10 Estimarea parametrilor: intervale de încredere

Intervalele de încredere pentru un parametru necunoscut θ al unei distribuţii (spre exemplu pentru media unei populaţii) sunt intervale (θ_1, θ_2) ce conţin parametrul θ, nu cu certitudine, si cu anumită probabilitate γ (spre exemplu cu probabilitate $\gamma = 95\%$ sau $\gamma = 99\%$). Valoarea lui γ se numeşte **nivel de încredere/nivel de semnificaţie** iar θ_1 şi θ_2 se numesc **limita de încredere inferioară şi superioară**.

Un astfel de interval (θ_1, θ_2) este calculat folosind o selecţie din populaţie. Valoarea $\gamma = 95\%$ înseamnă că probabilitatea ca intervalul (θ_1, θ_2) să nu-l conţină pe θ este $1 - \gamma = 5\% = 0.05$, adică aproximativ unul din 20 de intervale astfel determinate nu îl va conţine pe θ. Cu cât valoarea lui γ este mai apropiată de 1, cu atât este mai mică probabilitatea erorii $1-\gamma$ ca intervalul de încredere să nu-l conţină parametrul θ, dar în acelaşi timp lungimea $\theta_2 - \theta_1$ a intervalului de încredere devine mai mare (tinde la infinit pentru $\gamma \to 1$).

Intervalele de încredere sunt mai utile decît estimatorii punctuali ai parametrilor necunoscuţi, deoarece având intervalul de încredere putem considera ca estimator mijlocul $\frac{\theta_1+\theta_2}{2}$ al intervalului (în acest caz, probabilitatea ca θ să difere faţă de estimatorul considerat cu mai puţin de $\frac{\theta_2-\theta_1}{2}$ este egală cu γ).

Pentru a determina capetele θ_1 şi θ_2 ale intervalului de încredere, considerăm un eşantion $x_1, \ldots, x_n$ din populaţia dată. Aceste valori sunt valorile observate ale unei variabile aleatoare X (reprezentând populaţia studiată), dar ele pot fi privite ca valori ale n variabile aleatoare independente $X_1, \ldots, X_n$ având aceeaşi distribuţie cu variabila aleatoare X.

Atunci $\theta_1 = \theta_1(x_1, \ldots, x_n)$ şi $\theta_2 = \theta_2(x_1, \ldots, x_n)$ sunt valorile observate ale variabilelor aleatoare $\Theta_1 = \Theta_1(X_1, \ldots, X_n)$ şi $\Theta_2 = \Theta_2(X_1, \ldots, X_n)$. Condiţia ca intervalul (θ_1, θ_2) să-l conţină parametrul θ cu probabilitate γ se poate scrie deci sub forma

$$P(\Theta_1 \le \theta \le \Theta_2) = \gamma. \tag{57}$$

Pentru a determina variabilele aleatoare Θ_1 şi Θ_2 folosim următoarea.

Teorema 10.1 (Suma de variabile aleatoare normale independente) *Fie $X_1, \ldots, X_n$ variabile aleatoare normale independente, având fiecare medie μ şi dispersie σ^2. Atunci au loc următoarele.*

a) *Suma $X_1 + \ldots + X_n$ este o variabilă aleatoare normală cu medie $n\mu$ şi dispersie $n\sigma^2$.*

b) *Variabila aleatoare*

$$\overline{X} = \frac{X_1 + \ldots + X_n}{n} \tag{58}$$

este o variabilă aleatoare normală cu medie μ şi dispersie σ^2/n.

c) *Variabila aleatoare*

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{59}$$

este o variabilă aleatoare normală standard (cu medie 0 şi dispersie 1).

Demonstraţie. a) Este suficient să demonstrăm afirmaţia din enunţ în cazul $n = 2$ (demonstraţia poate fi apoi completată prin inducţie matematică).

Se poate arăta că deoarece variabilele aleatoare normale X_1 şi X_2 sunt independente, variabila aleatoare X_1+X_2 este de asemenea o variabilă aleatoare normală.

Avem

$$M(X_1 + X_2) = M(X_1) + M(X_2) = \mu + \mu = 2\mu,$$

şi deci variabila aleatoare $X_1 + X_2$ are medie 2μ.

De asemenea, deoarece variabilele aleatoare X_1 şi X_2 sunt independente, se poate arăta că dispersia sumei $X_1 + X_2$ este egală cu suma dispersiilor variabilelor aleatoare X_1 şi X_2, adică

$$\sigma^2(X_1 + X_2) = \sigma^2(X_1) + \sigma^2(X_2) = \sigma^2 + \sigma^2 = 2\sigma^2,$$

şi deci variabila aleatoare $X_1 + X_2$ are dispersie $2\sigma^2$.

Am arătat deci că $X_1 + X_2 \in \mathcal{N}(2\mu, 2\sigma^2)$, încheiând astfel prima parte a demonstraţiei.

b) Conform primei părţi a demonstraţie avem că $X_1 + \ldots + X_n \in \mathcal{N}(n\mu, n\sigma^2)$ este o variabilă aleatoare normală cu medie $n\mu$ şi dispersie $n\sigma^2$. Folosind Teorema 4.4 rezultă că variabila aleatoare $\overline{X} = \frac{X_1+\ldots+X_n}{n} \in \mathcal{N}\left(\frac{n\mu}{n}, \frac{n\sigma^2}{n^2}\right) = \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ este o variabilă aleatoare normală cu medie μ şi dispersie $\frac{\sigma^2}{n}$.

c) Cum $\overline{X} \in \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$, folosind din nou Teorema 4.4 obţinem că variabila aleatoare standardizată $Z = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \in \mathcal{N}(0,1)$ este o variabilă aleatoare normală cu medie 0 şi dispersie 1. ■

Folosind teorema anterioară, putem acum determina intervale de încredere pentru media şi dispersia unei populaţii normale după cum urmează.

10.1 Intervale de încredere pentru media μ unei populaţii normale cu dispersie σ^2 cunoscută

Dacă populaţia $X \in \mathcal{N}(\mu, \sigma^2)$ este normală cu medie μ (necunoscută) şi dispersie σ^2 (cunoscută), conform Teoremei 10.1 rezultă că variabila aleatoare

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \in \mathcal{N}(0,1)$$

este o variabilă aleatoare normală standard.

Pentru un nivel de încredere γ fixat, folosind Anexa 2 determinăm valoarea constantei c astfel încât

$$\gamma = P(-c \le Z \le c) = \Phi(c) - \Phi(-c).$$

Relaţia anterioară se poate scrie sub forma

$$P\left(-c \le \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \le c\right) = \gamma,$$

sau echivalent (rezolvând dubla inegalitate în raport cu μ)

$$P\left(\overline{X} - c\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + c\frac{\sigma}{\sqrt{n}}\right) = \gamma.$$

Considerând $\Theta_1 = \overline{X} - c\frac{\sigma}{\sqrt{n}}$ şi $\Theta_2 = \overline{X} + c\frac{\sigma}{\sqrt{n}}$, relaţia anterioară se poate scrie echivalent sub forma

$$P(\Theta_1 \le \mu \le \Theta_2) = \gamma.$$

Înlocuind variabilele aleatoare $X_1, \ldots, X_n$ prin valorile observate $x_1, \ldots, x_n$ ale eşantionului obţinem că un interval de încredere pentru media necunoscută μ a populaţiei este

$$(\theta_1, \theta_2) = \left(\overline{x} - c\frac{\sigma}{\sqrt{n}}, \overline{x} + c\frac{\sigma}{\sqrt{n}}\right) \tag{60}$$

Exemplul 10.2 *Să se determine un interval de* 95% *încredere pentru media necunoscută* μ *a unei populaţii normale cu dispersie cunoscută* $\sigma^2 = 9$*, folosind un eşantion de volum* $n = 100$ *cu medie* $\overline{x} = 5$.

Folosind Anexa 2 determinăm valoarea c *corespunzătoare probabilităţii* $\gamma = 0.95$ *astfel încât*

$$\Phi(c) - \Phi(-c) = \gamma = 0.95,$$

şi obţinem $c = 1.960$.

Folosind formula (60) determinăm intervalul de încredere pentru media necunoscută μ *a populaţiei*

$$(\theta_1, \theta_2) = \left(\overline{x} - c\frac{\sigma}{\sqrt{n}}, \overline{x} + c\frac{\sigma}{\sqrt{n}}\right) = (\theta_1, \theta_2) = \left(5 - 1.960\frac{3}{\sqrt{100}}, 5 + 1.960\frac{3}{\sqrt{100}}\right) = (4.412, 5.588).$$

O altă problemă în determinarea unui interval de încredere este legată de alegerea volumului selecţiei, ca în exemplul următor.

Exemplul 10.3 *Cât de mare trebuie ales volumul* n *al selecţiei în exemplu anterior dacă se doreşte obţinerea unui interval de* 95% *încredere de lungime cel mult* $L = 0.4$*?*

Intervalul de încredere pentru media μ *a populaţiei din formula (60) are lungimea*

$$\theta_2 - \theta_1 = \left(\overline{x} + c\frac{\sigma}{\sqrt{n}}\right) - \left(\overline{x} - c\frac{\sigma}{\sqrt{n}}\right) = 2c\frac{\sigma}{\sqrt{n}}.$$

Egalând această expresie cu valoarea L dorită, obţinem

$$L = 2c\frac{\sigma}{\sqrt{n}},$$

de unde rezolvând în raport cu n se obţine

$$n = \left(\frac{2c\sigma}{L}\right)^2 = \left(\frac{2 \cdot 1.960 \cdot 3}{0.4}\right)^2 = 864.36.$$

Pentru a obţine intervalul de încredere dorit, trebuie să alegem un volum de selecţie $n = 865$.

Observaţia 10.4 *Se poate arăta că pentru un nivel de încredere $\gamma \to 1$, lungimea intervalului de încredere corespunzător tinde către infinit $\theta_2 - \theta_1 \to \infty$.*

10.2 Intervale de încredere pentru media μ unei populaţii normale cu dispersie σ^2 necunoscută

În cazul în care dispersia σ^2 a populaţiei este necunoscută, nu mai putem utiliza formula (60) pentru a determina un interval de încredere pentru media necunoscută μ a populaţiei.

Deoarece valoarea abaterii pătratice medii σ din această formulă este necunoscută, înlocuim pe σ prin estimatorul

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} \tag{61}$$

al lui σ.

Fâcând însă această înlocuire, variabila aleatoare corespunzătoare

$$Z = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$$

nu mai este o variabilă aleatoare normală (este o variabilă aleatoare $T/Student$ cu $n-1$ grade de libertate, având funcţia de distribuţie $F(x)$ tabelată în Anexa 3), şi deci valoarea c trebuie determinată folosind acest tabel astfel încât

$$F(c) - F(-c) = \gamma.$$

Ca şi în cazul distribuţiei normale standard, distribuţia Student are o funcţie de densitate simetrică faţă de origine, şi deci $F(-x) = 1 - F(x)$ pentru orice $x \in \mathbb{R}$. Înlocuind în relaţia ant6erioară $F(-c)$ prin $1 - F(c)$ obţinem echivalent

$$F(c) - (1 - F(c)) = \gamma$$

sau

$$F(c) = \frac{1+\gamma}{2}. \tag{62}$$

Obţinem deci că în cazul unei populaţii având dispersie necunoscută σ^2, pentru un nivel de încredere γ fixat, intervalul de încredere pentru media necunoscută μ a populaţiei este

$$(\theta_1, \theta_2) = \left(\overline{x} - c\frac{s}{\sqrt{n}}, \overline{x} + c\frac{s}{\sqrt{n}}\right) \tag{63}$$

unde n este volumul selecţiei, $\overline{x} = \frac{x_1 + \ldots + x_n}{n}$ este media eşantionului, estimatorul s al abaterii pătratice medii σ este dat de formula (61), iar c este determinat din Anexa 3 astfel conform relaţiei (62) pentru $n-1$ grade de libertate.

Exemplul 10.5 *Cinci măsurători pentru temperatura de aprindere a motorinei (în grade Fahrenheit) sunt*

$$144, 147, 146, 142, 144.$$

Presupunând că temperatura de aprindere a motorinei este o variabilă aleatoare normală, să se determine un interval de 99% încredere pentru temperatura medie μ de aprindere a motorinei.

În acest caz, deoarece dispersia σ^2 este necunoscută, pentru a determina intervalul de încredere vom folosi formula (63).

Folosind datele din eşantion determinăm media de selecţie

$$\overline{x} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = 144.6$$

şi estimăm abaterea pătratică medie σ folosind formula (61):

$$s = \sqrt{\frac{\sum_{i=1}^{5} (x_i - \overline{x})^2}{5-1}} = \sqrt{3.8} = 1.95.$$

Folosind Anexa 3 pentru a determina valoarea lui c astfel încât să aibă loc relaţia (62), adică $F(c) = \frac{1+0.99}{2} = 0.995$ (alegem $n - 1 = 5 - 1 = 4$ grade de libertate), şi obţinem $c = 4.60$.

Înlocuind aceste valori în formula (63), obţinem că un interval de 99% pentru temperatura medie de aprindere a motorinei este

$$(\theta_1, \theta_2) = \left(\overline{x} - c\frac{s}{\sqrt{n}}, \overline{x} + c\frac{s}{\sqrt{n}} \right) = \left(144.6 - 4.60\frac{1.95}{\sqrt{5}}, 144.6 + 4.60\frac{1.95}{\sqrt{5}} \right) = (140.5, 148.7).$$

10.3 Intervale de încredere pentru dispersia σ^2 unei populaţii normale

Procedând similar Teoremei 10.1, şi folosind în acest caz estimatorul

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2$$

al dispersiei, se poate arăta că variabila aleatoare

$$Y = (n-1)\frac{S^2}{\sigma^2}$$

are o distribuţie χ^2 cu $n-1$ grade de libertate (valorile funcţiei de distribuţie sunt tabelate în Anexa 4).

Deoarece această distribuţie nu este simetrică (ca în cazul distribuţiei normale), în loc de $-c$ şi c determinăm două constante c_1 şi c_2 astfel încât S^2 se află între c_1 şi c_2 cu probabilitate γ, adică

$$P(c_1 \le Y \le c_2) = \gamma. \tag{64}$$

Spre exemplu, determinăm constantele c_1 şi c_2 astfel încât

$$F(c_1) = \frac{1-\gamma}{2} \qquad \text{şi} \qquad F(c_2) = \frac{1+\gamma}{2}, \tag{65}$$

unde $F(x)$ este în acest caz funcţia de distribuţie a variabilei aleatoare χ^2 cu $n-1$ grade de libertate din Anexa 4 (procedând astfel, aria de sub densitatea variabilei aleatoare χ^2 cuprinsă între c_1 şi c_2 va fi egală cu $\frac{1+\gamma}{2} - \frac{1-\gamma}{2} = \gamma$, şi are deci loc relaţia (64) de mai sus).

Relaţia (64) se mai poate scrie echivalent sub forma

$$P\left(c_1 \le (n-1)\frac{S^2}{\sigma^2} \le c_2 \right) = \gamma,$$

de unde rezolvând inegalităţile în raport cu σ^2 obţinem

$$P\left(\frac{n-1}{c_2} S^2 \le \sigma^2 \le \frac{n-1}{c_1} S^2 \right) = \gamma,$$

sau

$$P(\Theta_1 \le \mu \le \Theta_2) = \gamma,$$

unde $\Theta_1 = \frac{n-1}{c_2} S^2$ şi $\Theta_2 = \frac{n-1}{c_1} S^2$.

Înlocuind variabilele aleatoare $X_1, \ldots, X_n$ prin valorile observate $x_1, \ldots, x_n$ ale eşantionului obţinem că un interval de încredere pentru dispersia necunoscută σ^2 a populaţiei este

$$(\theta_1, \theta_2) = \left(\frac{n-1}{c_2} s^2, \frac{n-1}{c_1} s^2 \right), \tag{66}$$

unde n este volumul selecţiei, $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$ este estimatorul dispersiei σ^2, iar $c_{1,2}$ sunt determinate din Anexa 4 pentru $n-1$ grade de libertate conform relaţiilor (65).

Exemplul 10.6 *Să se determine un interval de 95% încredere pentru dispersia necunoscută σ^2 a eşantionului*

$$89, 84, 87, 81, 89, 86, 91, 90, 78, 89, 87, 99, 83, 89$$

reprezentând rezistenţa la rupere (kg/mm^2 a unei table de metal).

Folosind valorile eşantionului ($n = 14$ valori) determinăm media selecţiei $\overline{x} = \frac{x_1 + \ldots + x_{14}}{14} = 87.28$ şi estimatorul s^2 al dispersiei

$$s^2 = \frac{1}{14-1} \sum_{i=1}^{14} (x_i - \overline{x}) = 25.15.$$

Folosind Anexa 4 (cu $n - 1 = 14 - 1 = 13$ grade de libertate), determinăm constantele $c_{1,2}$ astfel încât $F(c_1) = \frac{1-\gamma}{2} = 0.025$ şi $F(c_2) = \frac{1+\gamma}{2} = 0.975$. Obţinem $c_1 = 5.01$ şi $c_2 = 24.74$.

Folosind formula (66) determinăm intervalul de încredere pentru dispersia necunoscută σ^2

$$(\theta_1, \theta_2) = \left(\frac{n-1}{c_2} s^2, \frac{n-1}{c_1} s^2 \right) = \left(\frac{14-1}{24.74} 25.15, \frac{14-1}{5.01} 25.15 \right) = (13.21, 65.25).$$

10.4 Intervale de încredere pentru alte distribuţii

Intervalele de încredere pentru media μ şi dispersia σ^2 determinate în secţiunile anterioare se aplică în cazul unei distribuţii normale.

Pentru a obţine aceste formule s-a folosit faptul că dacă $X_1, \ldots, X_n$ sunt variabile aleatoare normale independente cu medie μ şi dispersie σ^2, atunci variabila aleatoare

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

este o variabilă aleatoare normală standard.

Una din teoremele fundamentale ale Teoriei probabilităţilor (Teorema Limită centrală) afirmă că dacă $X_1, \ldots, X_n$ sunt variabile aleatoare independente cu medie μ şi dispersie σ^2 (nu neapărat normale), atunci pentru valori ale lui n ale volumului selecţiei suficient de mari, variabila aleatoare

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

este aproximativ o variabilă aleatoare normală standard.

Acest fapt permite ca formulele pentru intervalele de încredere pentru media μ şi dispersia σ^2 a unei populaţii normale din secţiunile anterioare să poată fi folosite şi în cazul altor distribuţii, nu neapărat normale. Pentru ca aproximarea să fie bună, volumul n al selecţiei trebuie să fie cel puţin 20 în cazul intervalelor de încredere pentru medie, şi cel puţin 50 în cazul intervalelor de încredere pentru dispersie.

Exerciţii

Exerciţiul 10.1 *De ce sunt intervalele de încredere mai utile decât estimatorii punctuali?*

Exerciţiul 10.2 *Ce se întâmplă cu intervalul de încredere din Examplul 10.2 dacă se foloseşte un nivel de încredere de 99% în loc de 95%? Dar dacă media de eşantionului este 8 în loc de 5?*

Exerciţiul 10.3 *Să se determine un interval de 99% încredere pentru media μ a unei populaţii normale cu abatere pătratică medie 2.5, folosind eşantionul*

$$30.8, 30.0, 29.9, 30.1, 31.7, 34.0.$$

Exerciţiul 10.4 *Să se determine un interval de* 95% *încredere pentru media* μ *a unei populaţii normale cu abatere pătratică medie* 1.2*, folosind eşantionul*

$$10, 10, 8, 12, 10, 11, 10, 11.$$

Exerciţiul 10.5 *Să se determine un interval de* 95% *încredere pentru media* μ *a unei populaţii normale cu dispersie* $\sigma^2 = 16$*, folosind un eşantion de volum* 200 *având medie* 74.81*.*

Exerciţiul 10.6 *Ce se întâmplă cu intervalul de încredere din exerciţiul anterior dacă volumul eşantionului se reduce la* 50*?*

Exerciţiul 10.7 *Ce volum al eşantionului este necesar pentru ca intervalul de* 95% *încredere pentru medie să aibă o lungimea* 2σ*?* σ*?*

Exerciţiul 10.8 *Să se determine un interval de* 99% *încredere pentru media unei populaţii normale cu dispersie* $\sigma^2 = 0.36$ *folosind un eşantion de volum* 290 *cu medie* 16.30*.*

Presupunând că populaţia din care următoarele eşantioane au fost extrase este normală, să se determine un interval de 99% încredere pentru media μ a populaţiei.

Exerciţiul 10.9 *Un eşantion de* 20 *de şuruburi cu medie* 15.50 *cm şi dispersie* 0.09 *cm*^{2}*.*

Exerciţiul 10.10 *Temperatura de topire a aluminiului (în grade Celsius) în cazul eşantionului*

$$658, 665, 652, 661, 660.$$

Exerciţiul 10.11 *Conţinutul de cupru (în procente) a unui aliaj de alamă în cazul eşantionului*

$$65, 65, 64, 63, 65, 66, 63, 64, 62, 63.$$

Exerciţiul 10.12 *Un interval de* 99% *încredere pentru parametrul* p *al distribuţiei binomiale folosind un eşantion de volum* $24,000$ *cu frecvenţă relativă de apariţie a succesului* $\frac{12,012}{24,000} = 0.5005$*.*

Exerciţiul 10.13 *Să se determine un interval de* 95% *încredere pentru procentul de maşini aflate pe o autostradă care au frâne aflate în condiţii necorespunzătoare, folosind folosind faptul că din* 500 *maşini verificate în mod aleator,* 87 *au avut frâne aflate în condiţii necorespunzătoare.*

Presupunând că populaţia din care următoarele eşantioane au fost extrase este normală, să se determine un interval de 95% încredere pentru dispersia σ^2 a populaţiei.

Exerciţiul 10.14 *Rezistenţa la rupere (kpsi) a unui aliaj de oţel aflat la temperatura camerei, folosind eşantionul*

$$251, 255, 258, 253, 253, 252, 250, 252, 255, 256.$$

Exerciţiul 10.15 *Emisia de monoxid de carbon (grame pe km) a unui anumit tip de maşină (având viteză medie de* 55 *km/h), folosind eşantionul*

$$17.3, 17.8, 18.0, 17.7, 18.2, 17.4, 17.6, 18.1$$

Exerciţiul 10.16 *Dacă* X *este o variabilă aleatoare normală cu medie* 40 *şi dispersie* 4*, ce fel de distribuţie au variabilele aleatoare* $3X$ *şi* $5X - 2$*?*

Exerciţiul 10.17 *Dacă* X_1 *şi* X_2 *sunt variabile aleatoare normale independente, cu medii* 16*, respectiv* 12*, şi dispersii* 8*, respectiv* 2*, care este distribuţia variabilei aleatoare* $4X_1 - X_2$*?*

Exerciţiul 10.18 *O maşină umple cutii cântărind* Y *kg cu* X *kg de sare, unde* X *şi* Y *sunt variabile aleatoare nromale cu medii* 200 *kg, respectiv* 10 *kg, şi dispersii* 2 *kg, respectiv 0.5 kg. Ce procent de cutii rezultate vor cântări între* 208 *kg şi* 212 *kg?*

Exerciţiul 10.19 *Dacă greutatea* X *a sacilor de ciment este distribuită normal cu medie* 40 *kg şi dispersie* 2 *kg, câţi saci de ciment poate căra un camion dacă probabilitatea ca greutatea totală să depăşească* $2,000$ *kg este de* 5%*?*