai mult. Se dă de obicei α , unde nivelul de încredere este $1 - \alpha$ (0.95 corespunde pragului de semnificație $\alpha = 0.05$).

Definiția 12.0.1 *Se numește interval de încredere pentru un parametru θ asociat unei populații orice interval $I = [a, b]$ pentru care se poate estima probabilitatea ca $\theta \in I$. Dacă α este un număr cuprins între 0 și 1 și dacă $P(\theta \in I) \geq 1 - \alpha$, se spune că I este un interval de încredere pentru θ cu un nivel de încredere $1 - \alpha$ (sau echivalent, cu un nivel de încredere $(1 - \alpha) 100\%$ sau cu eroare sub $\alpha 100\%$).*

În cele ce urmează vom construi intervale de încredere numai pentru caracteristici care urmează o distribuție normală

12.1 Intervale de încredere pentru medie în cazul σ cunoscut

Presupunem că realizăm o selecție populație a cărei caracteristică studiată urmează o distribuție normală, $N[m, \sigma]$, cu σ cunoscut, m necunoscut. Situația este mai puțin întâlnită în realitate deoarece în mod normal atât media cât și dispersia sunt necunoscute. Totuși vom prezenta în continuare și acest caz.

12.1.1 Construcția intervalului de încredere

Fie x_1, x_2, \dots, x_n valorile variabilelor de selecție X_1, X_2, \dots, X_n obținute dintr-o populație care urmează o distribuție normală, $N[m, \sigma]$, $\sigma > 0$ cunoscut, m necunoscut. Știm că $Z = \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} \in N([0, 1])$. Din această cauză putem scrie, (evident $z > 0$),

$$P(|Z| \leq z) = 1 - \alpha \Leftrightarrow 1 - \alpha = P\left(m \in \left[\bar{X} - z \frac{\sigma}{\sqrt{n}}, \bar{X} + z \frac{\sigma}{\sqrt{n}}\right]\right) \Leftrightarrow \\ 1 - \alpha = \Phi(z) - \Phi(-z) \Leftrightarrow 1 - \alpha = 1 - 2\Phi(-z) \Leftrightarrow \Phi(-z) = \frac{\alpha}{2}$$

Notăm cu $z_{\frac{\alpha}{2}}$ valoarea (pozitivă) a lui z obținută din relația $\Phi(-z) = \frac{\alpha}{2}$. Pentru determinarea acestei valori se folosește tabelul pentru funcția lui Laplace (a se vedea Anexa 1) sau programele Matlab sau Mathematica.

De îndată ce selecția a fost realizată și a fost calculată media de selecție $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ se obține intervalul,

$$\left[\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right] \quad (12.1)$$

Suntem tentați să spunem că $1 - \alpha$ este probabilitatea ca acest interval să cuprindă valoarea exactă a lui m , dar această afirmație nu este corectă. Trebuie să ținem seama de faptul că intervalul de încredere este un interval aleator, el depinde de selecția făcută, deci extremitățile sale sunt v. a. Prin urmare interpretarea corectă a lui $1 - \alpha$ este următoarea: dacă, facem un număr foarte mare de selecții și calculăm de fiecare dată intervalul de încredere cu nivelul de încredere $1 - \alpha$, atunci $(1 - \alpha) 100\%$ din aceste intervale vor conține valoarea exactă pentru m .

Observăm că intervalul de încredere pentru m este centrat în estimăția punctuală \bar{x} . Când n crește se obține un *interval mai scurt* pentru același coeficient de încredere. Un interval de încredere mai scurt indică o mai mare încredere în \bar{x} ca estimăție a lui m .

Exemplul 12.1.1 Punctajele obținute de studenți care au promovat examenul de matematică și care cuantifică cunoștințele lor sunt:

{64, 62, 76, 82, 66, 76, 72, 71, 74, 72, 71, 73, 70, 75, 77, 84, 92, 86, 62, 58, 78, 80, 79, 84, 83, 82, 66, 68, 68, 82, 84, 78, 76, 69, 77, 58, 62, 82, 85, 58, 78, 84, 94, 88, 77, 78, 88, 91, 70, 71, 78, 58, 65, 53, 60, 49, 68, 74, 71, 66, 68, 71, 73, 70, 85, 78, 65, 54, 51, 78, 89, 66, 68, 95, 94, 99, 81, 81, 92, 88, 99, 81, 81}.

Se presupune că se cunoaște $\sigma = 10.99$. Să se construiască intervalele de încredere pentru medie cu nivelele de încredere de 90%, 95% și 99%.

Rezolvare. Am calculat $\bar{x} = \frac{1}{83} \sum_{i=1}^{83} x_i = 75.0602$.

Calculăm intervalele de încredere cu nivelul de încredere de 90%, 95% și 99%.

Pentru 90% avem $\alpha = 0.1$ și

$$\Phi(-z) = 0.05 \Rightarrow -z = -1.6449 \Rightarrow z_{\frac{\alpha}{2}} = 1.6449.$$

Atunci, conform (12.1), intervalul

$$I = \left[\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right] = [73.0760; 77.0445].$$

este un interval de încredere pentru m cu 90% nivel de încredere.

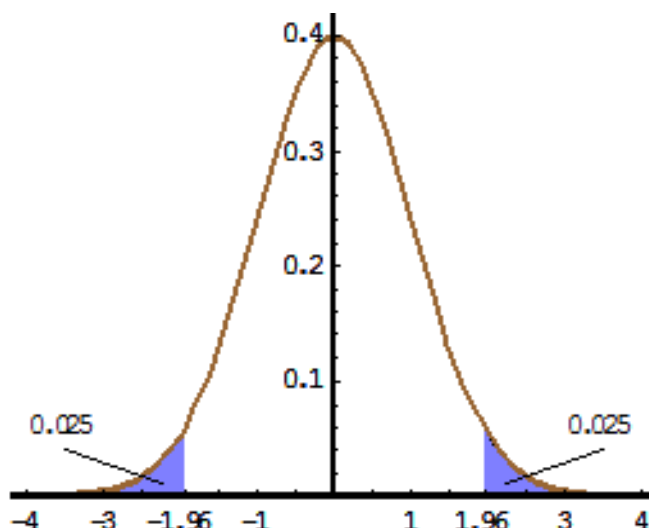
Pentru 95% avem $\alpha = 0.05$ și $\Phi(-z) = 0.025 \Rightarrow z_{\frac{\alpha}{2}} = 1.9599$. Intervalul de încredere pentru m este

$$I = \left[\bar{x} - 1.9599 \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.9599 \cdot \frac{\sigma}{\sqrt{n}} \right] = [72.6960; 77.4245].$$

Pentru 99% avem $\alpha = 0.01$ și $\Phi(-z) = 0.005 \Rightarrow z_{\frac{\alpha}{2}} = 2.5758$, intervalul de încredere pentru m este

$$I = \left[\bar{x} - 2.5758 \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + 2.5758 \cdot \frac{\sigma}{\sqrt{n}} \right] = [71.9530; 78.1675].$$

Observăm că dacă, de exemplu, nivelul de încredere este 0.95, atunci $z_{\frac{\alpha}{2}} = 1.9599$ trebuie să lase la dreapta sa o arie egală cu $\frac{\alpha}{2} = 0.025$, iar la stânga o arie egală cu $1 - \frac{\alpha}{2} = 1 - 0.025 = 0.975$.



Această modalitate de determinare a intervalului de încredere se poate sintetiza în **testul Z**.

Algoritmul testului Z

Presupunem dată o selecție de valori independente (de volum n) dintr-o populație de medie m necunoscută și dispersie σ^2 ($\sigma > 0$) cunoscută.

Pasul 1. Se calculează \bar{x} .

Pasul 2. Se consideră statistica $Z = \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}}$.

Pasul 3. Pentru un nivel de încredere prescris $(1 - \alpha) \cdot 100\%$ se determină $z_{\frac{\alpha}{2}} > 0$ astfel încât $\Phi(-z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$.

Pasul 4. Se determină intervalul de încredere pentru m

$$\left[\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right].$$

Comparăm intervalele obținute în exemplul de mai sus în funcție de nivelul de încredere

$(1 - \alpha) \cdot 100\%$	$\frac{\alpha}{2}$	$z_{\frac{\alpha}{2}}$	$\left[\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right]$
90%	0.05	1.6449	[73.0760; 77.0445]
95%	0.025	1.9599	[72.6960; 77.4245]
99%	0.005	2.5758	[71.9530; 78.1675]

Din tabel se observă că lungimea intervalului este invers proporțională cu nivelul de încredere.

Am putea spune că 95% dintre studenți au punctajele cuprinse în intervalul [72.6960; 77.4245]? Această interpretare nu este corectă deoarece valoarea exactă a mediei nu este cunoscută și afirmația $m \in [72.6960; 77.4245]$ poate fi corectă sau nu deoarece intervalul de încredere construit este aleator, el bazându-se pe o selecție aleatoare.

Interpretarea corectă este: dacă facem un număr mare de selecții și de fiecare dată calculăm intervalul de încredere pentru medie cu nivelul de încredere de 95%, atunci în 95% din aceste intervale vor conține valoarea corectă a mediei. Deci metoda folosită ne permite să obținem intervale pentru medie care vor conține în 95% din cazuri valoarea corectă.

Alegerea nivelului de încredere este arbitrară. Ne punem problema ce se întâmplă dacă mărim nivelul de încredere, de exemplu, la 99%? Este rezonabil să dorim să mărim nivelul de încredere. În acest caz, pentru exemplul considerat, intervalul de încredere va fi [71.9530; 78.1675], deci va fi mai mare decât în cazul nivelului de 95%. Dacă dimensiunea eșantionului și abaterea medie pătratică sunt păstrate constante, atunci un nivel mai înalt de încredere atrage un interval de încredere mai mare.

Lungimea intervalului de încredere este o măsură a preciziei estimării. Din cele prezentate rezultă că precizia este invers proporțională cu nivelul de încredere. Este preferabil să obținem un interval de încredere cât mai scurt pentru o problema pusă, dar cu un nivel de încredere adecvat. Un mod de a atinge acest scop este alegerea dimensiunii eșantionului astfel încât cu ajutorul acestei selecții să putem obține un interval de încredere de lungime specificată și cu nivelul de încredere dat.

Intervalele de încredere studiate până acum sunt bilaterale în sensul că dădeau ca rezultat un interval închis. Dacă există o informație relativă la valoarea medie de forma că aceasta nu este limitată superior, atunci intervalul de încredere devine de forma $\left(\bar{x} - z_{\alpha} \frac{\sigma}{\sqrt{n}}, \infty \right)$ și este un interval de încredere unilateral.

În acest caz

$$P(Z > -z) = 1 - \alpha \Leftrightarrow P\left(m \in \left(\bar{x} - z \frac{\sigma}{\sqrt{n}}, \infty\right)\right) = 1 - \alpha \Leftrightarrow 1 - \Phi(-z) = 1 - \alpha \Leftrightarrow \Phi(-z) = \alpha$$

Notăm cu z_{α} valoarea obținută din relația $\Phi(-z) = \alpha$.

O situație similară are loc dacă valoarea medie nu este limitată inferior, intervalul de încredere fiind $\left(-\infty, \bar{x} + z_{\alpha} \frac{\sigma}{\sqrt{n}}\right)$, iar valoarea z_{α} se obține din relația $\Phi(z) = 1 - \alpha$.

12.2 Intervale de încredere pentru medie în cazul σ necunoscut

Presupunem că populația studiată are o distribuție normală cu media m și dispersia σ^2 necunoscute. Facem o selecție de dimensiune n . Fie x_1, x_2, \dots, x_n valorile variabilelor de

selecție X_1, X_2, \dots, X_n . Putem calcula media de selecție $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ și dispersia de selecție modificată $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$.

Vrem să calculăm un interval de încredere pentru m . Dacă dispersia este cunoscută, știm că $Z = \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}}$ urmează o distribuție normală. Dacă σ este necunoscut o procedură normală

este de a înlocui σ cu s . Statistica devine acum $T = \frac{\bar{X} - m}{\frac{s}{\sqrt{n}}}$. O întrebare logică care se

pune este următoarea: care este efectul înlocuirii lui σ cu s asupra distribuției statisticii T ? Dacă n este suficient de mare, răspunsul la această întrebare este: efectul este "destul de mic" și putem considera că urmează o distribuție normală standard. În general n trebuie să fie cel puțin 40. Teorema limită centrală are loc pentru $n \geq 30$, dar mărirea eșantionului recomandată este la cel puțin 40, deoarece înlocuirea lui σ cu s în Z conduce la modificări suplimentare ale distribuției.

În acest caz intervalul de încredere se construiește astfel:

Pasul 1. Se calculează $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ și $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$.

Pasul 2. Se consideră statistica $Z = \frac{\bar{x} - m}{\frac{s}{\sqrt{n}}}$.

Pasul 3. Pentru un nivel de încredere prescris $(1 - \alpha) \cdot 100\%$ se determină $z_{\frac{\alpha}{2}} > 0$ astfel încât $\Phi(-z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$.

Pasul 4. Se determină intervalul de încredere pentru m ,

$$\left[\bar{x} - z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \right].$$

Dacă n este mic, cum se întâmplă în multe probleme din inginerie, trebuie folosită distribuția Student pentru construirea intervalului de încredere.

Testul Student

Presupunem că populația studiată are o distribuție normală cu media m și abaterea medie pătratică σ necunoscute. Facem o selecție de dimensiune n , n mic. Vrem să calculăm un interval de încredere pentru m .

Teorema 12.2.1 Fie X_1, X_2, \dots, X_n independente, care urmează o distribuție normală cu media m și dispersia σ^2 . Fie x_1, x_2, \dots, x_n , $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ și fie $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$, unde \bar{x} reprezintă media de selecție, s reprezintă abaterea medie de selecție. Statistica $T = \frac{\bar{X} - m}{\frac{s}{\sqrt{n}}}$ urmează o distribuție Student cu $n - 1$ grade de libertate.

În tabelul din Anexă pentru funcția de repartiție a distribuției Student pe prima linie sunt date valorile lui α iar pe coloană sunt trecute gradele de libertate. Astfel calculăm

$$F(t_{\alpha,n}) = P(T \leq t_{\alpha,n}) = \int_{-\infty}^{t_{\alpha,n}} f(x) dx = 1 - \alpha,$$

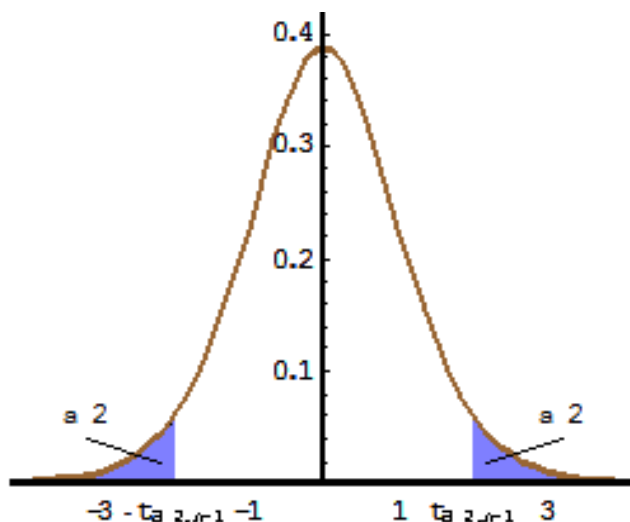
unde $f(x)$ este densitatea de probabilitate a distribuției Student.

Pentru valorile negative se folosește faptul că

$$F(-t_{\alpha,n}) = 1 - F(t_{\alpha,n}). \quad (12.2)$$

Deoarece distribuția Student este simetrică, avem $t_{1-\alpha,n} = -t_{\alpha,n}$, ceea ce înseamnă că în partea dreaptă a lui $t_{\alpha,n}$, dar și în partea stângă a lui $t_{1-\alpha,n}$, aria este α .

Pentru orice $\alpha \in (0, 1)$ se poate determina pragul $t_{\frac{\alpha}{2}, n-1} > 0$ astfel încât $P(|T_{n-1}| \leq t_{\frac{\alpha}{2}, n-1}) = 1 - \alpha$. Se alege $t_{\frac{\alpha}{2}, n-1}$ astfel încât ariile colorate din figură, să fie fiecare $\frac{\alpha}{2}$.



Înlocuind $T_{n-1} = \frac{(\bar{X} - m) \sqrt{n}}{s}$, rezultă

$$P(-t_{\frac{\alpha}{2}, n-1} \leq \frac{(\bar{X} - m) \sqrt{n}}{s} \leq t_{\frac{\alpha}{2}, n-1}) = 1 - \alpha \Leftrightarrow$$

$$P\left(m \in \left[\bar{x} - \frac{s}{\sqrt{n}} t_{\frac{\alpha}{2}, n-1}, \bar{x} + \frac{s}{\sqrt{n}} t_{\frac{\alpha}{2}, n-1}\right]\right) = 1 - \alpha.$$

Rezultă că intervalul

$$\left[\bar{x} - \frac{s}{\sqrt{n}} t_{\frac{\alpha}{2}, n-1}, \bar{x} + \frac{s}{\sqrt{n}} t_{\frac{\alpha}{2}, n-1}\right]$$

este un interval de încredere pentru media m cu coeficientul de încredere $100(1 - \alpha)\%$.

Algoritmul testului Student (mai este cunoscut sub denumirea de **testul T**)

Fie x_1, x_2, \dots, x_n o selecție de variabile de selecție X_1, X_2, \dots, X_n i.i.d. dintr-o populație normală cu media m și dispersia σ^2 necunoscute.

Pasul 1. Se calculează $\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k$ și $s = \left(\frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2\right)^{\frac{1}{2}}$.

Pasul 2. Se consideră statistica $T = \frac{\bar{X} - m}{\frac{s}{\sqrt{n}}}$.

Pasul 3. Pentru un coeficient de încredere prescris $(1 - \alpha) \cdot 100\%$ se determină din tabelul funcției de repartiție Student sau cu ajutorul softurilor numărul $t_{\frac{\alpha}{2}, n-1} > 0$ astfel încât $P(|T| \leq t_{\frac{\alpha}{2}, n-1}) = 1 - \alpha$.

Pasul 4. Se determină intervalul de încredere pentru m ,

$$\left[\bar{x} - \frac{s}{\sqrt{n}} t_{\frac{\alpha}{2}, n-1}, \bar{x} + \frac{s}{\sqrt{n}} t_{\frac{\alpha}{2}, n-1}\right].$$

Exemplul 12.2.2 Considerăm un eșantion de 36 studenți care au obținut punctajele: 64, 62, 76, 82, 66, 74, 72, 71, 73, 70, 75, 77, 84, 92, 86, 62, 58, 80, 79, 84, 83, 62, 78, 84, 94, 88, 77, 58, 65, 53, 60, 49, 68, 74, 78, 98.

Să se stabilească intervale de 90%, 95% și 99% încredere pentru media punctajelor obținute.

Rezolvare. Folosim testul Z .

Rezultă $\bar{x} = 73.7778$, $\sigma \simeq s = 11.6082$.

Statistica T este distribuită Student cu 35 grade de libertate (a se consulta anexa 2 cu tabelul valorilor funcției de repartiție Student în funcție de gradele de libertate).

Pentru 90% încredere (deci eroare sub 10%) avem $t_{0,05} = 1.68957$. Intervalul cerut are capetele $73.7778 \pm 1.68957 \cdot \frac{11.6082}{\sqrt{36}}$, adică [70.509; 77.0466].

Pentru 95% încredere (deci eroare sub 5%) avem $t_{0,025} = 2.03011$. Intervalul cerut are capetele $73.7778 \pm 2.03011 \cdot \frac{11.6082}{\sqrt{36}}$, adică [69.8501; 77.7054].

Pentru 99% încredere (deci eroare sub 1%) avem $t_{0,005} = 2.72381$. Intervalul cerut are capetele $73.7778 \pm 2.72381 \cdot \frac{11.6082}{\sqrt{36}}$, adică [68.5081; 79.0475].

Și în acest caz putem construi intervale de încredere de forma $\left(\bar{x} - t_{\alpha, n-1} \frac{s}{\sqrt{n}}, \infty\right)$ sau $\left(-\infty, \bar{x} + t_{\alpha, n-1} \frac{s}{\sqrt{n}}\right)$ pentru intervale de încredere unilaterale.

12.3 Intervale de încredere pentru dispersie

Uneori este necesar calculul intervalului de încredere pentru dispersia unei caracteristici studiate. Dacă populația este modelată de o distribuție normală putem aplica intervalele descrise în continuare.

Teorema 12.3.1 Fie variabilele de selecție X_1, X_2, \dots, X_n i.i.d. (independente și identic distribuite) cu $X \in N(m, \sigma)$, media m și dispersia σ^2 necunoscute. Statistica

$$\chi_{n-1}^2 = \frac{(n-1)s^2}{\sigma^2}$$

urmează o distribuție hi-pătrat cu $n-1$ grade de libertate.

Definim $x_{\alpha, n}$ ca fiind punctul pentru care este satisfăcută inegalitatea

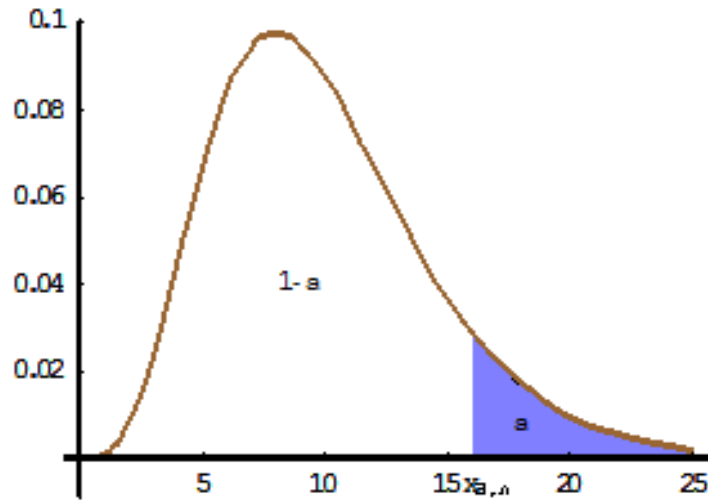
$$P(\chi^2 \leq x_{\alpha, n}) = \int_{-\infty}^{x_{\alpha, n}} f(t) dt = 1 - \alpha, \quad (12.3)$$

unde $f(t)$ este densitatea de probabilitate a distribuției $\chi^2(n)$.

Probabilitatea căutată este aria situată la stânga lui $x_{\alpha, n}$ din figura următoare.

Pentru a ilustra modul de utilizare a tabelului de valori ale funcției de repartiție hi-pătrat observăm că pe prima coloană sunt trecute gradele de libertate iar pe linie sunt trecute valorile lui α . De exemplu, pentru $n = 10$ și $\alpha = 0.05$ obținem $x_{0,05,10} = 18.31$, iar $x_{1-0,05,10} = 3.94$.

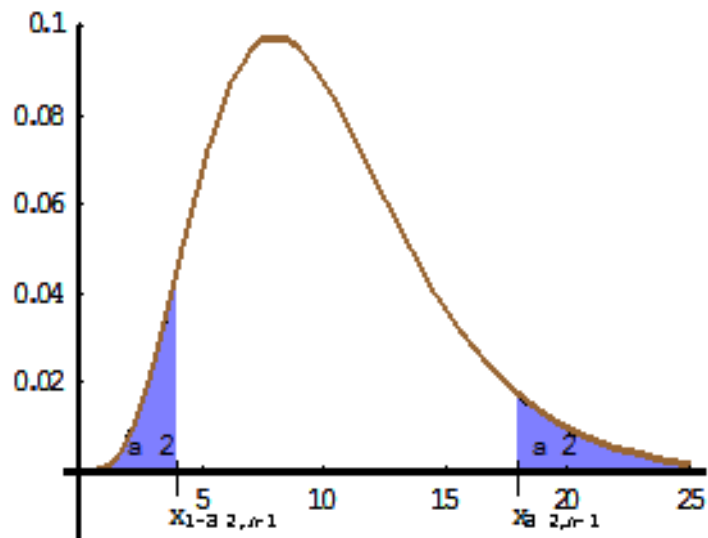
Deci $P(\chi^2 \leq x_{0,05,10}) = 1 - 0.05$, $P(\chi^2 \leq x_{0,95,10}) = 0.05$, $x_{0,05,10} = 18.31$, iar $x_{1-0,05,10} = 3.94$. \diamond



Pentru construcția intervalului de încredere pentru dispersie se folosește statistica $\chi_{n-1}^2 = \frac{(n-1)s^2}{\sigma^2} \in \chi^2(n-1)$. Pentru $\alpha \in (0, 1)$ dat determinăm $x_1(\alpha)$ și $x_2(\alpha)$ astfel încât

$$P(x_1(\alpha) \leq \chi_{n-1}^2 \leq x_2(\alpha)) = 1 - \alpha. \quad (12.4)$$

Numerele $x_1(\alpha)$ și $x_2(\alpha)$ nu sunt unic determinate și de obicei se aleg astfel încât $x_1(\alpha) = x_{1-\alpha/2, n-1}$ și $P(\chi_{n-1}^2 \leq x_{1-\alpha/2, n-1}) = \frac{\alpha}{2}$, iar $x_2(\alpha) = x_{\alpha/2, n-1}$ și $P(\chi_{n-1}^2 \geq x_{\alpha/2, n-1}) = \frac{\alpha}{2}$. Semnificația valorilor $x_{1-\alpha/2, n-1}$ și $x_{\alpha/2, n-1}$ poate fi văzută în figura următoare.



Înlocuind $\chi_{n-1}^2 = \frac{(n-1)s^2}{\sigma^2}$ în (12.4) și obținem:

$$P\left(x_{1-\alpha/2, n-1} \leq \frac{(n-1)s^2}{\sigma^2} \leq x_{\alpha/2, n-1}\right) = 1 - \alpha. \quad (12.5)$$

Relația (12.5) poate fi rearanjată astfel:

$$P\left(\frac{(n-1)s^2}{x_{\alpha/2, n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{x_{1-\alpha/2, n-1}}\right) = 1 - \alpha. \quad (12.6)$$

Am obținut astfel un interval de încredere pentru dispersie.

Algoritmul de determinare a intervalului de încredere pentru dispersie

Fie x_1, x_2, \dots, x_n o selecție de valori pentru variabilele de selecție X_1, X_2, \dots, X_n i.i.d. dintr-o populație normală cu media m și dispersia σ^2 necunoscute.

Pasul 1. Se calculează $\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k$ și $s = \left(\frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2 \right)^{\frac{1}{2}}$.

Pasul 2. Se alege statistica $\chi_{n-1}^2 = \frac{(n-1)s^2}{\sigma^2}$ despre care se știe că urmează o distribuție χ^2 cu $n-1$ grade de libertate.

Pasul 3. Pentru un nivel de încredere prescris $(1-\alpha) \cdot 100\%$ se determină, din tabelul valorilor funcției de repartiție χ^2 cu $n-1$ grade de libertate sau cu ajutorul softurilor Matlab sau Mathematica, numerele $x_{1-\alpha/2, n-1}$ și $x_{\alpha/2, n-1}$ astfel încât

$$P(\chi_{n-1}^2 \leq x_{1-\alpha/2, n-1}) = \frac{\alpha}{2} \quad \text{și} \quad P(\chi_{n-1}^2 \leq x_{\alpha/2, n-1}) = 1 - \frac{\alpha}{2}.$$

Pasul 4. Se determină intervalul $\left[\frac{(n-1)s^2}{x_{\alpha/2, n-1}}, \frac{(n-1)s^2}{x_{1-\alpha/2, n-1}} \right]$ și rezultă intervalul de încredere pentru σ^2 .

Observația 12.3.2 Este posibil să determinăm intervale nemărginite inferior sau superior pentru dispersie cu un anumit nivel de încredere $(1-\alpha) 100\%$. Acestea sunt $\left(-\infty, \frac{(n-1)s^2}{x_{1-\alpha, n-1}}\right)$, respectiv $\left(\frac{(n-1)s^2}{x_{\alpha, n-1}}, \infty\right)$.

Exemplul 12.3.3 Reluăm Exemplul 12.1.1. Dorim să construim intervale de încredere pentru dispersie.

Rezolvare. Avem $n = 83$. Calculăm $s = 11.3517$. Considerăm statistica $\chi_{82}^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{82s^2}{\sigma^2}$.

Pentru nivelul de încredere de 90% avem $x_{0.95, 83} = 62.1323$ și $x_{0.05, 83} = 104.139$. Intervalul de încredere pentru dispersie este $[101.468, 170.068]$.

Pentru nivelul de încredere de 95% avem $x_{0.975, 83} = 58.8446$ și $x_{0.025, 83} = 108.937$. Intervalul de încredere pentru dispersie este $[96.998, 179.57]$.

Pentru nivelul de încredere de 99% avem $x_{0.995, 83} = 52.7674$ și $x_{0.005, 83} = 118.726$. Intervalul de încredere pentru dispersie este $[89.0006, 200.251]$.

Exemplul 12.3.4 Media erorilor de măsurare a lungimilor unor baghete metalice este de 3 mm. Presupunem că aceste erori respectă legea normală cu media 3 mm și dispersia necunoscută. Se face o selecție de volum 6: $\{-1, 4, 4, 1, 3, 1\}$. Se cere un interval de estimare pentru dispersie cu nivel de încredere de 90%.

Rezolvare. Avem $n = 6, m = 3$. Calculăm

$$s^2 = \frac{1}{3} \left((-1-3)^2 + (4-3)^2 + (4-3)^2 + (1-3)^2 + (3-3)^2 + (1-3)^2 \right) = \frac{26}{3} = 8.6667$$

Considerăm statistica $\chi_5^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{5s^2}{\sigma^2}$.

Pentru nivelul de încredere de 90% avem $x_{0.95,5} = 1.14548$ și $x_{0.05,5} = 11.0705$.

Intervalul de încredere pentru dispersie este [2.34858; 22.698].

Se observă că intervalul este destul de mare, deci precizia pentru dispersie este mică, chiar dacă apare cu probabilitate mare.

12.4 Intervale de încredere pentru proporții

Pentru o populație a cărei membri pot fi clasificați în funcție de o anumită caracteristică în două categorii: fie p probabilitatea de a aparține unei categorii, numit succes și $1 - p$ probabilitatea de a aparține celeilalte categorii, numită eșec. Parametrul p poartă denumirea de proporția populației și ipotezele asupra lui p se fac numărând succesele, $X = \sum_{i=1}^n X_i$ ($\leq n$),

unde

$$X_i : \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}.$$

Atunci $\hat{P} = \frac{X}{n}$ un estimator punctual al lui p . Reamintim că n și p sunt parametrii unei distribuții binomiale. Mai mult, statistica P urmează o distribuție normală cu media p și dispersia $\frac{p(1-p)}{n}$ dacă p nu este aproape de 0 sau 1 și dacă n este relativ mare.

Aceasta înseamnă că pentru a folosi această aproximare este necesar ca $np \geq 5$, $n(1-p) \geq 5$.

Teorema 12.4.1 *Dacă n este astfel încât $np \geq 5$, $n(1-p) \geq 5$, atunci*

$$Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

urmează aproximativ o distribuție normală standard.

Pentru a construi un interval de încredere pentru p , observăm că

$$P(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}) = 1 - \alpha$$

sau

$$P\left(-z_{\frac{\alpha}{2}} \leq \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha,$$

unde $z_{\frac{\alpha}{2}} > 0$ astfel încât $\Phi(-z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$.

Această relație poate fi rearanjată astfel:

$$P\left(\hat{P} - z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{P} + z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}\right) = 1 - \alpha.$$

Cantitatea $\sqrt{\frac{p(1-p)}{n}}$ se numește *eroarea standard* a estimatorului punctual \hat{P} . Deoarece marginile intervalului conțin p care este necunoscut, o soluție satisfăcătoare este înlocuirea sa cu \hat{P} . Astfel obținem

$$P\left(\hat{P} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \leq p \leq \hat{P} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\right) \simeq 1 - \alpha.$$

Aceasta conduce la determinarea unui interval de încredere cu nivelul de încredere de $(1 - \alpha)100\%$ și acesta este

$$\left[\hat{P} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}, \hat{P} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \right].$$

Exemplul 12.4.2 Se presupune ca la 85 de autoturisme de o anumită marcă se studiază arborele cotit după un anumit timp de funcționare. Se constată că la 10 automobile acesta prezintă defecte ce impunea înlocuirea lui. Să se determine un estimator punctual al numărului ce reprezintă ce proporție dintre automobilele de acest tip prezintă această deficiență. Să se interpreteze acest rezultat și să se estimeze ce proporție dintre automobilele de acest tip prezintă această deficiență.

Rezolvare. Pentru fiecare autoturism studiat, se obține rezultatul că arborele cotit este defect sau nu. Putem aplica modelul Bernoulli și estimatorul de verosimilitate maximă va fi:

$$\hat{P} = \frac{10}{85} = 0.117655.$$

Eroarea standard a estimatorului punctual este

$$\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} = \sqrt{\frac{\frac{10}{85}(1-\frac{10}{85})}{85}} = 0.034946.$$

Nu este foarte clar cum interpretăm această valoare de 0.034946. Este această estimare foarte exactă, exactă sau nu?

Construim un interval de încredere cu un nivel de încredere de 95% bazat pe valoarea observată 0.117655.

$$\begin{aligned} \hat{P} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} &\leq p \leq \hat{P} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}, \\ 0.1176 - 1.96 \sqrt{\frac{0.1176 \cdot 0.88}{85}} &\leq p \leq 0.1176 + 1.96 \sqrt{\frac{0.1176 \cdot 0.88}{85}}, \\ 0.04915 &\leq p \leq 0.186141. \end{aligned}$$

Concluzia: în 95% din cazuri probabilitatea ca automobilul să prezinte defectul studiat este cuprinsă în intervalul [0.0491534, 0.186141].

12.4.1 Alegerea dimensiunii eșantionului

Deoarece \hat{P} este un estimator punctual al lui p , putem defini eroarea de estimare a lui p prin \hat{P} de forma $E = |p - \hat{P}|$. Observăm că suntem aproximativ $(1 - \alpha)100\%$ siguri că această eroare este mai mică decât $z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$. În exemplul anterior suntem 95% siguri că $\hat{P} = \frac{10}{85} = 0.117655$ diferă de valoarea exactă a lui p cu mai puțin de

$$1.95996 \sqrt{\frac{\frac{10}{85}(1-\frac{10}{85})}{85}} = 0.069084.$$

În situațiile în care dimensiunea eșantionului poate fi aleasă, putem lua n astfel încât să fim aproximativ $(1 - \alpha)100\%$ siguri că eroarea este mai mică decât o valoare specificată E .

Dacă considerăm $E = z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$ și rezolvăm ecuația în raport cu n obținem ca dimensiune a eșantionului

$$n = \left[\left(\frac{z_{\frac{\alpha}{2}}}{E} \right)^2 \hat{P}(1 - \hat{P}) \right] + 1. \quad (12.7)$$

O estimare a lui p este necesară. Pentru aceasta se poate lua un eșantion preliminar, se calculează \hat{P} și apoi folosind relația (12.7) putem determina câte observații mai trebuie făcute pentru a obține pentru p o estimare satisfăcătoare.

O altă soluție este de a ține seama că $p(1-p) \leq \frac{1}{4}$ și atunci

$$n = \left\lceil \frac{1}{4} \left(\frac{z_{\frac{\alpha}{2}}}{E} \right)^2 \right\rceil + 1. \quad (12.8)$$

Exemplul 12.4.3 Reluăm Exemplul 12.4.2. Cât de mare trebuie luat eșantionul astfel încât să fim 95% siguri că eroarea pe care o facem când folosim \hat{P} ca estimator a lui p să fie mai mică decât 0.05?

Rezolvare. Folosind $\hat{P} = \frac{10}{85}$ găsim

$$n = \left\lceil \left(\frac{z_{\frac{\alpha}{2}}}{E} \right)^2 \hat{p}(1-\hat{p}) \right\rceil + 1 = \left\lceil \left(\frac{1.96}{0.05} \right)^2 \frac{10}{85} \left(1 - \frac{10}{85} \right) \right\rceil + 1 = 160.$$

Dacă vrem să fim cel puțin 95% siguri că eroarea pe care o facem când folosim \hat{P} ca estimator a lui p să fie mai mică decât 0.05 putem folosi formula (12.8)

$$n = \left\lceil \frac{1}{4} \left(\frac{z_{\frac{\alpha}{2}}}{E} \right)^2 \right\rceil + 1 = \left\lceil \frac{1}{4} \left(\frac{1.96}{0.05} \right)^2 \right\rceil + 1 = 385.$$

Observăm că dacă avem o informație privind valoarea lui p , chiar dintr-un eșantion preliminar, obținem o dimensiune mai mică a eșantionului menținând precizia estimării și nivelul de încredere. \diamond

Este posibil să determinăm intervale deschise nemărginite inferior sau superior pentru proporția p cu un anumit nivel de încredere $(1-\alpha)100\%$. Acestea sunt

$$\left(-\infty, \hat{P} + z_{\alpha} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \right), \quad \left(\hat{P} - z_{\alpha} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}, \infty \right).$$

12.5 Predicția

În unele situații suntem interesați în a prevedea următoarele valori ale variabilei de selecție. Vom vedea cum se determină un interval de predicție cu $(1-\alpha)100\%$ nivel de încredere pentru următoarea valoare a unei variabile de selecție care urmează o distribuție normală.

Fie x_1, x_2, \dots, x_n o selecție de valori pentru variabile de selecție X_1, X_2, \dots, X_n i.i.d. dintr-o populație normală cu media m și dispersia σ^2 . Dorim să prevedem valoarea variabilei de selecție X_{n+1} la o singură observație viitoare.

Un punct de predicție este \bar{x} . Eroarea de predicție este $X_{n+1} - \bar{x}$. Media erorii de predicție este $M[X_{n+1} - \bar{x}] = m - m = 0$, iar dispersia este $D[X_{n+1} - \bar{x}] = \sigma^2 + \frac{\sigma^2}{n} = \sigma^2 \left(1 + \frac{1}{n} \right)$, deoarece X_{n+1} este independentă de media \bar{x} .

Predicția $X_{n+1} - \bar{x}$ este normal distribuită și atunci $Z = \frac{X_{n+1} - \bar{x}}{\sigma \sqrt{1 + \frac{1}{n}}}$ are o distribuție normală standard.

Dacă înlocuim σ prin s obținem $T = \frac{X_{n+1} - \bar{x}}{s \sqrt{1 + \frac{1}{n}}}$ care are o distribuție Student cu $n-1$ grade de libertate. Obținem astfel intervalele de predicție cu $(1-\alpha)100\%$ nivel de încredere

$$P \left(-t_{\frac{\alpha}{2}, n-1} \leq \frac{X_{n+1} - \bar{x}}{\sigma \sqrt{1 + \frac{1}{n}}} \leq t_{\frac{\alpha}{2}, n-1} \right) = 1 - \alpha \Rightarrow$$

$$\left[\bar{x} - z_{\frac{\alpha}{2}} \sigma \sqrt{1 + \frac{1}{n}} \leq X_{n+1} \leq \bar{x} + z_{\frac{\alpha}{2}} \sigma \sqrt{1 + \frac{1}{n}} \right]$$

și respectiv

$$\left[\bar{x} - t_{\frac{\alpha}{2}, n-1} s \sqrt{1 + \frac{1}{n}} \leq X_{n+1} \leq \bar{x} + t_{\frac{\alpha}{2}, n-1} s \sqrt{1 + \frac{1}{n}} \right].$$

Exemplul 12.5.1 Se studiază o caracteristică pe $n = 22$ unități statistice și se obțin valorile: 19.8, 15.4, 11.4, 19.5, 10.1, 18.5, 14.1, 8.8, 14.9, 7.9, 17.6, 13.6, 7.5, 12.7, 16.7, 11.9, 15.4, 11.9, 15.8, 11.4, 15.4, 11.4. Să se determine intervalul de predicție pentru cea de a 23-a valoare și intervalul de încredere pentru media valorilor obținute, nivelul de încredere fiind de 95%, iar caracteristica urmează o repartiție normală.

Rezolvare. Calculăm și obținem $\bar{x} = 13.71$ și $s = 3.55$. Intervalul de încredere pentru medie este

$$\begin{aligned} & \left[\bar{x} - \frac{s}{\sqrt{n}} t_{\frac{\alpha}{2}, n-1}, \bar{x} + \frac{s}{\sqrt{n}} t_{\frac{\alpha}{2}, n-1} \right] = \\ & = \left[13.7136 - \frac{3.55358}{\sqrt{22}} \cdot 2.07961, 13.7136 + \frac{3.55358}{\sqrt{22}} \cdot 2.07961 \right] = \\ & = [12.1381, 15.2892]. \end{aligned}$$

Intervalul de predicție cu 95% încredere este

$$\begin{aligned} & \left[\bar{x} - t_{\frac{\alpha}{2}, n-1} s \sqrt{1 + \frac{1}{n}}, \bar{x} + t_{\frac{\alpha}{2}, n-1} s \sqrt{1 + \frac{1}{n}} \right] = \\ & = \left[13.713 - 2.079 \cdot 3.553 \sqrt{1 + \frac{1}{22}}, 13.713 + 2.079 \cdot 3.553 \sqrt{1 + \frac{1}{22}} \right] = \\ & = [6.48601; 20.9413]. \end{aligned}$$

Rezultă că $6.48601 \leq X_{23} \leq 20.9413$.

Remarcăm că intervalul de predicție este considerabil mai lung decât cel de încredere. Lungimea intervalului de încredere, pentru $n \rightarrow \infty$, tinde la zero pe când lungimea intervalului de predicție tinde la $2t_{\frac{\alpha}{2}, n-1}s$.

12.6 Intervale de toleranță pentru caracteristici ce urmează o distribuție normală

Definiția 12.6.1 Intervalul de toleranță este un interval statistic în care, cu un nivel de încredere dat, caracteristica studiată a unei populații ia valori cu o probabilitate de acoperire specificată.

Capetele intervalului de toleranță se numesc *limite de toleranță*. În cazul intervalului de toleranță se cunoaște distribuția pe care o urmează caracteristica și, eventual, se cunosc parametrii distribuției și se determină intervalul cu o probabilitate de acoperire și cu un anumit nivel de încredere. Diferă de intervalul de încredere deoarece acesta din urmă furnizează limite pentru un anumit parametru al populației, necunoscut, de ex. medie, dispersie numai cu un anumit nivel de încredere.

În inginerie se specifică, de obicei, limitele de toleranță, de acceptibilitate ale caracteristicii unui produs și nu reflectă întotdeauna valoarea obținută prin măsurători.

Exemplul 12.6.2 Studiem o categorie de procesoare. Din istoricul studiului procesoarelor de acest tip se știe că frecvența procesoarelor urmează o distribuție normală cu media $m = 600$ megahertzi și abaterea medie pătratică $\sigma = 30$ megahertzi. Să se determine un interval de toleranță pentru frecvența procesoarelor cu un nivel de încredere de 95% și cu o probabilitate de acoperire de 95%.

Dacă media și dispersia sunt cunoscute, intervalul de toleranță se definește de forma $[m - z_{\frac{\alpha}{2}}\sigma, m + z_{\frac{\alpha}{2}}\sigma]$, în funcție de nivelul de încredere $100(1 - \alpha)\%$. Probabilitatea de acoperire este $1 - \alpha$.

În exemplul dat, pentru un nivel de încredere de 95% avem intervalul $[550.65; 649.35]$.

$$m - z_{\frac{\alpha}{2}}\sigma = 600 - 1.96 \cdot 30 = 541.2$$

$$m + z_{\frac{\alpha}{2}}\sigma = 600 + 1.96 \cdot 30 = 658.8$$

Ele reprezintă valorile acceptate pentru frecvența procesoarelor.

Dacă m și σ nu sunt cunoscuți, putem folosi \bar{x} și s pentru a calcula intervalul de toleranță, $[\bar{x} - z_{\frac{\alpha}{2}}s, \bar{x} + z_{\frac{\alpha}{2}}s]$.

Este de așteptat ca în acest caz, datorită introducerii lui \bar{x} și s , probabilitatea de acoperire să fie mai mică decât $1 - \alpha$ cu un nivel de încredere de $(1 - \alpha) \cdot 100\%$. Soluția este de a înlocui pe $z_{\frac{\alpha}{2}}$ cu o valoare care să asigure un nivel de încredere de $(1 - \alpha) \cdot 100\%$.

Un interval de toleranță care să aibă probabilitatea de acoperire p pentru o caracteristică care urmează o distribuție normală, cu un nivel de încredere de $(1 - \alpha) \cdot 100\%$ este $[\bar{x} - ks, \bar{x} + ks]$, unde k este un factor de toleranță.

Un exemplu de calcul al factorului de toleranță este dat de Howe, W. G. (1969).

Formula factorului de toleranță dată de Howe este

$$k = \sqrt{\frac{(n-1)(1 + \frac{1}{n})z_{(1+\alpha)/2}^2}{\chi_{1-p, n-1}}}$$

Algoritmul de calcul al factorului de toleranță, în funcție de proporția p și de probabilitatea γ , este:

1. Se determină $z_{(1+p)/2}$ astfel încât dacă Z urmează o distribuție normală, $P(Z \leq z_{(1+p)/2}) = (1 + p)/2$;
2. Se determină $\chi_{\alpha, n-1}$ astfel că dacă χ urmează o distribuție hi-pătrat cu $n - 1$ grade de libertate, $P(\chi \leq \chi_{\alpha, n-1}) = \alpha$;
3. Se înlocuiesc în formula lui k .

12.7 Intervale de încredere pentru caracteristicile a două populații care urmează o distribuție normală

Până acum au fost prezentate moduri de construcție ale intervalelor de încredere pentru un parametru corespunzător unei singure populații. Vom extinde în continuare aceste rezultate în cazul a două populații independente care urmează distribuții normale.

Presupunem că avem o populație care urmează o distribuție normală cu media m_1 și dispersia σ_1^2 , iar cea de a doua populație urmează o distribuție normală cu media m_2 și dispersia σ_2^2 . Intervalele vor fi construite pe baza a două eșantioane de volum n_1 respectiv n_2 . Fie $x_{11}, x_{12}, \dots, x_{1n_1}$ o selecție de valori pentru variabile de selecție $X_{11}, X_{12}, \dots, X_{1n_1}$ i.i.d. din prima populație distribuită normal și $x_{21}, x_{22}, \dots, x_{2n_2}$ o selecție de valori pentru variabile de selecție $X_{21}, X_{22}, \dots, X_{2n_2}$ i.i.r. din a doua populație distribuită normal.

12.7.1 Intervale de încredere pentru diferența mediilor a două populații ale căror dispersii sunt cunoscute

Construim intervale de încredere, cu nivelul de încredere $(1 - \alpha) \cdot 100\%$, pentru diferența mediilor a două populații care urmează o distribuție normală și ale căror dispersii sunt cunoscute.

Facem următoarele ipoteze:

1. Se consideră două populații independente care urmează o distribuție normală cu dispersiile cunoscute;

2. Se consideră $x_{11}, x_{12}, \dots, x_{1n_1}$ o selecție de n_1 valori pentru variabile de selecție $X_{11}, X_{12}, \dots, X_{1n_1}$ i.i.d. din prima populație;

3. Se consideră $x_{21}, x_{22}, \dots, x_{2n_2}$ o selecție de n_2 valori pentru variabile de selecție $X_{21}, X_{22}, \dots, X_{2n_2}$ i.i.d. din a doua populație.

Un estimator logic pentru diferența mediilor, $m_1 - m_2$ este diferența dintre mediile statistice ale celor două eșantioane, $\bar{X}_1 - \bar{X}_2$, unde $\bar{X}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_{1i}$ și $\bar{X}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} X_{2i}$. Folosind proprietățile mediei și dispersiei obținem că

$$\begin{aligned} M[\bar{X}_1 - \bar{X}_2] &= M[\bar{X}_1] - M[\bar{X}_2] = m_1 - m_2, \\ D[\bar{X}_1 - \bar{X}_2] &= D[\bar{X}_1] + D[\bar{X}_2] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}. \end{aligned}$$

Ținând seama de presupunerile făcute și de rezultatele anterioare putem afirma

Propoziția 12.7.1 *Statistica*

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (12.9)$$

urmează o distribuție normală standard.

Observația 12.7.2 *Dacă sunt îndeplinite condițiile Teoremei limită centrală pentru cele două populații, atunci rezultatul Propoziției 12.7.1 se păstrează.*

Ținând seama de rezultatul Propoziției 12.7.1 putem scrie

$$\begin{aligned} P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) &= 1 - \alpha \Leftrightarrow \\ P\left(-z_{\alpha/2} \leq \frac{\bar{X}_1 - \bar{X}_2 - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z_{\alpha/2}\right) &= 1 - \alpha \Leftrightarrow \\ P\left(\bar{X}_1 - \bar{X}_2 - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq m_1 - m_2 \leq \bar{X}_1 - \bar{X}_2 + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) &= 1 - \alpha. \end{aligned}$$

Dacă $\bar{x}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_{1i}$ și $\bar{x}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} x_{2i}$ atunci un interval de încredere pentru $m_1 - m_2$, cu un nivel de încredere $(1 - \alpha)\%$, este

$$\left[\bar{x}_1 - \bar{x}_2 - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \bar{x}_1 - \bar{x}_2 + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right]. \quad (12.10)$$

Exercițiul 1 Fie variabilele de selecție $X_{11}, X_{12}, \dots, X_{1n_1}$ i.i.d. care urmează o legea normală $N(m_1, \sigma_1)$ și reprezintă încasările în mii lei ale unui lanț de magazine din orasul A și variabile de selecție $X_{21}, X_{22}, \dots, X_{2n_2}$ i.i.d. care urmează o legea normală $N(m_2, \sigma_2)$ și reprezintă încasările în mii lei ale unui alt lanț de magazine din orașul B. Presupunem că cele două selecții sunt independente. S-au efectuat două sondaje, respectiv pentru X_1 și X_2 și s-au obținut următoarele date: pentru X_1 : 226.5, 224.1, 218.6, 220.1, 228.8, 229.6, 222.5 și pentru X_2 : 221.5, 230.2, 223.4, 224.3, 230.8, 223.8. Cu un nivel de încredere de respectiv 0.90%, 0.95%, 0.99% vrem să construim intervale de încredere pentru diferența mediilor, $m_1 - m_2$ dacă $\sigma_1 = 2.5, \sigma_2 = 3$, cunoscute.

Rezultă că intervalul de încredere pentru diferența mediilor încasărilor cu un nivel de încredere de 90% este $[-3.89678, 1.19202]$.

Observația 12.7.3 Și în acest caz putem construi intervale de încredere de forma

$$\left(\bar{x}_1 - \bar{x}_2 - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \infty \right) \quad \text{sau} \quad \left(-\infty, \bar{x}_1 - \bar{x}_2 + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$$

pentru diferența mediilor.

Alegerea dimensiunii eșantionului

Dacă dispersiile sunt cunoscute și dimensiunile eșantioanelor sunt egale, $n_1 = n_2 = n$, putem determina dimensiunea eșantionului astfel încât eroarea pe care o facem înlocuind $m_1 - m_2$ cu $\bar{x}_1 - \bar{x}_2$ să fie mai mică decât E cu un nivel de încredere $(1 - \alpha) 100\%$. Deoarece

$$|m_1 - m_2 - (\bar{x}_1 - \bar{x}_2)| \leq z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}} \leq E$$

rezultă

$$n \geq \left(\frac{z_{\alpha/2}}{E} \right)^2 \sqrt{\sigma_1^2 + \sigma_2^2}, \quad n \in \mathbb{N}.$$

12.7.2 Intervale de încredere pentru diferența mediilor a două populații ale căror dispersii sunt necunoscute

Extindem rezultatele anterioare pentru cazul în care cele două populații urmează o distribuție normală și dispersiile sunt necunoscute. Dacă volumele eșantioanelor n_1 și n_2 depășesc valoarea de 40, rezultatul anterior poate fi utilizat. Dacă eșantioanele sunt mai mici atunci construcția intervalului de încredere se bazează pe distribuția Student.

Construim intervalul de încredere cu nivelul de încredere egal cu $100(1 - \alpha)\%$, pentru diferența mediilor a două populații ale căror dispersii sunt necunoscute.

Două situații diferite trebuie tratate:

1. Cazul în care dispersiile sunt necunoscute, dar egale $\sigma_1^2 = \sigma_2^2 = \sigma^2$,
2. Cazul în care dispersiile sunt necunoscute și diferite.

Cazul 1. $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Se consideră $x_{11}, x_{12}, \dots, x_{1n_1}$ o selecție de n_1 valori pentru variabile de selecție $X_{11}, X_{12}, \dots, X_{1n_1}$ i.i.d. din prima populație și $x_{21}, x_{22}, \dots, x_{2n_2}$ o selecție de n_2 valori pentru variabile de selecție $X_{21}, X_{22}, \dots, X_{2n_2}$ i.i.d. din a doua populație. Fie mediile de selecție $\bar{X}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_{1i}$ și $\bar{X}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} X_{2i}$ și dispersiile de selecție $S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_{1i} - \bar{X}_1)^2$ și $S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (X_{2i} - \bar{X}_2)^2$.

Observăm că $M[\bar{X}_1 - \bar{X}_2] = M[\bar{X}_1] - M[\bar{X}_2] = m_1 - m_2$, deci $\bar{X}_1 - \bar{X}_2$ este un estimator nedeplasat pentru diferența mediilor. Dispersia lui $\bar{X}_1 - \bar{X}_2$ este

$$D[\bar{X}_1 - \bar{X}_2] = D[\bar{X}_1] + D[\bar{X}_2] = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right).$$

Combinăm cele două dispersii de selecție S_1^2 și S_2^2 pentru a forma un estimator al lui σ^2 . Acest estimator, notat S_p^2 , se definește astfel:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

Observăm că S_p^2 poate fi scris astfel

$$S_p^2 = \frac{n_1 - 1}{n_1 + n_2 - 2} S_1^2 + \frac{n_2 - 1}{n_1 + n_2 - 2} S_2^2 = \lambda S_1^2 + (1 - \lambda) S_2^2,$$

unde $0 < \lambda \leq 1$. S_p^2 este o combinație liniară de cele două dispersii de selecție în care λ depinde doar de dimensiunea eșantioanelor, n_1 și n_2 . Dacă $n_1 = n_2 = n$ atunci S_p^2 este media aritmetică a celor două dispersii de selecție.

Știm că $Z = \frac{\bar{X}_1 - \bar{X}_2 - (m_1 - m_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ urmează o distribuție normală standard. Înlocuind σ cu S_p obținem următorul rezultat:

Propoziția 12.7.4 *Statistica*

$$T = \frac{\bar{X}_1 - \bar{X}_2 - (m_1 - m_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

urmează o distribuție Student cu $n_1 + n_2 - 2$ grade de libertate.

Atunci

$$\begin{aligned} P(-t_{\alpha/2, n_1+n_2-2} \leq T \leq t_{\alpha/2, n_1+n_2-2}) &= 1 - \alpha \Leftrightarrow \\ P\left(-t_{\alpha/2, n_1+n_2-2} \leq \frac{\bar{X}_1 - \bar{X}_2 - (m_1 - m_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq t_{\alpha/2, n_1+n_2-2}\right) &= 1 - \alpha \Leftrightarrow \\ P\left(\bar{X}_1 - \bar{X}_2 - t_{\alpha/2, n_1+n_2-2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq m_1 - m_2 \leq \right. \\ \left. \leq \bar{X}_1 - \bar{X}_2 + t_{\alpha/2, n_1+n_2-2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right) &= 1 - \alpha. \end{aligned}$$

Dacă $\bar{x}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_{1i}$ și $\bar{x}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} x_{2i}$ și $s_1^2 = \frac{1}{n_1-1} \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2$ și $s_2^2 = \frac{1}{n_2-1} \sum_{i=1}^{n_2} (x_{2i} - \bar{x}_2)^2$ atunci un interval de încredere pentru $m_1 - m_2$, cu un nivel de încredere de $100(1 - \alpha)\%$ este

$$\left[\bar{x}_1 - \bar{x}_2 - t_{\alpha/2, n_1+n_2-2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \bar{x}_1 - \bar{x}_2 + t_{\alpha/2, n_1+n_2-2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right].$$

Cazul 2. $\sigma_1^2 \neq \sigma_2^2$

În unele situații nu putem presupune că dispersiile necunoscute sunt egale. Și în acest caz putem găsi un interval de încredere pentru diferența mediilor, cu un nivel de încredere de $100(1 - \alpha)\%$ folosind faptul că

$$T^* = \frac{\bar{X}_1 - \bar{X}_2 - (m_1 - m_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

urmează aproximativ o distribuție Student cu ν grade de libertate, unde

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}.$$

Dacă ν nu este întreg, se rotunjește prin lipsă la cel mai apropiat întreg.

Intervalul de încredere, în acest caz, este

$$\left[\bar{x}_1 - \bar{x}_2 - t_{\alpha/2, \nu} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, \bar{x}_1 - \bar{x}_2 + t_{\alpha/2, \nu} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right].$$

12.7.3 Intervale de încredere pentru raportul dispersiilor a două populații

Construim intervale de încredere, cu nivelul de încredere $(1 - \alpha) \cdot 100\%$, pentru raportul mediilor a două populații care urmează o distribuție normală și ale căror medii și dispersii sunt necunoscute.

Facem următoarele ipoteze:

1. Se consideră două populații independente care urmează o distribuție normală;
2. Se consideră $x_{11}, x_{12}, \dots, x_{1n_1}$ o selecție de n_1 valori pentru variabile de selecție $X_{11}, X_{12}, \dots, X_{1n_1}$ i.i.d. din prima populație;
3. Se consideră $x_{21}, x_{22}, \dots, x_{2n_2}$ o selecție de n_2 valori pentru variabile de selecție $X_{21}, X_{22}, \dots, X_{2n_2}$ i.i.d. din a doua populație.

Fie S_1^2 și S_2^2 dispersiile de selecție ale celor două populații.

Propoziția 12.7.5 *Statistica*

$$F = \frac{S_2^2/\sigma_2^2}{S_1^2/\sigma_1^2}$$

urmează o F -distribuție cu $n_2 - 1$ grade de libertate la numărător și $n_1 - 1$ grade de libertate la numitor.

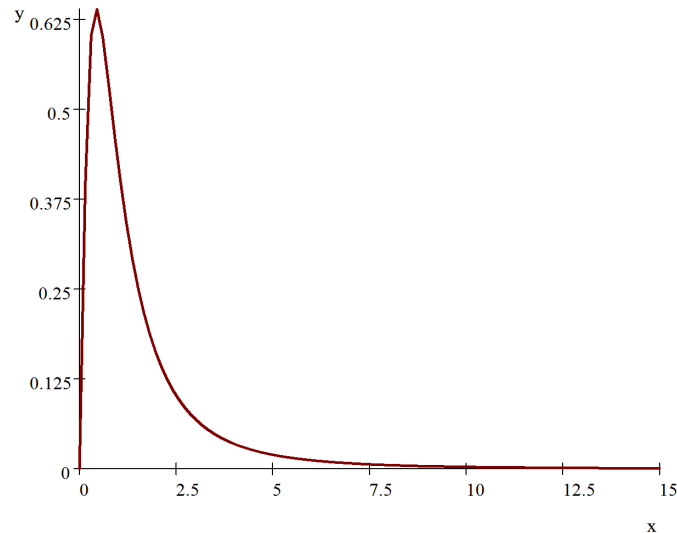
Definiția 12.7.6 *Densitatea de probabilitate a F -distribuției cu u grade de libertate la numărător și v grade de libertate la numitor este*

$$f(x) = \frac{\Gamma\left(\frac{u+v}{2}\right) \left(\frac{u}{v}\right)^{u/2} x^{(u/2)-1}}{\Gamma\left(\frac{u}{2}\right) \Gamma\left(\frac{v}{2}\right) \left[\left(\frac{u}{v}\right)x + 1\right]^{(u+v)/2}}, \quad 0 < x < \infty.$$

Propoziția 12.7.7 *Media și dispersia F -distribuției sunt:*

$$M[X] = \frac{v}{v-2}, \text{ pentru } v > 2,$$

$$D[X] = \frac{2v^2(u+v-2)}{u(v-2)^2(v-4)}, \text{ pentru } v > 4.$$



Graficul densității de probabilitate pentru $u = v = 5$ este dat în figura următoare

Graficul F-distribuției este foarte asemănător cu graficul distribuției hi-pătrat. Cei doi parametri asigură o flexibilitate a formei.

Valorile funcției de repartiție a F-distribuției pot fi calculate cu ajutorul tabelului din Anexa 4. Fie $f_{\alpha,u,v}$ punctul în care vrem să calculăm valoarea funcției de repartiție cu u grade de libertate la numărător și v grade de libertate la numitor. Atunci

$$F_{u,v}(f_{\alpha,u,v}) = \int_{f_{\alpha,u,v}}^{\infty} f(x)dx = \alpha.$$

Exemplul 12.7.8 Pentru $u = 5$ și $v = 10$ să se calculeze $f_{\alpha,u,v} = 3.33$.

În tabel la intersecția liniei $v = 10$ și coloanei $u = 5$ se găsește, în tabelul corespunzător lui $\alpha = 0.05$ valoarea 3.33.

$$F_{5,10}(3.33) = 0.95.$$

Tabelul conține probabilitățile pentru valorile lui α egale cu 0.25, 0.1, 0.05, 0.025, 0.01.

12.7.4 Intervale de încredere pentru diferența proporțiilor a două populații

Presupunem că avem două eșantioane de dimensiuni n_1 și respectiv n_2 extrase din două populații X_1 și X_2 reprezentând numărul de observații care aparțin unei clase care se studiază. Mai mult, presupunem aproximarea distribuției binomiale cu distribuția normală este aplicabilă (populația să aibă măcar 10 elemente), astfel încât estimatorii proporțiilor $\hat{P}_1 = X_1/n_1$ și $\hat{P}_2 = X_2/n_2$ urmează o distribuție normală.

Propoziția 12.7.9 *Statistica*

$$Z = \frac{\hat{P}_1 - \hat{P}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

este distribuită aproximativ normal standard.

Determinăm intervalul de încredere

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha \Leftrightarrow$$

$$P\left(-z_{\alpha/2} \leq \frac{\hat{P}_1 - \hat{P}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

Dacă \hat{p}_1 și \hat{p}_2 sunt proporțiile obținute din observațiile făcute asupra celor două eșantioane, atunci

$$-z_{\alpha/2} \leq \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \leq z_{\alpha/2} \Leftrightarrow$$

$$\hat{p}_1 - \hat{p}_2 - z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \leq p_1 - p_2 \leq$$

$$\hat{p}_1 - \hat{p}_2 + z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

Exercițiul 2 Un cercetător este interesat dacă persoanele care au făcut psihologia sunt capabile să rezolve o problemă care implică o anumită judecată. Cercetătorul este interesat în a estima diferența dintre proporțiile persoanelor din cele două populații care pot rezolva problema. Prima populație are 100 membrii din care 65 au rezolvat problema, iar a doua populație are 110 din care doar 45 au rezolvat problema. Să se construiască un interval de încredere pentru diferența proporțiilor cu un nivel de încredere de 99%.

Intervalul de încredere este $[0.0686441, 0.413174]$. Acest interval nu include 0, este pozitiv, ceea ce înseamnă că în prima populație avem o mai mulți cercetători care pot rezolva problema. \diamond