Curs 12

Intervale de încredere

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

Presupunem că studiem vâscozitatea unei anumite substanţe. Prin studierea unui eşantion s-a constatat că media acestei caracteristici este $\hat{\mu} = \overline{x} = 1000$. Dacă considerăm un alt eşantion este aproape imposibil să obţinem aceeaşi estimare numerică pentru media vâscozităţii. Nu putem spune nimic despre relaţia dintre cele două medii. Problema pe care o punem este următoarea: valoarea reală a vâscozităţii este cuprinsă între 900 şi 1100 sau între 990 şi 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 populaţiei studiate se numeşte *interval de încredere*. Intervalul de încredere constă din:

- *un interval*, obţinut cu ajutorul datelor furnizate de o selecţie,

- *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 semnificaţie $\alpha = 0.05$).

Definiţia 12.0.1 *Se numeşte* interval de încredere *pentru un parametru θ asociat unei populaţii orice interval $I = [a, b]$ pentru care se poate estima probabilitatea ca $\theta \in I$. Dacă α este un număr cuprins între 0 şi 1 şi 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 distribuţie normală

12.1 Intervale de încredere pentru medie în cazul σ cunoscut

Presupunem că realizăm o selecţie populaţie a cărei caracteristică studiată urmează o distribuţie normală, $N[m, \sigma]$, cu σ cunoscut, m necunoscut. Situaţia este mai puţin întâlnită în realitate deoarece în mod normal atât media cât şi dispersia sunt necunoscute. Totuşi vom prezenta în continuare şi acest caz.

12.1.1 Construcţia intervalului de încredere

Fie $x_1, x_2, \ldots, x_n$ valorile variabilelor de selecţie $X_1, X_2, \ldots, X_n$ obţinute dintr-o populaţie care urmează o distribuţie normală, $N[m, \sigma]$, $\sigma > 0$ cunoscut, m necunoscut. Ştim că $Z = \dfrac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} \in N([0,1]$. Din această cauză putem scrie, (evident $z > 0$),

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

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

De îndată ce selecţia a fost realizată şi a fost calculată media de selecţie $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ se obţine intervalul,

$$\left[\overline{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right] \tag{12.1}$$

Suntem tentaţi să spunem că $1-\alpha$ este probabilitatea ca acest interval să cuprindă valoarea exactă a lui m, dar această afirmaţie nu este corectă. Trebuie să ţinem seama de faptul că întervalul de încredere este un interval aleator, el depinde de selecţia făcută, deci extremităţile sale sunt v. a. Prin urmare interpretarea corecta a lui $1-\alpha$ este următoarea: dacă, facem un număr foarte mare de selecţii şi calculăm de fiecare dată întervalul de încredere cu nivelul de încredere $1-\alpha$, atunci $(1-\alpha)\,100\%$ din aceste intervale vor conţine valoarea exactă pentru m.

Observăm că intervalul de încredere pentru m este centrat în estimaţia punctuală $\overline{x}$. Când n creşte se obţine un *interval mai scurt* pentru acelaşi coeficient de încredere. Un interval de încredere mai scurt indică o mai mare încredere în $\overline{x}$ ca estimaţie a lui m.

Exemplul 12.1.1 Punctajele obţinute de studenţi care au promovat examenul de matematică şi care cuantifică cunoştinţele 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 cunoaşte $\sigma = 10.99$. Să se construiască intervalele de încredere pentru medie cu nivelele de încredere de 90%, 95% şi 99%.

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

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

Pentru 90% avem $\alpha = 0.1$ şi

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

Atunci, conform (12.1), intervalul

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

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

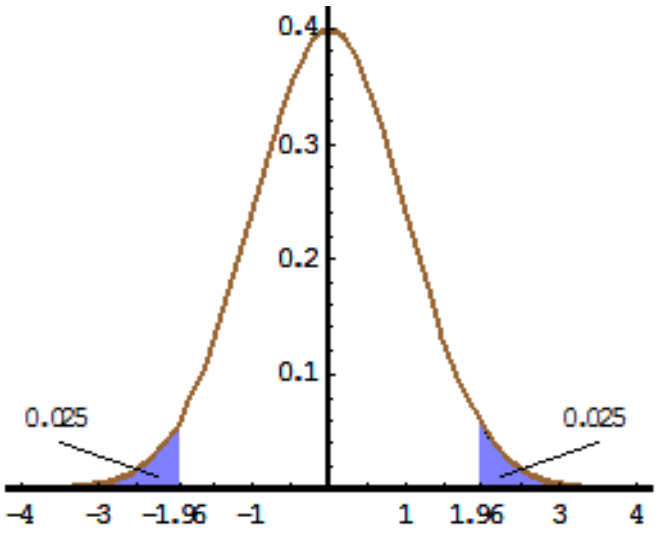
Pentru 95% avem $\alpha = 0.05$ şi $\Phi(-z) = 0.025 \Rightarrow z_{\frac{\alpha}{2}} = 1.9599$. Intervalul de încredere pentru m este

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

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

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

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 $\dfrac{\alpha}{2} = 0.025$, iar la stânga o arie egală cu $1 - \dfrac{\alpha}{2} = 1 - 0.025 = 0.975$.



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

Algoritmul testului Z

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

Pasul 1. Se calculează $\overline{x}$.

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

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

Pasul 4. Se determină intervalul de încredere pentru m

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

Comparăm intervalele obţinute în exemplul de mai sus în funcţie de nivelul de încredere

$(1-\alpha)\cdot 100\%$	$\frac{\alpha}{2}$	$z_{\frac{\alpha}{2}}$	$\left[\overline{x}-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{x}+z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$
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 proporţională cu nivelul de încredere.

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

Interpretarea corectă este: dacă facem un număr mare de selecţii şi de fiecare dată calculăm intervalul de încredere pentru medie cu nivelul de încredere de 95%, atunci în 95% din aceste intervale vor conţine valoarea corectă a mediei. Deci metoda folosită ne permite să obţinem intervale pentru medie care vor conţine în 95% din cazuri valoatrea 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 decît în cazul nivelului de 95%. Dacă dimensiunea eşantionului şi abaterea medie pătratică sunt păstrate constante, atunci un nivel mai înalt de încredere atrage un interval de încredere mai mare.

Lungimea intervaluilui de încredere este o măsură a preciziei estimării. Din cele prezentate rezultă că precizia este invers proporţională cu nivelul de încredere. Este preferabil să obţinem un interval de încredere cît mai scurt pentru o problema pusă, dar cu un nivel de încredere adecvat. Un mod de a atinge acest scop este alegerea dimensiunii eşantionului astfel încât cu ajutorul acestei selecţii să putem obţine un interval de încredere de lungime specificată şi 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 informaţie relativă la valoarea medie de forma că aceasta nu este limitată superior, atunci intervalul de încredere devine de forma $\left(\overline{x}-z_\alpha\frac{\sigma}{\sqrt{n}}, \infty\right)$ şi este un interval de încredere unilateral.

În acest caz

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

Notăm cu z_α valoarea obţinută din relaţia $\Phi(-z) = \alpha$.

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

12.2 Intervale de încredere pentru medie în cazul σ necunoscut

Presupunem că populaţia studiată are o distribuţie normală cu media m şi dispersia σ^2 necunoscute. Facem o selecţie de dimensiune n. Fie $x_1, x_2, \ldots, x_n$ valorile variabilelor de

selecţie $X_1, X_2, \ldots, X_n$. Putem calcula media de selecţie $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ şi dispersia de selecţie modificată $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$.

Vrem să calculăm un interval de încredere pentru m. Dacă dispersia este cunoscută, ştim că $Z = \dfrac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}}$ urmează o distribuţie normală. Dacă σ este necunoscut o procedură normală este de a înlocui σ cu s. Statistica devine acum $T = \dfrac{\overline{X} - m}{\frac{s}{\sqrt{n}}}$. O întrebare logică care se pune este următoarea: care este efectul înlocuirii lui σ cu s asupra distribuţiei statisticii T? Dacă n este suficient de mare, răspunsul la această întrebare este: efectul este "destul de mic" şi putem considera că urmează o distribuţie normală standard. În general n trebuie să fie cel puţin 40. Teorema limită centrală are loc pentru $n \geq 30$, dar mărirea eşantionului recomandată este la cel puţin 40, deoarece înlocuirea lui σ cu s în Z conduce la modificări suplimentare ale distribuţiei.

În acest caz intervalul de încredere se construieşte astfel:

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

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

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,

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

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

Testul Student

Presupunem că populaţia studiată are o distribuţie normală cu media m şi abaterea medie pătratică σ necunoscute. Facem o selecţie de dimensiune n, n mic. Vrem să calculăm un interval de încredere pentru m.

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

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

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

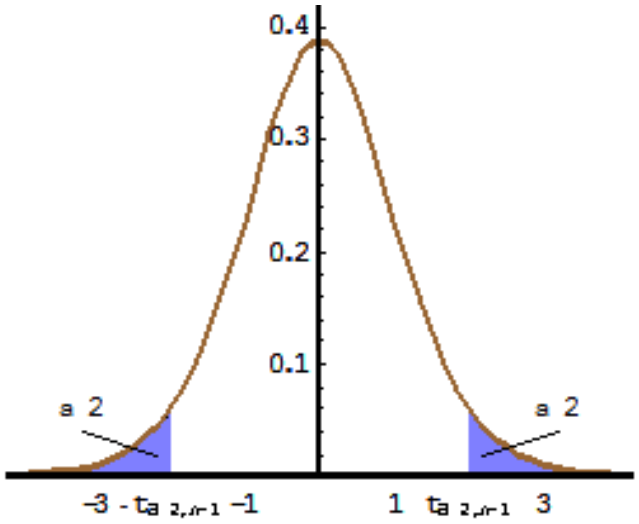
unde $f\left(x\right)$ este densitatea de probabilitate a distribuţiei Student.

Pentru valorile negative se foloseşte faptul că

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

Deoarece distribuţia Student este simetrică, avem $t_{1-\alpha,n} = -t_{\alpha,n}$, ceea ce înseamnă că în partea dreaptă a lui $t_{\alpha,n}$, dar şi 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 $\dfrac{\alpha}{2}$.



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

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

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

Rezultă că intervalul

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

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 $x_1, x_2, \ldots, x_n$ o selecţie de variabile de selecţie $X_1, X_2, \ldots, X_n$ i.i.d. dintr-o populaţie normală cu media m şi dispersia σ^2 necunoscute.

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

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

Pasul 3. Pentru un coeficient de încredere prescris $(1-\alpha)\cdot 100\%$ se determină din tabelul funcţiei de repartiţie 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,

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

Exemplul 12.2.2 Considerăm un eşantion de 36 studenţi care au obţinut 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% şi 99% încredere pentru media punctajelor obţinute.

Rezolvare. Folosim testul Z.

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

Statistica T este distribuită Student cu 35 grade de libertate (a se consulta anexa 2 cu tabelul valorilor funcţiei de repartiţie Student în funcţie 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 $[68.5081; 79.0475]$.

Şi în acest caz putem construi intervale de încredere de forma $\left(\overline{x} - t_{\alpha,n-1}\frac{s}{\sqrt{n}}, \infty\right)$ sau $\left(-\infty, \overline{x} + t_{\alpha,n-1}\frac{s}{\sqrt{n}}\right)$ 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 populaţia este modelată de o distribuţie normală putem aplica intervalele descrise în continuare.

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

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

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

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

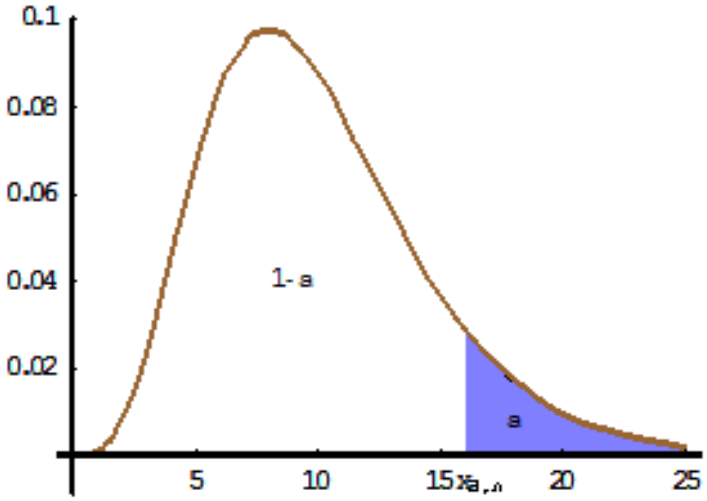
$$P\left(\chi^2 \leq x_{\alpha,n}\right) = \int_{-\infty}^{x_{\alpha,n}} f(t)dt = 1 - \alpha, \tag{12.3}$$

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

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 funcţiei de repartiţie hi-pătrat observăm că pe prima coloană sunt trecute gradele de libertate iar pe linie sunt trecute valorile lui α. De exemplu, pentru $n = 10$ şi $\alpha = 0.05$ obţinem $x_{0.05,10} = 18.31$, 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 $x_{1-0.05,10} = 3.94$. ◊

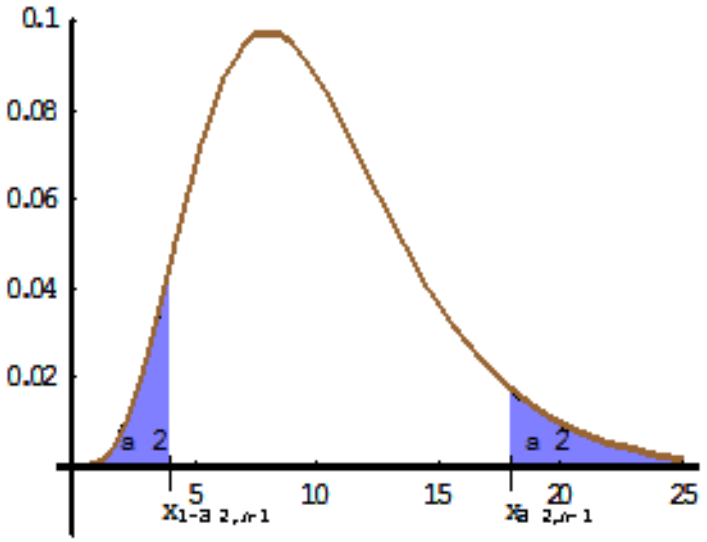


Pentru construcţia intervalului de încredere pentru dispersie se foloseşte statistica $\chi^2_{n-1} = \dfrac{(n-1)s^2}{\sigma^2} \in \chi^2(n-1)$. Pentru $\alpha \in (0,1)$ dat determinăm $x_1(\alpha)$ şi $x_2(\alpha)$ astfel încât

$$P\left(x_1(\alpha) \leq \chi^2_{n-1} \leq x_2(\alpha)\right) = 1-\alpha. \tag{12.4}$$

Numerele $x_1(\alpha)$ şi $x_2(\alpha)$ nu sunt unic determinate şi de obicei se aleg astfel încât $x_1(\alpha) = x_{1-\alpha/2,n-1}$ şi $P\left(\chi^2_{n-1} \leq x_{1-\alpha/2,n-1}\right) = \dfrac{\alpha}{2}$, iar $x_2(\alpha) = x_{\alpha/2,n-1}$ şi $P\left(\chi^2_{n-1} \geq x_{\alpha/2,n-1}\right) = \dfrac{\alpha}{2}$.

Semnificaţia valorilor $x_{1-\alpha/2,n-1}$ şi $x_{\alpha/2,n-1}$ poate fi vazută în figura următoare.



Înlocuind $\chi^2_{n-1} = \dfrac{(n-1)s^2}{\sigma^2}$ în (12.4) şi obţinem:

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

Relaţia (12.5) poate fi rearanjată astfel:

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

Am obţinut astfel un interval de încredere pentru dispersie.

Algoritmul de determinare a intervalului de încredere pentru dispersie

Fie $x_1, x_2, \ldots, x_n$ o selecţie de valori pentru variabilele de selecţie $X_1, X_2, \ldots, X_n$ i.i.d. dintr-o populaţie normală cu media m şi dispersia σ^2 necunoscute.

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

Pasul 2. Se alege statistica $\chi^2_{n-1} = \dfrac{(n-1)s^2}{\sigma^2}$ despre care se ştie că urmează o distribuţie 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 funcţiei de repartiţie χ^2 cu $n-1$ grade de libertate sau cu ajutorul softurilor Matlab sau Mathematica, numerele $x_{1-\alpha/2,n-1}$ şi $x_{\alpha/2,n-1}$ astfel încât

$$P\left(\chi^2_{n-1} \le x_{1-\alpha/2,n-1}\right) = \frac{\alpha}{2} \quad \text{şi} \quad P\left(\chi^2_{n-1} \le x_{\alpha/2,n-1}\right) = 1 - \frac{\alpha}{2}.$$

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

Observaţia 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* $\left(-\infty, \dfrac{(n-1)s^2}{x_{1-\alpha,n-1}}\right)$, *respectiv* $\left(\dfrac{(n-1)s^2}{x_{\alpha,n-1}}, \infty\right)$.

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 $\chi^2_{82} = \dfrac{(n-1)s^2}{\sigma^2} = \dfrac{82s^2}{\sigma^2}$.

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

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

Pentru nivelul de încredere de 99% avem $x_{0.995,83} = 52.7674$ şi $x_{0.005,83} = 118.726$. 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 şi dispersia necunoscută. Se face o selectie de volum 6: {-1, 4, 4, 1,3,1}. Se cere un interval de estimaţie pentru dispersie cu nivel de încredere de 90%.

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

$$s^2 = \frac{1}{3}\left((-1-3)^2 + (4-3)^2 + (4-3)^2 + (1-3)^2 + \right.$$
$$\left. +(3-3)^2 + (1-3)^2\right) = \frac{26}{3} = 8.666\,7$$

Considerăm statistica $\chi^2_5 = \dfrac{(n-1)s^2}{\sigma^2} = \dfrac{5s^2}{\sigma^2}$.

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

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 proporţii

Pentru o populaţie a cărei membri pot fi clasificaţi în funcţie de o anumită caracteristică în două categorii: fie p probabilitatea de a aparţine unei categorii, numit succes şi $1-p$ probabilitatea de a aparţine celeilalte categorii, numită eşec. Parametrul p poată denumirea de proporţia populaţiei şi 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 $\hat{P} = \dfrac{X}{n}$ un estimator punctual al lui p. Reamintim că n şi p sunt parametrii unei distribuţii binomiale. Mai mult, statistica P urmează o distribuţie normală cu media p şi dispersia $\dfrac{p(1-p)}{n}$ dacă p nu este aproape de 0 sau 1 şi 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 încăt $np \geq 5$, $n(1-p) \geq 5$, atunci*

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

urmează aproximativ o distribuţie normală standard.

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

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

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 $\Phi(-z_{\frac{\alpha}{2}}) = \dfrac{\alpha}{2}$.

Această relaţie 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 $\sqrt{\dfrac{p(1-p)}{n}}$ se numeşte *eroarea standard* a estimatorului punctual $\hat{P}$. Deoarece marginile intervalului conţin p care este necunoscut, o soluţie satisfăcătoare este înlocuirea sa cu $\hat{P}$. Astfel obţinem

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

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

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

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

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

$$\hat{P} = \frac{10}{85} = 0.117\,655.$$

Eroarea standard a estimatorului punctual este

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

Nu este foarte clar cum interpretăm această valoare de 0.034 946. Este această estimare foarte exactă, exactă sau nu?

Construim un interval de încredere cu un nivel de încredere de 95% bazat pe valoarea observată 0.117 655.

$$\begin{gathered}\hat{P} - z_{\frac{\alpha}{2}}\sqrt{\tfrac{\hat{P}(1-\hat{P})}{n}} \le p \le \hat{P} + z_{\frac{\alpha}{2}}\sqrt{\tfrac{\hat{P}(1-\hat{P})}{n}},\\ 0.117\,6 - 1.96\sqrt{\tfrac{0.117\,6\cdot 0.88}{85}} \le p \le 0.117\,6 + 1.96\sqrt{\tfrac{0.117\,6\cdot 0.88}{85}},\\ 0.04915 \le p \le 0.186141.\end{gathered}$$

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 eşantionului

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 $\hat{P} = \frac{10}{85} = 0.117\,655$ diferă de valoarea exactă a lui p cu mai puţin de

$$1.95996\sqrt{\frac{\frac{10}{85}(1-\frac{10}{85})}{85}} = 0.069\,084.$$

În situaţiile în care dimensiunea eşantionului 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 $E = z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$ şi rezolvăm ecuaţia în raport cu n obţinem ca dimensiune a eşantionului

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

O estimare a lui p este necesară. Pentru aceasta se poate lua un eşantion preliminar, se calculează $\hat{P}$ şi apoi folosind relaţia (12.7) putem determina câte observaţii mai trebuie făcute pentru a obţine pentru p o estimare satisfăcătoare.

O altă soluţie este de a ţine seama că $p(1-p) \leq \frac{1}{4}$ şi atunci

$$n = \left[\frac{1}{4}\left(\frac{z_{\frac{\alpha}{2}}}{E}\right)^2\right] + 1. \tag{12.8}$$

Exemplul 12.4.3 Reluăm Exemplul 12.4.2. Cât de mare trebuie luat eşantionul 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

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

Dacă vrem să fim cel puţin 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)

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

Observăm că dacă avem o informaţie privind valoarea lui p, chiar dintr-un eşantion preliminar, obţinem o dimensiune mai mică a eşantionului menţinând precizia estimării şi nivelul de încredere. ♢

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

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

12.5 Predicţia

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

Fie $x_1, x_2, \ldots, x_n$ o selecţie de valori pentru variabile de selecţie $X_1, X_2, \ldots, X_n$ i.i.d. dintr-o populaţie normală cu media m şi dispersia σ^2. Dorim să prevedem valoarea variabilei de selecţie X_{n+1} la o singură observaţie viitoare.

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

Predicţia $X_{n+1} - \overline{x}$ este normal distribuită şi atunci $Z = \dfrac{X_{n+1} - \overline{X}}{\sigma\sqrt{1 + \frac{1}{n}}}$ are o distribuţie normală standard.

Dacă înlocuim σ prin s obţinem $T = \dfrac{X_{n+1} - \overline{x}}{s\sqrt{1 + \frac{1}{n}}}$ care are o distribuţie Student cu $n-1$ grade de libertate. Obţinem astfel intervalele de predicţie cu $(1-\alpha)100\%$ nivel de încredere

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

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

şi respectiv

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

Exemplul 12.5.1 Se studiază o caracteristică pe $n = 22$ unităţi statistice şi se obţin 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 predicţie pentru cea de a 23-a valoare şi intervalul de încredere pentru media valorilor obţinute, nivelul de încredere fiind de 95%, iar caracteristica urmează o repartiţie normală.

Rezolvare. Calculăm şi obţinem $\overline{x} = 13.71$ şi $s = 3.55$. Intervalul de încredere pentru medie este

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

Intervalul de predicţie cu 95% încredere este

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

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

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

12.6 Intervale de toleranţă pentru caracteristici ce urmează o distribuţie normală

Definiţia 12.6.1 Intervalul de toleranţă *este un interval statistic în care, cu un nivel de încredere dat, caracteristica studiată a unei populaţii ia valori cu o probabilitate de acoperire specificată.*

Capetele intervalului de toleranţă se numesc *limite de toleranţă.* În cazul intervalului de toleranţă se cunoaşte distribuţia pe care o urmează caracteristica şi, eventual, se cunosc parametrii distribuţiei şi se determină intervalul cu o probabilitate de acoperire şi cu un anumit nivel de încredere. Diferă de intervalul de încredere deoarece acesta din urmă furnizează limite pentru un anumit parametru al populaţiei, necunoscut, de ex. medie, dispersie numai cu un anumit nivel de încredere.

În inginerie se specifică, de obicei, limitele de toleranţă, de acceptibilitate ale caracteristicii unui produs şi nu reflectă întotdeauna valoarea obţinută prin măsurători.

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

Dacă media şi dispersia sunt cunoscute, intervalul de toleranţă se defineşte de forma $\left[m - z_{\frac{\alpha}{2}}\sigma, m + z_{\frac{\alpha}{2}}\sigma\right]$, în funcţie 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 frecvenţa procesoarelor.

Dacă m şi σ nu sunt cunoscuţi, putem folosi $\overline{x}$ şi s pentru a calcula intervalul de toleranţă, $\left[\overline{x} - z_{\frac{\alpha}{2}}s, \overline{x} + z_{\frac{\alpha}{2}}s\right]$.

Este de aşteptat ca în acest caz, datorită introducerii lui $\overline{x}$ şi s, probabilitatea de acoperire să fie mai mică decât $1 - \alpha$ cu un nivel de încredere de $(1-\alpha)\cdot 100\%$ Soluţia 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 toleranţă care să aibă probabilitatea de acoperire p pentru o caracteristică care urmează o distribuţie normală, cu un nivel de încredere de $(1-\alpha)\cdot 100\%$ este $[\overline{x} - ks, \overline{x} + ks]$, unde k este un factor de toleranţă.

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

Formula factorului de toleranţă dată de Howe este

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

Algoritmul de calcul al factorului de toleranţă, în funcţie de proporţia p şi de probabilitatea γ, este:

1. Se determină $z_{(1+p)/2}$ astfel încât dacă Z urmează o distribuţie normală, $P\left(Z \leq z_{(1+p)/2}\right) = (1+p)/2$;

2. Se determină $\chi_{\alpha,n-1}$ astfel că dacă χ urmează o distribuţie hi-pătrat cu $n-1$ grade de libertate, $P\left(\chi \leq \chi_{\alpha,n-1}\right) = \alpha$;

3. Se înlocuiesc în formula lui k.

12.7 Intervale de încredere pentru caracteristicile a două populaţii care urmează o distribuţie normală

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

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

12.7.1 Intervale de încredere pentru diferenţa mediilor a două populaţii ale căror dispersii sunt cunoscute

Construim intervale de încredere, cu nivelul de încredere $(1-\alpha)\cdot\cdot 100\%$, pentru diferenţa mediilor a două populaţii care urmează o distribuţie normală şi ale căror dispersii sunt cunoscute.

Facem următoarele ipoteze:

1. Se consideră două populaţii independente care urmează o distribuţie normală cu dispersiile cunoscute;

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

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

Un estimator logic pentru diferenţa mediilor, $m_1 - m_2$ este diferenţa dintre mediile statistice ale celor două eşantioane, $\overline{X}_1 - \overline{X}_2$, unde $\overline{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1i}$ şi $\overline{X}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2i}$. Folosind proprietăţile mediei şi dispersiei obţinem că

$$M\left[\overline{X}_1 - \overline{X}_2\right] = M\left[\overline{X}_1\right] - M\left[\overline{X}_2\right] = m_1 - m_2,$$
$$D\left[\overline{X}_1 - \overline{X}_2\right] = D\left[\overline{X}_1\right] + D\left[\overline{X}_2\right] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Ţinând seama de presupunerile făcute şi de rezultatele anterioare putem afirma

Propoziţia 12.7.1 *Statistica*

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

urmează o distribuţie normală standard.

Observaţia 12.7.2 *Dacă sunt îndeplinite condiţiile Teoremei limită centrală pentru cele două populaţii, atunci rezultatul Propoziţiei 12.7.1 se păstrează.*

Ţinând seama de rezultatul Propoziţiei 12.7.1 putem scrie

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

$$P\left(-z_{\alpha/2} \le \frac{\overline{X}_1 - \overline{X}_2 - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \le z_{\alpha/2}\right) = 1-\alpha \Leftrightarrow$$

$$P\left(\overline{X}_1 - \overline{X}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le m_1 - m_2 \le \overline{X}_1 - \overline{X}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = 1-\alpha.$$

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

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

Exerciţiul 1 Fie variabilele de selecţie $X_{11}, X_{12}, \ldots, X_{1n_1}$ i.i.d. care urmează o legea normală $N(m_1, \sigma_1)$ şi reprezintă încasările în mii lei ale unui lanţ de magazine din orasul A şi variabile de selecţie $X_{21}, X_{22}, \ldots, X_{2n_2}$ i.i.r. care urmează o legea normală $N(m_2, \sigma_2)$ şi reprezintă încasările în mii lei ale unui alt lanţ de magazine din oraşul B. Presupunem că cele două selecţii sunt independente. S-au efectuat două sondaje, respectiv pentru X_1 şi X_2 şi s-au obţinut următoarele date: pentru X_1 : 226.5, 224.1, 218.6, 220.1, 228.8, 229.6, 222.5 şi 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 diferenţa mediilor, $m_1 - m_2$ dacă $\sigma_1 = 2.5, \sigma_2 = 3$, cunoscute.

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

Observaţia 12.7.3 *Şi în acest caz putem construi intervale de încredere de forma*

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

pentru diferenţa mediilor.

Alegerea dimensiunii eşantionului

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

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

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 diferenţa mediilor a două populaţii ale căror dispersii sunt necunoscute

Extindem rezultatele anterioare pentru cazul în care cele două populaţii urmează o distribuţie normală şi dispersiile sunt necunoscute. Dacă volumele eşantioanelor n_1 şi n_2 depăşesc valoarea de 40, rezultatul anterior poate fi utilizat. Dacă eşantioanele sunt mai mici atunci construcţia intervalului de încredere se bazează pe distribuţia Student.

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

Două situaţii 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 şi diferite.

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

Se consideră $x_{11}, x_{12}, \ldots, x_{1n_1}$ o selecţie de n_1 valori pentru variabile de selecţie $X_{11}, X_{12}, \ldots, X_{1n_1}$ i.i.d. din prima populaţie şi $x_{21}, x_{22}, \ldots, x_{2n_2}$ o selecţie de n_2 valori pentru variabile de selecţie $X_{21}, X_{22}, \ldots, X_{2n_2}$ i.i.d. din a doua populaţie. Fie mediile de selecţie $\overline{X}_1 = \dfrac{1}{n_1}\sum_{i=1}^{n_1} X_{1i}$ şi $\overline{X}_2 = \dfrac{1}{n_2}\sum_{i=1}^{n_2} X_{2i}$ şi dispersiile de selecţie $S_1^2 = \dfrac{1}{n_1 - 1}\sum_{i=1}^{n_1}\left(X_{1i} - \overline{X}_1\right)^2$ şi $S_2^2 = \dfrac{1}{n_2 - 1}\sum_{i=1}^{n_1}\left(X_{2i} - \overline{X}_2\right)^2$.

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

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

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

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

Obsevăm 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 combinaţie liniară de cele două dipersii de selecţie în care λ depinde doar de dimensiunea eşantioanelor, n_1 şi n_2. Dacă $n_1 = n_2 = n$ atunci S_p^2 este media aritmetică a celor două dispersii de selecţie.

Ştim că $Z = \dfrac{\overline{X}_1 - \overline{X}_2 - (m_1 - m_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ urmează o distribuţie normală standard. Înlocuind σ cu S_p obţinem următorul rezultat:

Propoziţia 12.7.4 *Statistica*

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

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

Atunci

$$P\left(-t_{\alpha/2, n_1+n_2-2} \leq T \leq t_{\alpha/2, n_1+n_2-2}\right) = 1 - \alpha \Leftrightarrow$$

$$P\left(-t_{\alpha/2, n_1+n_2-2} \leq \frac{\overline{X}_1 - \overline{X}_2 - (m_1 - m_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq t_{\alpha/2,}\right) = 1 - \alpha \Leftrightarrow$$

$$P\left(\overline{X}_1 - \overline{X}_2 - t_{\alpha/2, n_1+n_2-2} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq m_1 - m_2 \leq \right.$$

$$\left. \leq \overline{X}_1 - \overline{X}_2 + t_{\alpha/2, n_1+n_2-2} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right) = 1 - \alpha.$$

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

$$\left[\overline{x}_1 - \overline{x}_2 - t_{\alpha/2, n_1+n_2-2} s_p\sqrt{\tfrac{1}{n_1} + \tfrac{1}{n_2}}, \overline{x}_1 - \overline{x}_2 + t_{\alpha/2, n_1+n_2-2} s_p\sqrt{\tfrac{1}{n_1} + \tfrac{1}{n_2}}\right].$$

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

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

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

urmează aproximativ o distribuţie 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 rotunjeşte prin lipsă la cel mai apropiat întreg.

Intervalul de încredere, în acest caz, este

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

12.7.3 Intervale de încredere pentru raportul dispersiilor a două populaţii

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

Facem următoarele ipoteze:

1. Se consideră două populaţii independente care urmează o distribuţie normală;

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

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

Fie S_1^2 şi S_2^2 dispersiile de selecţie ale celor două populaţii.

Propoziţia 12.7.5 *Statistica*

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

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

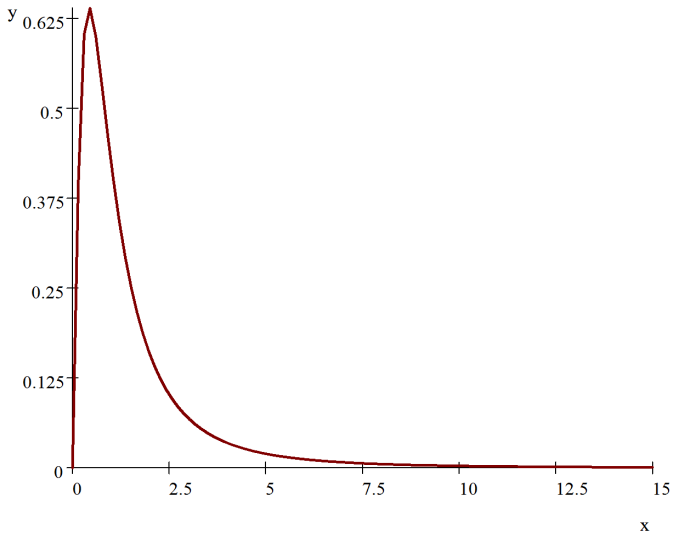
Definiţia 12.7.6 *Densitatea de probabilitate a F-distribuţiei cu u grade de libertate la numărător şi v grade de libertate la numitor este*

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

Propoziţia 12.7.7 *Media şi dispersia F-distribuţiei sunt:*

$M[X] = \dfrac{v}{v-2}$, *pentru* $v > 2$,

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



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

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

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

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

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

În tabel la intersecţia liniei $v = 10$ şi coloanei $u = 5$ se găseşte, în tabelul corespunzător lui $\alpha = 0.05$ valoarea 3.33.

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

Tabelul conţine probabilităţile pentru valorile lui α egale cu 0.25, 0.1, 0.05, 0.025, 0.01.

12.7.4 Intervale de încredere pentru diferenţa proporţiilor a două populaţii

Presupunem că avem două eşantioane de dimensiuni n_1 şi respectiv n_2 extrase din două populaţii X_1 şi X_2 reprezentând numărul de observaţii care aparţin unei clase care se studiază. Mai mult, presupunem aproximarea distribuţiei binomiale cu distribuţia normală este aplicabilă (populaţia să aibă măcar 10 elemente), astfel încât estimatorii proporţiilor $\hat{P}_1 = X_1/n_1$ şi $\hat{P}_2 = X_2/n_2$ urmează o distribuţie normală.

Propoziţia 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

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

Dacă $\hat{p}_1$ şi $\hat{p}_2$ sunt proporţiile obţinute din observaţiile făcute asupra celor două eşantioane, atunci

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

Exerciţiul 2 Un cercetăror 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 diferenţa dintre proporţiile persoanelor din cele două populaţii care pot rezolva problema. Prima populaţie are 100 membrii din care 65 au rezolvat problema, iar a doua populatie are 110 din care doar 45 au rezolvat problema. Să se construiască un interval de încredere pentru diferenţa proporţiilor 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 populaţie avem o mai mulţi cercetători care pot rezolva problema. $\Diamond$