

2.2 Intervale de încredere

În cursul anterior am determinat diverse estimări $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ ale parametrului necunoscut θ al densității $f = f(x, \theta)$ a unei populații X folosind o selecție X_1, \dots, X_n de volum n a acestei populații.

În practică, valoarea $\hat{\theta}(x_1, \dots, x_n)$ calculată folosind valorile observate x_1, \dots, x_n ale variabilelor aleatoare X_1, \dots, X_n nu coincide aproape niciodată cu valoarea reală a parametrului θ .

Se pune problema cât de apropiat este $\hat{\theta}$ de valoarea reală a lui θ , în sensul determinării unui interval (L, U) ($L = L(X_1, \dots, X_n)$ și $U = U(X_1, \dots, X_n)$ sunt variabile aleatoare ce depinde de selecția X_1, \dots, X_n aleasă), astfel încât $\theta \in (L, U)$ cu o probabilitatea dată, adică

$$P(L < \theta < U) = 1 - \alpha,$$

unde $\alpha \in (0, 1)$ este o valoare dată.

Înlocuind variabilele aleatoare X_1, \dots, X_n prin valorile observate x_1, \dots, x_n , obținem $l = L(x_1, \dots, x_n)$ și $u = U(x_1, \dots, x_n)$, și numim intervalul (l, u) un interval de $100(1 - \alpha)\%$ încredere pentru parametrul necunoscut θ .

În general, se poate determina un interval de încredere pentru parametrul necunoscut θ al densității unei populații X în următoarele condiții:

există o variabilă aleatoare $g = g(X_1, \dots, X_n, \theta)$ cu următoarele proprietăți:

- 1) g depinde netrivial de selecția X_1, \dots, X_n și parametrul θ
- 2) distribuția variabilei aleatoare g nu depinde de θ (sau de alți parametri necunoscuți).

În aceste condiții, determinarea unui interval de $100(1 - \alpha)\%$ încredere pentru θ se face astfel: se determină constanțele c_L și c_U astfel încât să avem

$$P(c_L < g(X_1, \dots, X_n, \theta) < c_U) = 1 - \alpha,$$

și se rezolvă inegalitățile anterioare în raport cu θ , pentru a obține:

$$P(L(X_1, \dots, X_n) < \theta < U(X_1, \dots, X_n)) = 1 - \alpha.$$

Variabilele aleatoare L și U astfel obținute determină un interval (l, u) de $100(1 - \alpha)\%$ încredere pentru parametrul necunoscut θ .

APLICAȚII

2.2.1 Cazul mediei distribuției normale (cu dispersie cunoscută)

Să presupunem că populația X are o distribuție normală $\mathcal{N}(\mu, \sigma^2)$ cu medie μ necunoscută și dispersie σ^2 cunoscută.

Dacă X_1, \dots, X_n este o selecție de volum n din populația dată, atunci X_1, \dots, X_n sunt variabile aleatoare $\mathcal{N}(\mu, \sigma^2)$ independente, și se poate arăta

că $X_1 + \dots + X_n$ este o variabilă aleatoare $\mathcal{N}(n\mu, n\sigma^2)$ cu medie $n\mu$ și dispersie $n\sigma^2$.

Rezultă că variabila aleatoare

$$\frac{X_1, \dots, X_n - n\mu}{\sigma\sqrt{n}} = \frac{\frac{X_1, \dots, X_n}{n} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

este o variabilă aleatoare $\mathcal{N}(\mu, \sigma^2)$, și deci variabila aleatoare $g = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ verifică ipotezele precizate anterior.

Să notăm cu $z_\alpha > 0$ punctul cu proprietatea că aria cuprinsă între axa orizontală și graficul densității normale $\mathcal{N}(0, 1)$, situate la dreapta punctului z_α este egală cu α , adică

$$\int_{z_\alpha}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \alpha,$$

sau echivalent

$$\Phi(z_\alpha) = \int_{-\infty}^{z_\alpha} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 1 - \int_{z_\alpha}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 1 - \alpha. \quad (2.1)$$

De asemenea, să observăm că din simetria funcției de densitate normală, aria de sub graficul densității, aflate la stânga punctului $-z_\alpha$ este de asemenea egală cu α , adică

$$\Phi(-z_\alpha) = \int_{-\infty}^{-z_\alpha} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \alpha. \quad (2.2)$$

Observația 2.2.1 Pentru un nivel de încredere $\alpha \in (0, 1)$ dat, valorile lui z_α se determină din tabele de valori ale funcției de distribuție a densității normale $\mathcal{N}(0, 1)$, folosind formula (2.1) sau (2.2).

Cu notația anterioară, avem

$$\begin{aligned} P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < z_{\alpha/2}\right) &= \int_{-z_{\alpha/2}}^{z_{\alpha/2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \\ &= \Phi(z_{\alpha/2}) - \Phi(-z_{\alpha/2}) \\ &= 1 - \alpha/2 - (\alpha/2) \\ &= 1 - \alpha, \end{aligned}$$

de unde rezolvând în raport cu parametrul μ necunoscut obținem

$$P\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha,$$

și deci avem

$$\begin{cases} L = L(X_1, \dots, X_n) = \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \\ U = U(X_1, \dots, X_n) = \bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \end{cases}$$

Am obținut deci următoarea:

Propoziția 2.2.2 Dacă X_1, \dots, X_n este o selecție de volum n dintr-o populație normală $\mathcal{N}(\mu, \sigma^2)$ cu dispersie σ^2 cunoscută, atunci un interval de $100(1 - \alpha)\%$ încredere pentru parametrul necunoscut μ este

$$(l, u) = \left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right),$$

unde $\bar{x} = \frac{x_1 + \dots + x_n}{n}$ este media de selecție a valorilor observate x_1, \dots, x_n ale selecției X_1, \dots, X_n și $z_{\alpha/2}$ este determinat astfel încât $\Phi(z_{\alpha/2}) = 1 - \alpha/2$.

Observația 2.2.3 Să presupunem că într-un exemplu concret obținem pentru μ intervalul (5.9; 6.1) de $100(1 - \alpha)\%$ încredere. Este tentant (dar incorect!) să se considere că probabilitatea ca μ să aparțină intervalului (5.9; 6.1) este $1 - \alpha$. Cum μ are o valoare determinată, această probabilitate este sau 0 sau 1, și deci nu poate fi eglă cu $1 - \alpha$ (pentru $\alpha \in (0, 1)$).

Interpretarea corectă constă în a realiza că intervalul de încredere este un interval aleator (capetele intervalului depind de selecția aleasă), cu proprietatea că $P(L < \mu < U) = 1 - \alpha$, adică μ se găsește în $100(1 - \alpha)\%$ din intervalele (l, u) calculate, unde $l = L(x_1, \dots, x_n)$ și $u = U(x_1, \dots, x_n)$ iar x_1, \dots, x_n sunt valorile observate ale variabilelor aleatoare X_1, \dots, X_n .

Observația 2.2.4 (Alegerea volumului n al selecției) Dacă se dorește ca eroarea de estimare $|\bar{X} - \mu|$ să nu depășească o valoare E dată cu încredere $100(1 - \alpha)\%$, atunci

$$P(-E < \bar{X} - \mu < E) \geq 1 - \alpha,$$

și cum

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

rezultă prin comparare că trebuie să avem

$$E \geq z_{\alpha/2} \frac{\sigma}{\sqrt{n}},$$

de unde obținem că volumul selecției trebuie să verifice inegalitatea

$$n \geq \left(\frac{\sigma z_{\alpha/2}}{E} \right)^2$$

Observația 2.2.5 În mod similar construcției anterioare, atunci când ne interesează numai estimarea parametrului necunoscut printr-o valoare limită superioară-inferioară, se pot construi intervale de $100(1 - \alpha)\%$ încredere de mărginire:

- superioară: $(l, u) = \left(-\infty, \bar{x} + z_{\alpha} \frac{\sigma}{\sqrt{n}}\right)$;
- inferioară: $(l, u) = \left(\bar{x} - z_{\alpha} \frac{\sigma}{\sqrt{n}}, +\infty\right)$.

2.2.2 Intervale de încredere pentru media selecțiilor mari

Considerăm cazul estimării mediei μ a unei populații X , nu neapărat normal distribuită. Reamintim următoarea:

Teorema 2.2.6 (Teorema limită centrală) *Dacă X_1, X_2, \dots sunt variabile aleatoare independente și identic distribuite, cu medie $M(X_i) = \mu$ și dispersie $\sigma^2(X_i) = \sigma^2$, atunci funcția de distribuție a variabilei aleatoare*

$$\frac{X_1 + \dots + X_n - n\mu}{\sigma\sqrt{n}}$$

ține pentru $n \rightarrow \infty$ către funcția de distribuție normală $\mathcal{N}(0, 1)$, adică

$$P\left(\frac{X_1 + \dots + X_n - n\mu}{\sigma\sqrt{n}} < a\right) \xrightarrow{n \rightarrow \infty} \int_{-\infty}^a \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx.$$

Din Teorema limită centrală rezultă deci că pentru valori n ale volumului selecției mari (spre exemplu $n \geq 30$), putem considera că variabila aleatoare $\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ este aproximativ o variabilă aleatoare normală $\mathcal{N}(0, 1)$.

În plus, dacă dispersia σ^2 este necunoscută, atunci înlocuind pe σ prin estimatorul

$$\hat{\sigma}(X_1, \dots, X_n) = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}},$$

pentru valori n suficient de mari (spre exemplu pentru $n \geq 40$), avem

$$\frac{\bar{X} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}} \approx \mathcal{N}(0, 1)$$

este aproximativ o variabilă aleatoare normală cu medie 0 și dispersie 1. Procedând similar cazului anterior, obținem următoarea:

Propoziția 2.2.7 *Pentru valori n ale volumului selecției X_1, \dots, X_n suficient de mari, variabila aleatoare*

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

este aproximativ o variabilă aleatoare $\mathcal{N}(0, 1)$, și deci un interval de $100(1 - \alpha)\%$ încredere pentru media μ este

$$(l, u) = \left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right),$$

unde $\bar{x} = \frac{x_1 + \dots + x_n}{n}$ este media de selecție a valorilor observate x_1, \dots, x_n ale selecției X_1, \dots, X_n și $z_{\alpha/2}$ este determinat astfel încât $\Phi(z_{\alpha/2}) = 1 - \alpha/2$.