Statistica matematică

► Statistica matematică este o ramură a matematicii aplicate, care se ocupă de colectarea, gruparea, analiza şi interpretarea datelor referitoare la anumite fenomene în scopul obţinerii unor previziuni;
• statistica descriptivă: metode de colectare, organizare, sintetizare, prezentare şi descriere a datelor numerice (sau nenumerice) într-o formă convenabilă
• statistica inferenţială: metode de interpretare a rezultatelor obţinute prin metodele statisticii descriptive, utilizate apoi pentru luarea deciziilor.
► O *colectivitate* sau *populaţie statistică* $\mathcal{C}$ este o mulţime de elemente care au anumite însuşiri comune ce fac obiectul analizei statistice. Numărul elementelor populaţiei se numeşte volumul populaţiei.
Exemple de populaţii statistice: mulţimea persoanelor dintr-o anumită ţară, localitate, zonă etc. într-un anumit an, multimea gospodăriilor din Romania la un moment dat, mulţimea consumatorilor unui produs, mulţimea societăţilor care produc un anumit produs, angajaţii unei societăţi, studenţii unei facultăţi.
► *Eşantionul* $\mathcal{E}$ reprezintă o submulţime a unei populaţii statistice $\mathcal{E} \subset \mathcal{C}$, constituită dupa criterii bine stabilite:
a) să fie aleatoare;
b) toate elementele colectivităţii să aibe aceeaşi şansă de a fi alese în eşantion;
c) eşantionul să fie reprezentativ (structura eşantionului să fie apropiată de structura populaţiei);
d) volumul eşantionului să fie suficient de mare.
► *Unitatea statistică* (indivizii) este elementul, entitatea de sine stătătoare a unei populaţii statistice, care posedă o serie de trăsături caracteristice ce-i conferă apartenenţa la populaţia studiată.
De exemplu: *unitatea statistică simplă*: un salariat, un student, un agent economic, o trăsătură, o părere; *unitatea statistică complexă*: o grupă de studenţi sau o echipă de salariaţi, o familie sau o gospodărie, o categorie de mărfuri.
► *Variabila statistică* sau *carcteristica* reprezintă o însuşire, o proprietate măsurabilă a unei unităţi statistice, întâlnită la toate unităţile care aparţin aceleiaşi colectivităţi şi care prezintă variabilitate de la o unitate statistică la alta. Caracteristica sau variabila statististică corespunde unei variabile aleatoare.
Exemple de caracteristici: vârsta, salariul, preferinţele politice, preţul unui produs, calitatea unor servicii, nivelul de studii.
• caracteristici continue (greutatea, înălţimea, valoarea glicemiei, temperatura aerului)
• caracteristici discrete (numări elevi ai unei şcoli, numărul liceelor existente într-un oraş)
• caracteristici nominale (culoarea ochilor, ramura de activitate, religia)
• caracteristici nominale ordonate (starea de sănătate / calitatea unor servicii - precară, mai bună, bună, foarte bună)
• caracteristici dichotomiale (binare) (stagiul militar - satisfăcut/nesatisfăcut, starea civilă - căsătorit/necăsătorit, genul - masculin/feminin)
► *Datele statistice* reprezintă observaţiile rezultate dintr-o cercetare statistică, sau ansamblul valorilor colectate în urma unei cercetări statistice.
De exemplu: un angajat al unei companii are o vechime de 6 ani în muncă. Angajatul reprezintă unitatea statistică, vechimea în muncă este caracteristica (variabila) cercetată, iar 6 este valoarea acestei caracteristici.

O *colectivitate* (populaţie) $\mathcal{C}$ este cercetatată din punctul de vedere al caracteristicii (variabilei statistice) X.

► Fie $\mathcal{E} \subset \mathcal{C}$ un eşantion. Se numesc *date de selecţie* relative la caracteristica X datele statistice $x_1, \ldots, x_n$ obţinute prin cercetarea indivizilor care fac parte din eşantionul $\mathcal{E}$.
► Datele de selecţie $x_1, \ldots, x_n$ pot fi considerate ca fiind valorile unor variabile aleatoare $X_1, \ldots, X_n$, numite variabile de selecţie şi care se consideră a fi variabile aleatoare independente şi având aceeaşi distribuţie ca X.
► Fie $x_1, \ldots, x_n$ datele statistice pentru caracteristica cercetată X, notăm cu $X_1, \ldots, X_n$ variabilele de selecţie corespunzătoare. Fie $g : \mathbb{R}^n \to \mathbb{R}$ o funcţie astfel încât $g(X_1, \ldots, X_n)$ este o variabilă aleatoare.
$g(X_1, \ldots, X_n)$ se numeşte *funcţie de selecţie* sau *estimator*
$g(x_1, \ldots, x_n)$ se numeşte valoarea funcţiei de selecţie sau *valoarea estimatorului*.

Distribuţia caracteristicii X poate fi

1) complet specificată (de ex.: $X \sim Exp(3)$, $X \sim Bin(10, 0.3)$, $X \sim N(0, 1)$)

2) specificată, dar depinzând de unul sau mai mulţi parametri necunoscuţi
(de ex.: $X \sim Exp(\lambda)$, $X \sim Bin(10, p)$, $X \sim N(m, \sigma^2)$)

3) necunoscută: $X\sim ?$

• în cazurile 2) şi 3) parametrii necunoscuţi sau distribuţia necunoscută

$\hookrightarrow$ se estimează $\rightarrow$ teoria estimaţiei

$\hookrightarrow$ se testează $\rightarrow$ teste statistice

• Exemple de estimatori (funcţii de selecţie) sunt: media de selecţie, dispersia de selecţie, funcţia de repartiţie empirică (de selecţie).

Fie $x_1,\dots,x_n$ datele statistice pentru caracteristica cercetată X, notăm cu $X_1,\dots,X_n$ variabilele de selecţie corespunzătoare

▶ media de selecţie

$$\bar{X}_n=\frac{1}{n}\left(X_1+\cdots+X_n\right)$$

▶ valoarea mediei de selecţie

$$\bar{x}_n=\frac{1}{n}\left(x_1+\cdots+x_n\right)$$

▶ varianţa (dispersia) de selecţie

$$\tilde{S}_n^2=\frac{1}{n-1}\sum_{k=1}^{n}(X_k-\bar{X}_n)^2$$

▶ valoarea varianţei (dispersiei) de selecţie

$$\tilde{s}_n^2=\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{x}_n)^2$$

▶ abaterea standard de selecţie

$$\tilde{S}_n=\left(\frac{1}{n-1}\sum_{k=1}^{n}(X_k-\bar{X}_n)^2\right)^{\frac{1}{2}}$$

▶ valoarea abaterii standard de selecţie

$$\tilde{s}_n=\left(\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{x}_n)^2\right)^{\frac{1}{2}}$$

▶ funcţia de repartiţie empirică $\hat{F}_n:\mathbb{R}\times\Omega\rightarrow\mathbb{R}$

$$\hat{F}_n(x)=\frac{\#\{i\in\{1,...,n\}:X_i\le x\}}{n},x\in\mathbb{R}$$

▶ valoarea funcţiei de repartiţie empirice

$$\hat{F}_n(x)=\frac{\#\{i\in\{1,...,n\}:x_i\le x\}}{n},x\in\mathbb{R}$$

▶ $g(X_1,\dots,X_n)$ este estimator nedeplasat pentru parametrul necunoscut θ, dacă

$$E(g(X_1,\dots,X_n))=\theta.$$

▶ $g(X_1,\dots,X_n)$ este estimator consistent pentru parametrul necunoscut θ, dacă

$$g(X_1,\dots,X_n)\overset{a.s.}{\rightarrow}\theta.$$

▶ Fie $g_1(X_1,\dots,X_n)$ şi $g_2(X_1,\dots,X_n)$ estimatori nedeplasaţi pentru parametrul necunoscut θ. $g_1(X_1,\dots,X_n)$ este mai eficient decât $g_2(X_1,\dots,X_n)$, dacă $V(g_1)<V(g_2)$.

▶ Fie $\alpha\in(0,1)$ nivelul de semnificaţie (probabilitatea de risc). Cuantila de ordin α pentru distribuţia caracteristicii cercetate X este numărul $z_\alpha\in\mathbb{R}$ pentru care

$$P(X<z_\alpha)\le\alpha\le P(X\le z_\alpha).$$

• dacă X este v.a. continuă, atunci: z_α este cuantilă de ordin $\alpha\iff P(X\le z_\alpha)=\alpha\iff F_X(z_\alpha)=\alpha$

• $\alpha\cdot 100\%$ din valorile lui X sunt mai mici sau egale cu z_α

Exemple:
1) Media de selecţie $\bar{X}_n$ este estimator nedeplasat şi consistent pentru **media teoretică** $E(X)$ a caracteristicii X.
2) Varianţa (dispersia) de selecţie $\tilde{S}_n^2$ este estimator nedeplasat şi consistent pentru **varianţa teoretică** $V(X)$ a caracteristicii X.
3) **Funcţia de repartiţie de selecţie** calculată în $x \in \mathbb{R}$: $\hat{F}_n(x)$ este estimator nedeplasat şi consistent pentru $F(x)$ valoarea funcţiei de repartiţie teoretice în x.

Teste statistice

Fie $x_1, \ldots, x_n$ datele statistice pentru caracteristica cercetată X, notăm cu $X_1, \ldots, X_n$ variabilele de selecţie corespunzătoare.
► Ipoteza statistică este o presupunere relativă la un parametru necunoscut θ
► Metoda de stabilire a veridicităţii unei ipoteze statistice se numeşte test (criteriu de verificare).
► Rezultatul testării se foloseşte apoi pentru luarea unor decizii (cum ar fi: eficienţa unor medicamente, strategii de marketing, alegerea unui produs etc.).
► Se formulează ipoteza nulă H_0 şi ipoteza alternativă H_1, privind parametrul θ; fie θ_0 o valoare dată

$$\text{I. } H_0 : \theta = \theta_0 \quad H_1 : \theta \neq \theta_0$$

$$\text{II. } H_0 : \theta \leq \theta_0 \quad H_1 : \theta > \theta_0$$

$$\text{III. } H_0 : \theta \geq \theta_0 \quad H_1 : \theta < \theta_0$$

Se dă $\alpha \in (0, 1)$ nivelul de semnificaţie (probabilitatea de risc). Formularea unui test revine la construirea unei regiuni critice $U \subset \mathbb{R}^n$ (pentru cazurile I, II, respectiv III) astfel încât

$$P((X_1, \ldots, X_n) \in U | H_0) = \alpha$$

ceea ce este echivalent cu

$$P((X_1, \ldots, X_n) \notin U | H_0) = 1 - \alpha$$

Concluzia testului:
$(x_1, \ldots, x_n) \notin U \Rightarrow$ ipoteza H_0 este admisă
$(x_1, \ldots, x_n) \in U \Rightarrow$ ipoteza H_0 este respinsă, în favoarea ipotezei H_1

► O colectivitate este testată în raport cu caracteristica X.
- test pentru valoarea medie $E(X)$
  - ▷ când varianţa teoretică $V(X)$ este cunoscută: testul lui Gauss (testul Z)
  - ▷ când varianţa teoretică $V(X)$ este necunoscută: Student Test (testul T)
- test pentru abaterea standard teoretică $\sqrt{V(X)}$ sau pentru varianţa teoretică $V(X)$: testul χ^2
- test asupra proporţiei (testul Gauss aproximativ)

Paşii în efectuarea unui test statistic:
- Care parametru se testează? Care test este potrivit?
- Care este ipoteza nulă H_0 şi care este ipoteza alternativă H_1?
- Care este nivelul de semnificaţie (probabilitatea de risc) α ?
- Calculul valorii estimatorului pe baza datelor statistice
- Concluzia testului

Test pentru media $m = E(X)$ caracteristicii cercetate X, când varianţa $\sigma^2 = V(X)$ este cunoscută

► se dau $\alpha \in (0, 1)$, m_0, σ, datele statistice $x_1, \ldots, x_n$
► dacă $X \sim N(m, \sigma^2)$ sau $n > 30$ şi X are o distribuţie necunoscută, atunci $\dfrac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$

▶ folosind datele statistice $x_1, \ldots, x_n$, se calculează $z = \dfrac{\bar{x}_n - m_0}{\frac{\sigma}{\sqrt{n}}}$

▶ cuantilele legii normale $N(0,1)$:
$z_{1-\frac{\alpha}{2}} = norminv(1 - \frac{\alpha}{2}, 0, 1)$, $z_{1-\alpha} = norminv(1-\alpha, 0, 1)$, $z_\alpha = norminv(\alpha, 0, 1)$

	I. H_0: $m = m_0$ H_1: $m \neq m_0$	II. H_0: $m \leq m_0$ H_1: $m > m_0$	III. H_0: $m \geq m_0$ H_1: $m < m_0$
Se acceptă H_0 dacă	$\lvert z\rvert < z_{1-\frac{\alpha}{2}}$	$z < z_{1-\alpha}$	$z > z_\alpha$
Se respinge H_0 în favoarea lui H_1, dacă	$\lvert z\rvert \geq z_{1-\frac{\alpha}{2}}$	$z \geq z_{1-\alpha}$	$z \leq z_\alpha$

▶ în Octave/Matlab: *ztest*

▶ regiunea critică $U \subset \mathbb{R}^n$

I. $U = \left\{ (u_1, \ldots, u_n) \in \mathbb{R}^n : \left| \dfrac{\bar{u}_n - m_0}{\frac{\sigma}{\sqrt{n}}} \right| \geq z_{1-\frac{\alpha}{2}} \right\}$, unde $\bar{u}_n = \dfrac{1}{n}(u_1 + \cdots + u_n)$

II. $U = \left\{ (u_1, \ldots, u_n) \in \mathbb{R}^n : \dfrac{\bar{u}_n - m_0}{\frac{\sigma}{\sqrt{n}}} \geq z_{1-\alpha} \right\}$

III. $U = \left\{ (u_1, \ldots, u_n) \in \mathbb{R}^n : \dfrac{\bar{u}_n - m_0}{\frac{\sigma}{\sqrt{n}}} \leq z_\alpha \right\}$

Test pentru media $m = E(X)$ caracteristicii cercetate X, când varianţa $\sigma^2 = V(X)$ este necunoscută

▶ se dau $\alpha \in (0,1)$, m_0, datele statistice $x_1, \ldots, x_n$

▶ dacă $X \sim N(m, \sigma^2)$ sau $n > 30$ şi X are o distribuţie necunoscută, atunci $\dfrac{\bar{X}_n - m}{\frac{\bar{S}_n}{\sqrt{n}}} \sim Student(n-1)$

▶ folosind datele statistice $x_1, \ldots, x_n$, se calculează $t = \dfrac{\bar{x}_n - m_0}{\frac{\bar{s}_n}{\sqrt{n}}}$

▶ cuantilele legii Student cu $n-1$ grade de libertate:
$t_{1-\frac{\alpha}{2}} = tinv(1 - \frac{\alpha}{2}, n-1)$, $t_{1-\alpha} = tinv(1-\alpha, n-1)$, $t_\alpha = tinv(\alpha, n-1)$

	I. H_0: $m = m_0$ H_1: $m \neq m_0$	II. H_0: $m \leq m_0$ H_1: $m > m_0$	III. H_0: $m \geq m_0$ H_1: $m < m_0$
Se acceptă H_0 dacă	$\lvert t\rvert < t_{1-\frac{\alpha}{2}}$	$t < t_{1-\alpha}$	$t > t_\alpha$
Se respinge H_0 în favoarea lui H_1, dacă	$\lvert t\rvert \geq t_{1-\frac{\alpha}{2}}$	$t \geq t_{1-\alpha}$	$t \leq t_\alpha$

▶ în Octave/Matlab: *ttest*

▶ regiunea critică $U \subset \mathbb{R}^n$

I. $U = \left\{ (u_1, \ldots, u_n) \in \mathbb{R}^n : \left| \dfrac{\bar{u}_n - m_0}{\frac{\tilde{\sigma}_n}{\sqrt{n}}} \right| \geq t_{1-\frac{\alpha}{2}} \right\}$, unde $\bar{u}_n = \dfrac{1}{n}(u_1 + \cdots + u_n)$, $\tilde{\sigma}_n = \left(\dfrac{1}{n-1} \displaystyle\sum_{k=1}^{n} (u_k - \bar{u}_n)^2 \right)^{\frac{1}{2}}$

II. $U = \left\{ (u_1, \ldots, u_n) \in \mathbb{R}^n : \dfrac{\bar{u}_n - m_0}{\frac{\tilde{\sigma}_n}{\sqrt{n}}} \geq t_{1-\alpha} \right\}$

III. $U = \left\{ (u_1, \ldots, u_n) \in \mathbb{R}^n : \dfrac{\bar{u}_n - m_0}{\frac{\tilde{\sigma}_n}{\sqrt{n}}} \leq t_\alpha \right\}$

Test asupra proporţiei p pentru caracteristica $X \sim Bernoulli(p)$ (testul Gauss aproximativ)

- se dau $\alpha \in (0,1)$, p_0, datele statistice $x_1, \ldots, x_n$
- dacă $X \sim Bernoulli(p)$ şi $np(1-p) \geq 10$, atunci $\dfrac{\bar{X}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$
- folosind datele statistice $x_1, \ldots, x_n$, se calculează $z = \dfrac{\bar{x}_n - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
- cuantilele legii normale $N(0,1)$:
  $z_{1-\frac{\alpha}{2}} = norminv(1-\frac{\alpha}{2}, 0, 1)$, $z_{1-\alpha} = norminv(1-\alpha, 0, 1)$, $z_\alpha = norminv(\alpha, 0, 1)$

	I. H_0: $p = p_0$ H_1: $p \neq p_0$	II. H_0: $p \leq p_0$ H_1: $p > p_0$	III. H_0: $p \geq p_0$ H_1: $p < p_0$
Se acceptă H_0 dacă	$\lvert z\rvert < z_{1-\frac{\alpha}{2}}$	$z < z_{1-\alpha}$	$z > z_\alpha$
Se respinge H_0 în favoarea lui H_1, dacă	$\lvert z\rvert \geq z_{1-\frac{\alpha}{2}}$	$z \geq z_{1-\alpha}$	$z \leq z_\alpha$

- regiunea critică $U \subset \mathbb{R}^n$

I. $U = \left\{(u_1, \ldots, u_n) \in \mathbb{R}^n : \left|\dfrac{\bar{u}_n - p_0}{\frac{\sigma}{\sqrt{n}}}\right| \geq z_{1-\frac{\alpha}{2}}\right\}$, unde $\bar{u}_n = \dfrac{1}{n}(u_1 + \cdots + u_n)$

II. $U = \left\{(u_1, \ldots, u_n) \in \mathbb{R}^n : \dfrac{\bar{u}_n - p_0}{\frac{\sigma}{\sqrt{n}}} \geq z_{1-\alpha}\right\}$

III. $U = \left\{(u_1, \ldots, u_n) \in \mathbb{R}^n : \dfrac{\bar{u}_n - p_0}{\frac{\sigma}{\sqrt{n}}} \leq z_\alpha\right\}$

Test pentru abaterea standard teoretică $\sigma = \sqrt{V(X)}$ a caracteristicii cercetate X / Test pentru varianţa teoretică $\sigma^2 = V(X)$ a caracteristicii cercetate X

- se dau $\alpha \in (0,1)$, σ_0, datele statistice $x_1, \ldots, x_n$
- dacă $X \sim N(m, \sigma^2)$, atunci $\frac{n-1}{\sigma^2}\tilde{S}_n^2 \sim \chi^2(n-1)$
- folosind datele statistice $x_1, \ldots, x_n$, se calculează $c = \dfrac{n-1}{\sigma_0^2} \cdot \tilde{s}_n^2$
- cuantilele χ^2 (Chi-pătrat) cu $n-1$ grade de libertate:
  $c_{1-\frac{\alpha}{2}} = chi2inv(1-\frac{\alpha}{2}, n-1)$, $c_{\frac{\alpha}{2}} = chi2inv(\frac{\alpha}{2}, n-1)$, $c_{1-\alpha} = chi2inv(1-\alpha, n-1)$, $c_\alpha = chi2inv(\alpha, n-1)$

	I. H_0: $\sigma = \sigma_0$ H_1: $\sigma \neq \sigma_0$	II. H_0: $\sigma \leq \sigma_0$ H_1: $\sigma > \sigma_0$	III. H_0: $\sigma \geq \sigma_0$ H_1: $\sigma < \sigma_0$
Se acceptă H_0, dacă	$c_{\frac{\alpha}{2}} < c < c_{1-\frac{\alpha}{2}}$	$c < c_{1-\alpha}$	$c > c_\alpha$
Se respinge H_0 în favoarea lui H_1, dacă	$c \notin (c_{\frac{\alpha}{2}}, c_{1-\frac{\alpha}{2}})$	$c \geq c_{1-\alpha}$	$c \leq c_\alpha$

- în Octave/Matlab: *vartest*
- regiunea critică $U \subset \mathbb{R}^n$

I. $U = \left\{(u_1, \ldots, u_n) \in \mathbb{R}^n : \dfrac{1}{\sigma_0^2}\sum_{k=1}^{n}(u_k - \bar{u}_n)^2 \notin \left(c_{\frac{\alpha}{2}}, c_{1-\frac{\alpha}{2}}\right)\right\}$, unde $\bar{u}_n = \dfrac{1}{n}(u_1 + \cdots + u_n)$

II. $U = \left\{(u_1, \ldots, u_n) \in \mathbb{R}^n : \dfrac{1}{\sigma_0^2}\sum_{k=1}^{n}(u_k - \bar{u}_n)^2 \geq c_{1-\alpha}\right\}$

III. $U = \left\{ (u_1, \ldots, u_n) \in \mathbb{R}^n : \frac{1}{\sigma_0^2} \sum_{k=1}^{n} (u_k - \bar{u}_n)^2 \leq c_\alpha \right\}$

Probleme:

1. Specificaţiile unui anumit medicament indică faptul că fiecare comprimat conţine în medie 2.4 g de substanţă activă. 100 de comprimate alese la întâmplare din producţie sunt analizate şi se constată că ele conţin în medie 2.5 g de substanţă activă cu o deviaţie standard de 0.2 g. Se poate spune că medicamentul respectă specificaţiile (cu $\alpha = 0.01$)?
Soluţie: H_0: $m = 2.4$ cu H_1: $m \neq 2.4$, testul Student.

2. Un manager este suspicios că un utilaj, care umple anumite cutii cu ceai, trebuie înlocuit cu unul mult mai precis. 121 de cutii cu ceai sunt cântărite. S-a obţinut o medie de 196.6 g şi o deviaţie standard de 2.09 g pentru acest eşantion.
a) Să se testeze dacă abaterea standard a utilajului este de 2 g.
b) Sunt datele suficiente pentru a concluziona, că utilajul trebuie reglat pentru că nu pune 200 g de ceai într-o cutie? ($\alpha = 0.01$)
Soluţie: a) H_0: $\sigma = 2$ cu H_1: $\sigma \neq 2$, testul pentru abaterea standard
b) H_0: $m = 200$ cu H_1: $m \neq 200$, testul Student.

Intervale de încredere

Fie $x_1, \ldots, x_n$ datele statistice pentru caracteristica cercetată X, a cărei distribuţie depinde de parametrul necunoscut θ; notăm cu $X_1, \ldots, X_n$ variabilele de selecţie corespunzătoare. Fie $\alpha \in (0, 1)$ nivelul de semnificaţie; $1-\alpha$ se numeşte nivelul de încredere. Se caută doi estimatori $g_1(X_1, \ldots, X_n)$ şi $g_2(X_1, \ldots, X_n)$ astfel încât

$$P(g_1(X_1, \ldots, X_n) < \theta < g_2(X_1, \ldots, X_n)) = 1 - \alpha \;\Leftrightarrow\; P\Big(\theta \notin \Big(g_1(X_1, \ldots, X_n), g_2(X_1, \ldots, X_n)\Big)\Big) = \alpha$$

► $\Big(g_1(X_1, \ldots, X_n), g_2(X_1, \ldots, X_n)\Big)$ se numeşte **interval de încredere pentru parametrul necunoscut** θ
► $\Big(g_1(x_1, \ldots, x_n), g_2(x_1, \ldots, x_n)\Big)$ este valoarea intervalului de încredere pentru parametrul necunoscut θ
► $g_1(X_1, \ldots, X_n)$ este limita inferioară a intervalului de încredere, valoarea sa este $g_1(x_1, \ldots, x_n)$
► $g_2(X_1, \ldots, X_n)$ este limita superioară a intervalului de încredere, valoarea sa este $g_2(x_1, \ldots, x_n)$
► probabilitatea ca parametrul necunoscut θ să fie în intervalul $\Big(g_1(X_1, \ldots, X_n), g_2(X_1, \ldots, X_n)\Big)$ este $1-\alpha$ (nivelul de încredere)

Interval de încredere pentru media $m = E(X)$ caracteristicii cercetate X, când varianţa $\sigma^2 = V(X)$ este cunoscută

► se dau $\alpha \in (0, 1)$, σ, datele statistice $x_1, \ldots, x_n$
► dacă $X \sim N(m, \sigma^2)$ sau $n > 30$ şi X are o distribuţie necunoscută, atunci $\dfrac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$
► cuantilele legii normale $N(0, 1)$:
$z_{1-\frac{\alpha}{2}} = norminv(1 - \frac{\alpha}{2}, 0, 1)$, $z_{1-\alpha} = norminv(1 - \alpha, 0, 1)$, $z_\alpha = norminv(\alpha, 0, 1)$

• *interval de încredere bilateral*: să se indice doi estimatori $g_1(X_1, \ldots, X_n)$ şi $g_2(X_1, \ldots, X_n)$ astfel încât pentru media teoretică $m = E(X)$ să avem:

$$P\Big(g_1(X_1, \ldots, X_n) < m < g_2(X_1, \ldots, X_n)\Big) = 1 - \alpha$$

▷ din $P\left(\left|\dfrac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}}\right| < z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$ avem : $\left|\dfrac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}}\right| < z_{1-\frac{\alpha}{2}} \Leftrightarrow \bar{X}_n - \dfrac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}} < m < \bar{X}_n + \dfrac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}$

• *intervalul de încredere (bilateral)* pentru $m = E(X)$ (media teoretică) este $\left(\bar{X}_n - \dfrac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}\, ,\; \bar{X}_n + \dfrac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}\right)$

▷ se calculează valoarea intervalului de încredere $\left(\bar{x}_n - \dfrac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}\, ,\; \bar{x}_n + \dfrac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}\right)$

• *interval de încredere unilateral*: să se indice doi estimatori $g_1(X_1, \ldots, X_n)$, $g_2(X_1, \ldots, X_n)$ astfel încât pentru media teoretică

$m = E(X)$ să avem:

$$P(g_1(X_1, \dots, X_n) < m) = 1 - \alpha, \quad P(m < g_2(X_1, \dots, X_n)) = 1 - \alpha$$

▷ din $P\left(\dfrac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} < z_{1-\alpha}\right) = 1 - \alpha$ avem : $\dfrac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} < z_{1-\alpha} \Leftrightarrow m > \bar{X}_n - \dfrac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha}$

• $g_1(X_1, \dots, X_n) = \bar{X}_n - \dfrac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha}$; valoarea intervalului de încredere este $\left(\bar{X}_n - \dfrac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha}\, , \infty\right)$

▷ din $P\left(\dfrac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} > z_{\alpha}\right) = 1 - \alpha$ avem : $\dfrac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} > z_{\alpha} \Leftrightarrow m < \bar{X}_n - \dfrac{\sigma}{\sqrt{n}} \cdot z_{\alpha}$

• $g_2(X_1, \dots, X_n) = \bar{X}_n - \dfrac{\sigma}{\sqrt{n}} \cdot z_{\alpha}$; valoarea intervalul de încredere este $\left(-\infty\, , \bar{x}_n - \dfrac{\sigma}{\sqrt{n}} \cdot z_{\alpha}\right)$

Interval de încredere pentru media $m = E(X)$ caracteristicii cercetate X, când varianţa $\sigma^2 = V(X)$ este necunoscută

▶ se dau $\alpha \in (0, 1)$, datele statistice $x_1, \dots, x_n$

▶ dacă $X \sim N(m, \sigma^2)$ sau $n > 30$ şi X are o distribuţie necunoscută, atunci $\dfrac{\bar{X}_n - m}{\frac{\tilde{S}_n}{\sqrt{n}}} \sim Student(n-1)$

▶ cuantilele legii $Student(n-1)$:
$t_{1-\frac{\alpha}{2}} = norminv(1 - \frac{\alpha}{2}, 0, 1)$, $t_{1-\alpha} = norminv(1 - \alpha, 0, 1)$, $t_{\alpha} = norminv(\alpha, 0, 1)$

• *interval de încredere bilateral*: să se indice doi estimatori $g_1(X_1, \dots, X_n)$ şi $g_2(X_1, \dots, X_n)$ astfel încât pentru media teoretică $m = E(X)$ să avem:

$$P\Big(g_1(X_1, \dots, X_n) < m < g_2(X_1, \dots, X_n)\Big) = 1 - \alpha$$

▷ din $P\left(\left|\dfrac{\bar{X}_n - m}{\frac{\tilde{S}_n}{\sqrt{n}}}\right| < t_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$ avem : $\left|\dfrac{\bar{X}_n - m}{\frac{\tilde{S}_n}{\sqrt{n}}}\right| < t_{1-\frac{\alpha}{2}} \Leftrightarrow \bar{X}_n - \dfrac{\tilde{S}_n}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}} < m < \bar{X}_n + \dfrac{\tilde{S}_n}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}$

• *intervalul de încredere (bilateral)* pentru $m = E(X)$ (media teoretică) este $\left(\bar{X}_n - \dfrac{\tilde{S}_n}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}\, ,\ \bar{X}_n + \dfrac{\tilde{S}_n}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}\right)$

▷ se calculează valoarea intervalului de încredere $\left(\bar{x}_n - \dfrac{\tilde{s}_n}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}\, ,\ \bar{x}_n + \dfrac{\tilde{s}_n}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}\right)$

• *interval de încredere unilateral*: să se indice doi estimatori $g_1(X_1, \dots, X_n)$, $g_2(X_1, \dots, X_n)$ astfel încât pentru media teoretică $m = E(X)$ să avem:

$$P(g_1(X_1, \dots, X_n) < m) = 1 - \alpha, \quad P(m < g_2(X_1, \dots, X_n)) = 1 - \alpha$$

▷ din $P\left(\dfrac{\bar{X}_n - m}{\frac{\tilde{S}_n}{\sqrt{n}}} < t_{1-\alpha}\right) = 1 - \alpha$ avem : $\dfrac{\bar{X}_n - m}{\frac{\tilde{S}_n}{\sqrt{n}}} < t_{1-\alpha} \Leftrightarrow m > \bar{X}_n - \dfrac{\tilde{S}_n}{\sqrt{n}} \cdot t_{1-\alpha}$

• $g_1(X_1, \dots, X_n) = \bar{X}_n - \dfrac{\tilde{S}_n}{\sqrt{n}} \cdot t_{1-\alpha}$; valoarea intervalului de încredere este $\left(\bar{X}_n - \dfrac{\tilde{S}_n}{\sqrt{n}} \cdot t_{1-\alpha}\, , \infty\right)$

▷ din $P\left(\dfrac{\bar{X}_n - m}{\frac{\tilde{S}_n}{\sqrt{n}}} > t_{\alpha}\right) = 1 - \alpha$ avem : $\dfrac{\bar{X}_n - m}{\frac{\tilde{S}_n}{\sqrt{n}}} > t_{\alpha} \Leftrightarrow m < \bar{X}_n - \dfrac{\tilde{S}_n}{\sqrt{n}} \cdot t_{\alpha}$

• $g_2(X_1, \dots, X_n) = \bar{X}_n - \dfrac{\tilde{S}_n}{\sqrt{n}} \cdot t_{\alpha}$; valoarea intervalul de încredere este $\left(-\infty\, , \bar{x}_n - \dfrac{\tilde{s}_n}{\sqrt{n}} \cdot t_{\alpha}\right)$

Interval de încredere pentru varianţa (dispersia) $\sigma^2 = V(X)$ caracteristicii cercetate X

▶ se dau $\alpha \in (0, 1)$, datele statistice $x_1, \dots, x_n$
▶ dacă $X \sim N(m, \sigma^2)$, atunci $\frac{n-1}{\sigma^2}\tilde{S}_n^2 \sim \chi^2(n-1)$
▶ cuantilele χ^2 (Chi-pătrat) cu $n-1$ grade de libertate:

$c_{1-\frac{\alpha}{2}} = chi2inv(1-\frac{\alpha}{2}, n-1)$, $c_{\frac{\alpha}{2}} = chi2inv(\frac{\alpha}{2}, n-1)$, $c_{1-\alpha} = chi2inv(1-\alpha, n-1)$, $c_{\alpha} = chi2inv(\alpha, n-1)$

• *interval de încredere bilateral*: să se indice doi estimatori $g_1(X_1,\ldots,X_n)$ şi $g_2(X_1,\ldots,X_n)$ astfel încât pentru varianţa teoretică $m = E(X)$ să avem:

$$P\Big(g_1(X_1,\ldots,X_n) < \sigma^2 < g_2(X_1,\ldots,X_n)\Big) = 1-\alpha$$

▷ din $P\Big(c_{\frac{\alpha}{2}} < \frac{n-1}{\sigma^2}\cdot \tilde{S}_n^2 < c_{1-\frac{\alpha}{2}}\Big) = 1-\alpha$ avem : $c_{\frac{\alpha}{2}} < \frac{n-1}{\sigma^2}\cdot \tilde{S}_n^2 < c_{1-\frac{\alpha}{2}} \Leftrightarrow \frac{n-1}{c_{1-\frac{\alpha}{2}}}\cdot \tilde{S}_n^2 < \sigma^2 < \frac{n-1}{c_{\frac{\alpha}{2}}}\cdot \tilde{S}_n^2$

• *intervalul de încredere (bilateral)* pentru $\sigma^2 = V(X)$ (varianţa teoretică) este $\left(\frac{n-1}{c_{1-\frac{\alpha}{2}}}\cdot \tilde{S}_n^2 < \sigma^2 < \frac{n-1}{c_{\frac{\alpha}{2}}}\cdot \tilde{S}_n^2\right)$

▷ se calculează valoarea intervalului de încredere $\left(\frac{n-1}{c_{1-\frac{\alpha}{2}}}\cdot \tilde{s}_n^2 < \sigma^2 < \frac{n-1}{c_{\frac{\alpha}{2}}}\cdot \tilde{s}_n^2\right)$

• *interval de încredere unilateral*: să se indice doi estimatori $g_1(X_1,\ldots,X_n)$, $g_2(X_1,\ldots,X_n)$ astfel încât pentru varianţa teoretică $\sigma^2 = V(X)$ să avem:

$$P(g_1(X_1,\ldots,X_n) < \sigma^2) = 1-\alpha, \quad P(\sigma^2 < g_2(X_1,\ldots,X_n)) = 1-\alpha$$

▷ din $P\left(\frac{n-1}{\sigma^2}\cdot \tilde{S}_n^2 < c_{1-\alpha}\right) = 1-\alpha$ avem : $\frac{n-1}{\sigma^2}\cdot \tilde{S}_n^2 < c_{1-\alpha} \Leftrightarrow \sigma^2 > \frac{n-1}{c_{1-\alpha}}\cdot \tilde{S}_n^2$

• $g_1(X_1,\ldots,X_n) = \frac{n-1}{c_{1-\alpha}}\cdot \tilde{S}_n^2$; valoarea intervalului de încredere este $\left(\frac{n-1}{c_{1-\alpha}}\cdot \tilde{S}_n^2\,,\, \infty\right)$

▷ din $P\left(\frac{n-1}{\sigma^2}\cdot \tilde{S}_n^2 > c_{\alpha}\right) = 1-\alpha$ avem : $\frac{n-1}{\sigma^2}\cdot \tilde{S}_n^2 > c_{\alpha} \Leftrightarrow \sigma^2 < \frac{n-1}{c_{\alpha}}\cdot \tilde{S}_n^2$

• $g_2(X_1,\ldots,X_n) = \frac{n-1}{c_{\alpha}}\cdot \tilde{S}_n^2$; valoarea intervalul de încredere este $\left(-\infty\,,\, \frac{n-1}{c_{\alpha}}\cdot \tilde{s}_n^2\right)$

Metoda momentelor pentru estimarea parametrilor necunoscuţi $\theta = (\theta_1,\ldots,\theta_r)$ pentru distribuţia caracteristicii cercetate X

de exemplu:

$X \sim Exp(\lambda)$ parametrul necunoscut: $\theta = \lambda$
$X \sim N(m,\sigma^2)$ parametri necunoscuţi: $(\theta_1,\theta_2) = (m,\sigma)$
$X \sim Unif[a,b]$ parametri necunoscuţi: $(\theta_1,\theta_2) = (a,b)$

Fie $x_1,\ldots,x_n$ datele statistice pentru caracteristica cercetată X şi fie $X_1,\ldots,X_n$ variabilele de selecţie corespunzătoare.
Se rezolvă sistemul

$$\begin{cases} E(X^k) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i^k, \\ k = \{1,...,r\} \end{cases}$$

cu necunoscutele $\theta_1,\ldots,\theta_r$.
Soluţia sistemului $\hat{\theta}_1,\ldots,\hat{\theta}_r$ este estimatorul pentru parametrii necunoscuţi ai distribuţiei caracteristicii X.

Exemplu: Folosind metoda momentelor, să se estimeze parametrul necunoscut $\theta := a$ pentru $X \sim Unif[0,a]$; se dau datele statistice: 0.1,0.3,0.9,0.49,0.12,0.31,0.98,0.73, 0.13,0.62.

Avem cazul: $r = 1$, calculăm $E(X) = \frac{a}{2}$, $n = 10$, $\bar{x}_n = 0.468$. Se rezolvă

$$E(X) = \frac{1}{n}\sum_{i=1}^{n} x_i \Longrightarrow \frac{a}{2} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Estimatorul pentru parametrul necunoscut a este

$$\hat{a}(X_1,...,X_n) = \frac{2}{n}\sum_{i=1}^{n} X_i,$$

unde $X_1,\ldots,X_n$ sunt variabilele de selecţie. Valoarea estimatorului este

$$\hat{a}(x_1,...,x_n) = \frac{2}{n}\sum_{i=1}^{n} x_i = 0.936\,.$$

Parametrul necunoscut a este estimat cu valoarea 0.936.

► Este $\hat{a}(X_1, ..., X_n)$ un estimator nedeplasat pentru parametrul a?

Metoda verosimilităţii maxime pentru estimarea parametrului necunoscut θ al distribuţiei caracteristicii cercetate X

Fie $x_1, \ldots, x_n$ datele statistice pentru caracteristica cercetată X şi fie $X_1, \ldots, X_n$ variabilele de selecţie corespunzătoare. Notăm

$$L(x_1, \ldots, x_n; \theta) = \begin{cases} P(X = x_1) \cdot \ldots \cdot P(X = x_n), \text{ dacă } X \text{ e v.a. discretă} \\ f_X(x_1) \cdot \ldots \cdot f_X(x_n), \text{ dacă } X \text{ e v.a. continuă.} \end{cases}$$

Aceasta este funcţia de verosimilitate pentru parametrul θ şi datele statistice $x_1, \ldots, x_n$.

Metoda verosimilităţii maxime se bazează pe principiul că valoarea cea mai verosimilă (cea mai potrivită) a parametrului necunoscut θ este aceea pentru care funcţia de verosimilitate $L(x_1, \ldots, x_n; \theta)$ ia valoarea maximă:

$$(1) \qquad L(x_1, \ldots, x_n; \hat{\theta}) = \max_{\theta} L(x_1, \ldots, x_n; \theta).$$

Se rezolvă sistemul $\frac{\partial L}{\partial \theta} = 0$ şi se arată că $\frac{\partial^2 L}{\partial \theta^2} < 0$. Deseori este mai practic să se considere varianta transformată $\frac{\partial lnL}{\partial \theta} = 0$ cu $\frac{\partial^2 lnL}{\partial \theta^2} < 0$. În unele situaţii (1) se rezolvă prin alte metode.

Observaţie: Dacă distribuţia caracteristicii cercetate depinde de k parametri necunoscuţi $(\theta_1, \ldots, \theta_k)$ atunci se rezolvă sistemul

$$\frac{\partial L}{\partial \theta_j} = 0, j = \overline{1,k} \text{ şi se arată că matricea } \Big(\frac{\partial^2 L}{\partial \theta_i \partial \theta_j}\Big)_{1 \le i \le j \le k} \text{ este negativ definită.}$$

Se poate lucra şi cu varianta transformată:

$$\frac{\partial lnL}{\partial \theta_j} = 0, j = \overline{1,k} \text{ şi se arată că matricea } \Big(\frac{\partial^2 lnL}{\partial \theta_i \partial \theta_j}\Big)_{1 \le i \le j \le k} \text{ este negativ definită.}$$

O matrice M este negativ definită dacă $y^t M y < 0$ pentru orice $y \in \mathbb{R}^n \setminus \{0_n\}$.

Exemplu: Folosind metoda verosimilităţii maxime să se estimeze parametrul $\theta := p \in (0, 1)$ al distribuţiei Bernoulli, $X \sim \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$, cu datele statistice: 0,1,1,0,0,0,1,0,1,0.

$\Rightarrow n = 10, x_1 = 0, x_2 = 1, x_3 = 1, x_4 = 0...; P(X = x) = p^x(1-p)^{1-x}, x \in \{0, 1\}$

$$\Rightarrow L(x_1, \ldots, x_n; p) = P(X = x_1) \cdot \ldots \cdot P(X = x_n) = p^{x_1+\cdots+x_n}(1-p)^{n-(x_1+\cdots+x_n)}$$

$$\Rightarrow lnL(x_1, \ldots, x_n; p) = (x_1 + \cdots + x_n)ln(p) + (n - (x_1 + \cdots + x_n))ln(1-p)$$

$$\frac{\partial lnL}{\partial p} = 0 \Rightarrow p = \frac{1}{n}(x_1 + \cdots + x_n).$$

Are loc: $\frac{\partial^2 lnL}{\partial p^2} < 0$.

Estimatorul de verosimilitate maximă pentru parametrul necunoscut p este

$$\hat{p}(X_1, \ldots, X_n) = \frac{1}{n}(X_1 + \cdots + X_n) = \bar{X}_n,$$

unde $X_1, \ldots, X_n$ sunt variabilele de selecţie. Valoarea estimată este

$$\hat{p}(x_1, \ldots, x_n) = \frac{1}{n}(x_1 + \cdots + x_n) = \bar{x}_n = \frac{4}{10} = 0.4\,.$$

► Este $\hat{p}(X_1, ..., X_n)$ un estimator nedeplasat pentru parametrul p?