Curs 11

Statistică inferenţială

Statistica inferenţială se concentrază asupra luării deciziilor despre caracteristicele unei populaţii bazate pe informaţiile obţinute prin studiul datelor uneia sau mai multor selecţii aleatoare şi pe calcule probabilistice.

Statistica inferenţială se împarte în două mari domenii:

– estimarea parametrilor şi

– ipoteze statistice.

Metodele de inferenţă se bazează pe respectarea caracterului aleator al selecţiei.

Fie X o v. a. cu *legea de probabilitate* $f(x,\theta)$, $x \in \mathbb{R}$, $\theta \in \Theta \subset \mathbb{R}$, care poate fi funcţia de frecvenţă, în cazul discret, respectiv densitatea de probabilitate, în cazul continuu.

Definiţia 11.0.1 *Mulţimea legilor de probabilitate $f(x,\theta)$ care conţin parametrul necunoscut θ se numeşte* model probabilistic.

În practică există numeroase astfel de familii de legi de probabilitate care pot fi alese ca modele probabilistice: normală, exponentială, Poisson etc. Alegerea unei astfel de familii, ca model al unui fenomen particular real se face pe baza experienţei anterioare în studiul fenomenelor asemănătoare sau în urma unui studiu preliminar al rezultatelor experienţei sau observaţiei.

În momentul în care am ales o lege de probabilitate de o anumită formă, incertitudinea legată de rezultatul particular al experimentului s-a transferat în incertitudinea legată de valorile parametrului (parametrilor).

Definiţia 11.0.2 Repartiţia selecţiei $X = (X_1, X_2, \ldots, X_n)$ *este definită ca repartiţia comună a variabilelor de selecţie* $X_1,\ X_2,\ \ldots,\ X_n$. *Dacă* $x_1,\ x_2, \ldots, x_n$ *sunt valorile luate de variabilele de selcţie* $X_1, X_2, \ldots, X_n$ *atunci* legea de probabilitate a selecţiei *este notată prin*

$$f(x_1, x_2, \ldots, x_n; \theta), (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n.$$

Repartiţia selecţiei depinde de θ şi încorporează atât selecţia cât şi modelul probabilistic.

Cea mai folosită formă de selecţie este selecţia aleatoare şi este bazată pe ideea experimentului aleator.

Definiţia 11.0.3 *Vom spune că* $(X_1, X_2, \ldots, X_n)$ *este o* selecţie aleatoare *asupra v. a. X care are legea de probabilitate $f(x,\theta)$ dacă* $X_1,\ X_2,\ \ldots,\ X_n$ *sunt independente şi identic distribuite (i.i.d.) ca şi X.*

În cazul selecţiei aleatoare legea de probabilitate a selecţiei este

$$f(x_1, x_2, \ldots, x_n; \theta) = \prod_{j=1}^{n} f(x_j, \theta).$$

O selecţie aleatoare poate fi constituită prin repetarea unui experiment aleator de n ori. Realizarea selecţiei aleatoare (datele obţinute în urma selecţiei) se notează cu $(x_1, x_2, \ldots, x_n)$ şi mulţimea tuturor realizărilor defineşte spaţiul observaţiilor.

Definiţia 11.0.4 *Modelul probabilist $f(x, \theta)$ împreună cu selecţia aleatoare*

$$X = (X_1, X_2, \ldots, X_n)$$

definesc modelul statistic.

Modelul statistic împreună cu datele observate impun considerarea următoarelor întrebări:

1. datele observate sunt consistente cu modelul statistic postulat?

2. presupunând că modelul statistic postulat este consistent cu datele observate, ce putem spune despre parametrii necunoscuţi $\theta \in \Theta$?

a) putem descreşte incertitudinea asupra lui θ prin reducerea spaţiului parametrilor Θ la Θ_0 unde Θ_0 este o submulţime a lui Θ? (estimaţie prin intervale de încredere)

b) putem descreşte incertitudinea asupra lui θ prin alegerea unei valori particulare $\hat{\theta}$ din Θ ca având cea mai reprezentativă valoare a lui θ? (estimaţie punctuală)

c) putem considera întrebarea: θ aparţine unei submulţimi Θ_0 a lui Θ? (verificarea ipotezelor statistice)

3. presupunând că a fost aleasă o valoare particulară $\hat{\theta}$ reprezentativă a lui θ, ce putem spune despre observaţiile viitoare? (predicţie)

Definiţia 11.0.5 *Dată o populaţie de volum N, se numeşte* statistică *a acelei populaţii orice mărime semnificativă calculată cu ajutorul datelor obţinute dintr-un eşantion aleator al acelei populaţii.*

Exemple de statistici frecvent folosite: media de selecţie statistică, dispersia de selecţie statistică, dispersia de selecţie modificată.

Definiţia 11.0.6 *Pentru un parametru al repartiţiei, se numeşte* estimator *al acelui parametru orice statistică care aproximează acel parametru.*

Dacă considerăm selecţia aleatoare $X = (X_1, X_2, \ldots, X_n)$ cu datele obţinute în urma selecţiei $(x_1, x_2, \ldots, x_n)$, vom nota estimatorul parametrului θ cu $\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n)$ şi atunci putem scrie

$$\forall \varepsilon > 0 : \lim_{n \to \infty} P\left(\left|\hat{\theta}(x_1, x_2, \ldots, x_n) - \theta\right| > \varepsilon\right) = 0$$

Cu alte cuvinte, pentru $\varepsilon > 0$ dat, pentru valori mari ale lui n este foarte puţin probabil ca $\hat{\theta}(x_1, x_2, \ldots, x_n)$ să ia valori în afara intervalului $[\theta - \varepsilon, \theta + \varepsilon]$, adică este foarte puţin probabil ca numărul $\hat{\theta}(x_1, x_2, \ldots, x_n)$ să fie în afara intervalului $[\theta - \varepsilon, \theta + \varepsilon]$. În aceste condiţii, după un număr de n experienţe, considerăm pe $\hat{\theta}(x_1, x_2, \ldots, x_n)$ ca o aproximaţie bună pentru θ. Este posibil să ne înşelăm, dar probabilitatea de a ne înşela este mică, pentru n mare. Statistica nu ne oferă răspunsuri sigure ci doar aproximaţii în care putem avea un grad mai mic sau mai mare de încredere. Se acceptă acele aproximaţii în care avem un grad mai mare de încredere.

Se utilizează *estimări punctuale* sau *estimări cu ajutorul intervalelor de încredere*. Estimatorul punctual atribuie o valoare parametrului. Intervalele de încredere generează un interval în care parametrul aparţine cu un anumit nivel de încredere.

Exemplul 11.0.7 Considerăm populaţia diametrelor unor bile de rulmenţi dintr-un lot. Ca statistică a acestei populaţii se poate lua media de selecţie $\overline{x}$ a unui şir $(x_1, x_2, \ldots, x_n)$ format de diametrele unui eşantion de bile din acel lot. Considerând alt lot, media de selecţie $\overline{x}$ va avea altă valoare.

Exemplul 11.0.8 O societate de telefoane poate considera populaţia tuturor abonaţilor. Un eşantion se poate obţine luând pe sărite din lista abonaţilor, de exemplu, din 25 în 25. Presupunem că societatea este interesată în estimarea unui parametru caracteristic; de exemplu, gradul de satisfacţie g al abonaţilor relativ la serviciile oferite de socitate. Se cere o apreciere cu note de la 1 la 5. Este solicitat răspunsul abonaţilor din eşantionul ales şi se obţine un estimator al lui g.

Dacă, de exemplu, θ reprezintă media atunci putem considera ca statistică a mediei, $\theta = \overline{X} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$. Fie $x_1, x_2, \ldots, x_n$ sunt valorile luate de $X_1, X_2, \ldots, X_n$ atunci estimatorul lui θ este $\hat{\theta} = \overline{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$. În condiţiile în care măsurătorile sunt realizate cu acurateţe şi precizie rezonabile, atunci $X_1, X_2, \ldots, X_n$ se pot considera independente şi identic repartizate.

Vom prezenta exemple de modele statistce care apar adesea în practică.

11.1 Modele statistice

Modelul Bernoulli

Presupunem că $X_1, X_2, \ldots, X_n$ este o selecţie aleatoare a unei populaţii care urmează o distribuţie Bernoulli cu parametrul $\theta \in [0, 1]$ necunoscut. Ştim că distribuţia Bernoulli poate fi utilizată la observarea unui articol, rezultat al unui proces industrial, caz în care $X_i = 1$ indică faptul că articolul i este bun şi $X_i = 0$ dacă articolul este defect. În studii medicale rezultatul $X_i = 1$ indică faptul că tratamentul aplicat pacientului i a avut succes iar $X_i = 0$ în caz contrar. În aceste cazuri vrem să ştim valoarea lui θ.

Spaţiul parametrului θ este $[0, 1]$. Dacă $x_1, x_2, \ldots, x_n$ sunt valorile luate de $X_1, X_2, \ldots, X_n$ atunci *funcţia de frecvenţă* pentru X_i este

$$f(x_i, \theta) = \theta^{x_i}(1 - \theta)^{1-x_i},$$

şi reprezintă $P(X = x_i)$, θ este parametrul necunoscut, iar legea de probabilitate a selecţiei este dată de

$$\prod_{i=1}^{n} f(x_i, \theta) = \prod_{i=1}^{n} \theta^{x_i}(1 - \theta)^{1-x_i} = \theta^{n\overline{x}}(1 - \theta)^{n(1-\overline{x})}.$$

Modelul Poisson

Presupunem că $X_1, X_2, \ldots, X_n$ este o selecţie aleatoare a unei populaţii care urmează o distribuţie Poisson cu parametrul $\theta \in \mathbb{R}_+$ necunoscut. Ştim că distribuţia Repartiţia Poisson a apărut odată cu studiul variabilelor aleatoare cu un număr foarte mare de valori. Exemple clasice de astfel de variabile aleatoare sunt cele legate de **evenimentele care apar rar** într-un interval de timp. În aceste cazuri vrem să ştim valoarea lui θ.

Spaţiul parametrului θ este $\mathbb{R}_+$. Dacă $x_1, x_2, \ldots, x_n$ sunt valorile luate de $X_1, X_2, \ldots, X_n$ atunci *funcţia de frecvenţă* pentru X_i este

$$f(x_i, \theta) = \frac{\theta^{x_i}}{x_i!} e^{-\theta},$$

şi reprezintă $P(X = x_i)$, θ este parametrul necunoscut, iar legea de probabilitate a selecţiei este dată de

$$\prod_{i=1}^{n} f(x_i, \theta) = \prod_{i=1}^{n} \frac{\theta^{x_i}}{x_i!} e^{-\theta} = \frac{\theta^{n\overline{x}}}{x_1! x_2! ... x_n!} e^{n}.$$

Modelul normal

Presupunem că $X_1, X_2, \ldots, X_n$ este o selecţie aleatoare a unei populaţii care urmează o distribuţie normală $N(\mu, \sigma)$ cu $\theta = (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+$ necunoscuţi, unde $\mathbb{R}^+ = (0, \infty)$. De exemplu, putem avea observaţii asupra înălţimii în centimetrii a unei populaţii şi intuim că este rezonabil să presupunem că distribuţia înălţimii este normală cu media şi dispersia necunoscute. Dacă $x_1, x_2, \ldots, x_n$ sunt valorile luate de $X_1, X_2, \ldots, X_n$ atunci densitatea de probabiltate a unui eşantion este dată de

$$\prod_{i=1}^{n} f(x_i; \mu, \sigma^2) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right)$$

Deoarece

$$\sum_{i=1}^{n} (x_i - \mu)^2 = n(\overline{x} - \mu)^2 + \sum_{i=1}^{n} (x_i - \overline{x})^2,$$

unde $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ este media de selecţie şi $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$ este dispersia de selecţie modificată (deviaţia standard), obţinem

$$\prod_{i=1}^{n} f(x_i; \mu, \sigma^2) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{n}{2\sigma^2} (\overline{x} - \mu)^2 - \frac{n-1}{2\sigma^2} s^2 \right). \tag{11.1}$$

Ne putem pune întrebarea dacă presupunerea că înălţimea populaţiei urmează o distribuţie normală este corectă. Această ipoteză trebuie verificată. Procedurile care duc la verificarea unor astfel de ipoteze se numesc teste statistice.

Modelul exponenţial

Presupunem că timpul de viaţă a unui aparat este distribuit $Exp[\theta]$ unde $\theta \in (0, \infty)$ este necunoscut. Presupunem că $X_1, X_2, \ldots, X_n$ este o selecţie aleatoare a unei populaţii care urmează o distribuţie exponenţială. Dacă $x_1, x_2, \ldots, x_n$ sunt valorile luate de $X_1, X_2, \ldots, X_n$ atunci densitatea de probabiltate a unui eşantion este dată de

$$\prod_{i=1}^{n} f(x_i; \theta) = \theta^n \exp\{-n\overline{x}\theta\}.$$

Modelul multinomial

În capitolul II am introdus distribuţia multinomială. Aceasta apare în aplicaţii în care în urma unui experiment se pot realiza k evenimente diferite independente $A_1, A_2, \ldots, A_k$. Aceste evenimente se produc cu probabilităţile $p_1, p_2, \ldots, p_k$, $\sum_{j=1}^{k} p_j = 1$, $0 \leq p_j \leq 1$.

Facem presupunerea că, de exemplu, $k = 3$. În acest caz nu cunoaştem parametrii (p_1, p_2, p_3). Dacă x_1, x_2, x_3 sunt valorile luate de X_1, X_2, X_3 atunci densitatea de probabiltate a unui eşantion este dată de

$$\prod_{i=1}^{3} f(x_i; p_1, p_2, p_3) = p_1^{x_1} p_2^{x_2} p_3^{x_3}.$$

Construcţia histogramei şi încercarea de a intui modelul statistic sunt metode folosite de statistician în încercarea de a studia distribuţia populaţiei. Deseori nu este clar ce model statistic trebuie folosit. Mai mult, dacă am intuit modelul statistic, nu ştim nimic despre valoarea parametrilor care intra în definiţia modelului statistic. De exemplu, ştim că înălţimea urmează o distribuţie normală, dar nu ştim media şi dispersia. Folosind intuiţia am putea presupune că media poate fi aproximată cu $\overline{x}$. Pentru a justifica această alegere avem nevoie să studiem estimatorii punctuali.

11.2 Estimatori punctuali

Estimarea punctuală este valoarea atribuită unui parametru pe baza statisticii construite din eşantion.

Definiţia 11.2.1 *Se consideră o populaţie de volum N şi un parametru θ al acestei populaţii. Fie $X_1, X_2, \ldots, X_n$ este o selecţie aleatoare dintr-un eşantion reprezentativ ($n << N$) al populaţiei care ia valorile $x_1, x_2, \ldots, x_n$; un estimator punctual $\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n)$ a lui θ se numeşte* estimator nedeplasat sau absolut corect *dacă $M[\hat{\theta}] = \theta$.*

În practică se recomandă $n \geq 30$ şi $n \leq \frac{N}{3}$.

Definiţia 11.2.2 *Dacă $\lim\limits_{n\to\infty} M[\hat{\theta}] = \theta$ şi $\lim\limits_{n\to\infty} D[\hat{\theta}] = 0$ atunci $\hat{\theta}$ este un* estimator corect sau deplasat *al lui θ (în cazul estimatorului consistent limita trebuie să fie oricât de mică atunci când n creşte).*

Teorema 11.2.3 *Dacă statistica $\hat{\theta}(x_1, x_2, \ldots, x_n)$ este un estimator corect a lui θ atunci este un estimator al lui θ, adică pentru orice $\varepsilon > 0$ avem*

$$\lim_{n\to\infty} P\left(\left|\hat{\theta}(x_1, x_2, \ldots, x_n) - \theta\right| > \varepsilon\right) = 0.$$

Demonstraţie. Conform inegalităţii lui Cebâşev,

$$P\left(\left|\hat{\theta}(x_1, x_2, \ldots, x_n) - M[\hat{\theta}]\right| > \varepsilon\right) \leq \frac{D[\hat{\theta}]}{\varepsilon^2},$$

dar $\lim\limits_{n\to\infty} M[\hat{\theta}] = \theta$ şi $\lim\limits_{n\to\infty} D[\hat{\theta}] = 0$ de unde rezultă concluzia. □

Arătăm că statisticile introduse anterior sunt estimatori ale parametrilor corespunzători.

Teorema 11.2.4 *Fie $x_1, x_2, \ldots, x_n$ ($n \geq 2$) o selecţie de valori ale variabilelor $X_1, X_2, \ldots, X_n$ (v. a. independente şi identic distribuite ca şi X). Notăm*

$m = M[X]$ *media teoretică,*

$\sigma^2 = D[X]$ *dispersia teoretică,*

$\overline{X} = \dfrac{1}{n}\sum\limits_{i=1}^{n} X_i$ *media de selecţie,*

$\overline{\sigma^2} = \dfrac{1}{n}\sum\limits_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ *dispersia de selecţie şi*

$S^2 = \dfrac{1}{n-1}\sum\limits_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ *dispersia de selecţie modificată.*

Atunci media de selecţie $\overline{X}$ şi dispersia de selecţie S^2 sunt estimatori punctuali pentru media, respectiv dispersia teoretică. În plus,

1. *media de selecţie $\overline{X}$ este un estimator absolut corect al lui m;*
2. *dispersia de selecţie $\overline{\sigma^2}$ este un estimator corect al lui σ^2.*
3. *dispersia de selecţie modificată S^2 este un estimator absolut corect al lui σ^2.*

Demonstraţie. 1. Trebuie verificate condiţiile din definiţia estimatorului absolut corect.

$$M\left[\overline{X}\right] = \frac{1}{n} M\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n} nM\left[X\right] = M\left[X\right].$$

$$D\left[\overline{X}\right] = D\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2} D\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} D\left[X_i\right] = \frac{\sigma^2}{n}.$$

2.

$$\begin{aligned} M\left[\overline{\sigma^2}\right] &= \frac{1}{n} M\left[\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right] = \\ &= \frac{1}{n} M\left[\sum_{i=1}^{n} X_i^2 - 2\overline{X}\sum_{i=1}^{n} X_i + n\overline{X}^2\right] = \\ &= \frac{1}{n} M\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right] = M\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2\right] - M\left[\overline{X}^2\right]. \end{aligned}$$

Dar

$$M\left[\overline{X}^2\right] = \frac{1}{n^2} M\left[\left(\sum_{i=1}^{n} X_i\right)^2\right] = \frac{1}{n^2} M\left[\sum_{i=1}^{n} X_i^2\right] + \frac{2}{n^2}\sum_{i<j}^{n} M\left[X_i X_j\right].$$

Apoi $M\left[X_i^2\right] = M\left[X^2\right]$ şi din ipoteza de independenţă

$$M\left[X_i X_j\right] = M\left[X_i\right] M\left[X_j\right] = \left(M\left[X\right]\right)^2,$$

rezultă

$$M\left[\overline{X}^2\right] = \frac{1}{n^2} M\left[X^2\right] + \frac{n-1}{n}\left(M\left[X\right]\right)^2.$$

Deci

$$\begin{aligned} M\left[\overline{\sigma^2}\right] &= \frac{1}{n} M\left[X^2\right] - \left(\frac{1}{n^2} M\left[X^2\right] + \frac{n-1}{n}\left(M\left[X\right]\right)^2\right) = \\ &= \frac{n-1}{n}\left(M\left[X^2\right] - \left(M\left[X\right]\right)^2\right) = \frac{n-1}{n}\sigma^2 \to \sigma^2. \end{aligned}$$

Rezultă că $\overline{\sigma^2}$ nu estimează absolut corect dispersia v. a. X.

3. Reţinem că $M\left[\overline{\sigma^2}\right] = \dfrac{n-1}{n}\sigma^2$.

Rezultă că $M\left[S^2\right] = M\left[\dfrac{n}{n-1}\overline{\sigma^2}\right] = \sigma^2$. Se poate demonstra că

$$D\left[S^2\right] = \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^4 - \frac{n-3}{n-1} D\left[X\right]^2\right), \tag{11.2}$$

de unde rezultă că $D\left[S^2\right] \to 0$. □

Observaţia 11.2.5 *Deşi dispersia de selecţie S^2 este un estimator nedeplasat a lui σ^2, S este un estimator deplasat a lui σ. Pentru valori mari $n \geq 10$ se poate considera că $\hat{\sigma} = \left(1 + \frac{1}{[4(n-1)]}\right) S$ este un estimator nedeplasat a lui σ.*

Observaţia 11.2.6 *În practică se foloseste S^2 în locul lui $\overline{\sigma^2}$ deoarece dă rezultate mai bune după cum ne arată Teorema 11.2.4. Totuşi formula (11.2) ne spune că pentru n suficient de mare si statistica $\overline{\sigma^2}$ poate fi folosită ca estimator al dispersiei v.a. X.*

Următorul tabel sintetizează relaţia dintre parametrii necunoscuţi şi estimatorii punctuali asociaţi. Astfel μ notează media, σ^2 este dispersia, p este probabilitatea repartiţiei binomiale, $p = \frac{k}{n}$, unde k este numărul realizărilor într-un eşantion de volum n, $\mu_1 - \mu_2$ estimează diferenţa dintre mediile a două eşantioane independente a două statistici aleatoare, $p_1 - p_2$ estimează diferenţa dintre proporţiile a două eşantioane independente a două statistici aleatoare.

Param. nec.	Statistica	Estim. punctual
μ	$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
σ^2	$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$
p	$\hat{p} = \frac{1}{n}X$	$\hat{p} = \frac{1}{n}x$
$\mu_1 - \mu_2$	$\overline{X_1} - \overline{X_2} = = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1i} - \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2i}$	$\overline{x}_1 - \overline{x}_2 = = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i} - \frac{1}{n_2}\sum_{i=1}^{n_2} x_{2i}$
$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2 = \frac{1}{n_1}X_1 - \frac{1}{n_2}X_2$	$\hat{p}_1 - \hat{p}_2 = \frac{1}{n_1}x_1 - \frac{1}{n_2}x_2$

Observaţia 11.2.7 *Dacă $X \in Poiss\,[\lambda]$, atunci $M\,[X] = D\,[X] = \lambda$ şi atunci $\overline{x}$ şi s^2 sunt estimatorii nedeplasaţi pentru λ. Se pune întrebarea pe care îl vom alege. Cum dispersia este o măsură a împrăştierii, intuiţia sugerează să-l alegem pe acela care are cea mai mică dispersie şi aceasta deoarece el are o repartiţie mai concentrată în jurul lui λ.*

Dintre doi estimatori nedeplasaţi $\hat{\theta}$ şi $\hat{\theta}_1$ pentru acelaşi parametru θ, se consideră că este mai precis cel care are dispersie mai mică.

Teorema 11.2.8 *Fie X o v. a. cu media m şi abaterea medie pătratică σ. Dacă $x_1, x_2, \ldots, x_n$ $(n \geq 2)$ o selecţie de valori ale lui X (v. a. independente la fel distribuite ca şi X), atunci pentru n suficient de mare se poate considera că $\overline{X} \in N\left(m, \frac{\sigma}{\sqrt{n}}\right)$.*

Demonstraţie. Conform teoremei limită centrală avem că

$$S_n = X_1 + X_2 + \cdots + X_n \Rightarrow S_n \in N\left(nm, \sigma\sqrt{n}\right).$$

Deoarece

$$M\left[\frac{1}{n}S_n\right] = \frac{1}{n}nm = m, \quad D\left[\frac{1}{n}S_n\right] = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n},$$

rezultă că $\frac{1}{n}S_n \in N\left(m, \frac{\sigma}{\sqrt{n}}\right)$. □

Corolarul 11.2.9 *În condiţiile Teoremei 11.2.8 pentru orice $a < b$ avem*

$$P\left(a \leq \overline{X} < b\right) \approx \Phi\left(\frac{b-m}{\frac{\sigma}{\sqrt{n}}}\right) - \Phi\left(\frac{a-m}{\frac{\sigma}{\sqrt{n}}}\right). \tag{11.3}$$

Statisticienii recomandă folosirea formulei (11.3) pentru populaţii statistice de volum N şi eşantioane de volum n unde $n \geq 30$ şi $n \leq \frac{N}{3}$. Dacă $n < 30$ formula este utilă dacă populaţia iniţială nu este departe de a fi normal distribuită. Pentru eşantioane mici statisticienii propun o corecţie care să înlocuiască $\frac{\sigma}{\sqrt{n}}$ prin $\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$.

Exerciţiul 1 Considerăm greutăţile populaţiei de 350 studenţi dintr-o facultate s-a constatat că media este $m = 70$ şi abaterea media pătratică este $\sigma = 10$.

a) Să se determine probabilitatea ca un student luat la întâmplare să cântărească între 65 şi 70 kg?

b) Să se determine probabilitatea ca media maselor să fie între 65 şi 75 kg.

c) Să se determine probabilitatea ca extrăgând un eşantion de 36 studenţi din acea populaţie, media eşantionului să fie cuprinsă între 65 şi 75 kg.

d) Extrăgînd un eşantion de 100 de studenţi care este probabilitatea ca media maselor să fie sub 65 kg?

Rezolvare. *a*) Pobabilitatea cerută nu poate fi calculată deoarece $X =$ greutatea populaţiei nu este repartizată normal.

La fel şi la *b*).

c) $n = 36$, $m = 70$, $\sigma_{\overline{X}} = \frac{10}{\sqrt{36}} = 1.6667$,

$$P\left(65 \leq \overline{X} \leq 75\right) = \Phi\left(\frac{75-70}{1.6667}\right) - \Phi\left(\frac{65-70}{1.6667}\right) = \\ = 0.9986 - 0.0014 = 0.9972.$$

d) $n = 100$, $m = 70$, $\sigma_{\overline{X}} = \frac{10}{\sqrt{100}} = 1$,

$$P\left(\overline{X} \leq 65\right) = \Phi\left(\frac{65-70}{1}\right) = 0.0000003. \qquad \Diamond$$

11.3 Metoda momentelor

Ideea metodei este de a egala momentele distribuţiei cu momentele statistice şi de aici obţinem estimarea parametrilor. Momentele populaţiei vor depinde de parametrii necunoscuţi.

De exemplu momentul iniţial de ordin întâi este media, iar media statistică este momentul statistic de ordin întâi. Egalăm acestea şi obţinem estimarea punctuală $\hat{\mu} = \overline{X}$. Deci media statistică este un estimator punctual al mediei populaţiei.

Fie $x_1, x_2, \ldots, x_n$ $(n \geq 2)$ o selecţie de valori ale variabilelor X_1, X_2, $\ldots$, X_n care au o distribuţie exponenţială cu parametru λ. Avem un singur parametru de estimat. Ştim că $M[X] = \frac{1}{\lambda}$. De aici rezultă că $\frac{1}{\lambda} = \overline{X} \Rightarrow \hat{\lambda} = \frac{1}{\overline{x}}$ este un estimator punctual al parametrului λ obţinut cu metoda momentelor.

Exemplul 11.3.1 Ca exemplu, presupunem că este testat timpul de viaţă a unui modul electronic folosit în industria automobilelor. Timpul de viaţă urmează o distribuţie exponenţială. S-au făcut teste şi s-au obţinut următoarele date: $x_1 = 11.96$, $x_2 = 5.03$, $x_3 = 67.40$, $x_4 = 16.07$, $x_5 = 31.50$, $x_6 = 7.73$, $x_7 = 11.10$, $x_8 = 22.38$. Deoarece $\overline{x} = 21.65$, estimarea punctuală a parametrului λ obţinut cu metoda momentelor este $\hat{\lambda} = \frac{1}{21.65} = 0.046189$.

Fie $x_1, x_2, \ldots, x_n$ $(n \geq 2)$ o selecţie de valori ale variabilelor X_1, X_2, $\ldots$, X_n care au o distribuţie normală de parametrii μ şi σ. În acest caz avem $M[X] = \mu$, $M[X^2] = \mu^2 + \sigma^2$.

Rezultă că

$$\mu = \frac{1}{n}\sum_{i=1} x_i, \mu^2 + \sigma^2 = \frac{1}{n}\sum_{i=1} x_i^2 \Rightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1} x_i,$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1} x_i^2 - \left(\frac{1}{n}\sum_{i=1} x_i\right)^2 \Rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1} (x_i - \overline{x})^2 .$$

Reamintim că acest estimator pentru σ^2 este un estimator corect sau deplasat.

11.4 Metoda verosimilităţii maxime

Una din cele mai bune metode de a obţine un estimator punctual pentru un parametrul θ este metoda verosimilităţii maxime.

Metoda verosimilităţii maxime a fost dezvoltată în 1920 de matematicianul englez R. A Fisher şi extinsă de Cramer, Rao şi Wald. Este cea mai folosită metodă de estimare şi joacă un rol important în verificarea ipotezelor statistice.

Fie v. a. X cu densitatea de probabilitate $f(x, \theta)$, unde θ este parametrul necunoscut. Considerăm o selecţie $X_1, X_2, \ldots, X_n$ i.i.d cu v. a. X şi fie $x_1, x_2, \ldots, x_n$ valorile variabilelor de selecţie. Atunci *funcţia de verosimilitate maximă* corespunzătoare selecţiei este

$$L(x_1, x_2, \ldots, x_n; \theta) = \prod_{j=1}^{n} f(x_j, \theta).$$

Această funcţie se utilizează numai dacă avem un singur parametru necunoscut θ. Estimatorul de verosimilitate maximă este acea valoare a lui θ care maximizează funcţia $L(x_1, x_2, \ldots, x_n; \theta)$. În cazul unei v. a. discrete interpretarea funcţiei este

$$L(x_1, x_2, \ldots, x_n; \theta) = P\left(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right),$$

adică reprezintă tocmai probabilitatea de a obţine în urma selecţiei valorile $x_1, x_2, \ldots, x_n$.

Metoda verosimiltăţii maxime constă în următorul principiu (axiomă): valoarea "cea mai verosimilă" (cea mai potrivită în acest sens!) a parametrului este aceea pentru care funcţia $L(x_1, x_2, \ldots, x_n; \theta)$ este maximă. Aceasta implică θ punct în care

$$\frac{\partial L(x_1, x_2, \ldots, x_n; \theta)}{\partial \theta} = 0. \tag{11.4}$$

Ecuaţia (11.4) în practică se dovedeşte a fi dificilă. De aceea se foloseşte observaţia: $L(x_1, x_2, \ldots, x_n; \theta)$ este maximă dacă şi numai dacă $\ln L(x_1, x_2, \ldots, x_n; \theta)$ este maximă.

Introducem funcţiile: *logaritmul funcţiei de verosimilitate*
$\ln L(x_1, x_2, \ldots, x_n; \theta)$ şi *funcţia scor*

$$\frac{\partial L(x_1, x_2, \ldots, x_n; \theta)}{\partial \theta} = s\left(x_1, x_2, \ldots, x_n; \theta\right).$$

Acestea încoporează aceeaşi informaţie ca şi $L(x_1, x_2, \ldots, x_n; \theta)$. Dacă avem oricare dintre ele putem să o obţinem pe cealaltă.

Definiţia 11.4.1 $t_n(x_1, x_2, \ldots, x_n) : S \to \Theta$ *se numeşte* estimaţie de verosimilitate maximă *pentru parametrul* θ *dacă este punct de maxim pentru funcţia de verosimilitate, adică*

$$\ln L(x_1, x_2, \ldots, x_n; t_n) \geq \ln L(x_1, x_2, \ldots, x_n; \theta)$$

pentru toţi $\theta \in \Theta$.

În cele ce urmează presupunem că $f(x, \theta)$ este derivabilă până la ordinul doi inclusiv în raport cu θ. Atunci există $\dfrac{\partial L(x_1, x_2, \ldots, x_n; \theta)}{\partial \theta}$ şi estimaţia de verosimilitate maximă este soluţie a ecuaţiei

$$\frac{\partial L(x_1, x_2, \ldots, x_n; \theta)}{\partial \theta} = 0 \tag{11.5}$$

sau a ecuaţiei

$$\frac{\partial \ln L(x_1, x_2, \ldots, x_n; \theta)}{\partial \theta} = 0. \tag{11.6}$$

Fie această soluţie θ'. Pentru a fi estimator de verosimilitate maximă trebuie să satisfacă şi condiţia

$$\left.\frac{\partial^2 \ln L(x_1, x_2, \ldots, x_n; \theta)}{\partial \theta^2}\right|_{\theta=\theta'} < 0.$$

Definiţia 11.4.2 *Ecuaţiile (11.5) şi (11.6) se numesc* ecuaţii de verosimilitate.

Exemplul 11.4.3 Fie $x_1, x_2, \ldots, x_n$ $(n \geq 2)$ o selecţie de valori ale variabilelor $X_1,\ X_2,\ \ldots,\ X_n$ care au o distribuţie binomilală, i. i. r. cu v. a. $X \in Bernoulli\,[\theta]$. Atunci

$$f(x, \theta) = \begin{cases} \theta^x (1-\theta)^{1-x}, & x = 0, 1 \\ 0, & \text{în rest.} \end{cases}.$$

θ este parametrul care trebuie estimat. Funcţia de verosimilitate maximă este, pentru selecţia făcută,

$$\begin{aligned} L(x_1, x_2, \ldots, x_n; \theta) &= \theta^{x_1}(1-\theta)^{1-x_1}\theta^{x_2}(1-\theta)^{1-x_2}\ldots\theta^{x_n}(1-\theta)^{1-x_n} = \\ &= \prod_{j=1}^{n}\theta^{x_j}(1-\theta)^{1-x_j} = \theta^{n\overline{x}}(1-\theta)^{n-n\overline{x}}. \end{aligned}$$

Observăm că θ minimizează $L(x_1, x_2, \ldots, x_n; \theta)$ dacă minimizează şi $\ln L(x_1, x_2, \ldots, x_n; \theta)$. Atunci

$$\begin{aligned} \ln L(x_1, x_2, \ldots, x_n; \theta) &= n\overline{x}\ln\theta + (n - n\overline{x})\ln(1-\theta). \\ \frac{\partial \ln L(x_1, x_2, \ldots, x_n; \theta)}{\partial \theta} &= \frac{n\overline{x}}{\theta} - \frac{n - n\overline{x}}{1-\theta} \\ \frac{n\overline{x}}{\theta} - \frac{n - n\overline{x}}{1-\theta} &= 0 \Rightarrow \hat{\theta} = \overline{x}. \end{aligned}$$

De aici rezultă că estimatorul de verosimilitate maximă pentru p este

$$\hat{\theta} = \frac{1}{n}\sum_{j=1}^{n} x_j. \qquad \Diamond$$

Exemplul 11.4.4 Fie $x_1, x_2, \ldots, x_n$ $(n \geq 2)$ o selecţie de valori ale variabilelor de selecţie $X_1, X_2, \ldots, X_n$ care au o distribuţie normală, i. i. d. cu v. a. $X \in N(m, \sigma)$.

a) Dacă σ^2 este cunoscută şi media m este necunoscută cu $\Theta = R$, atunci funcţia de verosimilitate maximă este

$$L(x_1, x_2, \ldots, x_n; m) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2\right\}.$$

Deoarece funcţia $\ln L$ este strict crescătoare în L şi valoarea lui m care maximizează funcţia $\ln L(x_1, x_2, \ldots, x_n; m)$ maximizează şi $L(x_1, x_2, \ldots, x_n; m)$, este mai uşor de lucrat cu logaritmul lui L decât cu L. Avem

$$\ln L(x_1, x_2, \ldots, x_n; m) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2.$$

Dar $\dfrac{\partial \ln L(x_1, x_2, \ldots, x_n; m)}{\partial m} = \dfrac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - m),$

deci $\dfrac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - m) = 0 \Rightarrow \hat{m} = \overline{x}.$

Observăm că $\dfrac{\partial^2 \ln L(x_1, x_2, \ldots, x_n; m)}{\partial m^2} = -\dfrac{n}{\sigma^2} \leq 0$, $\forall m \in \mathbb{R}$, deci $\hat{m} = \overline{x}$ este punct de maxim. Astfel $t_n(x_1, x_2, \ldots, x_n) = \overline{x}$ este punct de verosimilitate maximă.

b) Dacă m este cunoscută şi $\sigma^2 \in \Theta = (0, \infty)$ este necunoscut folosim forma funcţiei de verosimilitate maximă dedusă în (11.1) şi obţinem

$$\begin{aligned} L(x_1, x_2, \ldots, x_n; m, \sigma^2) = \\ = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\overline{x} - m)^2 - \frac{n-1}{2\sigma^2}s^2\right\} = \\ = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\overline{x} - m)^2\right\} \exp\left\{-\frac{n-1}{2\sigma^2}s^2\right\}. \end{aligned}$$

Fixăm $m = \overline{x}$ şi obţinem

$$L(x_1, x_2, \ldots, x_n; \sigma^2) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{n-1}{2\sigma^2}s^2\right\} \Rightarrow$$

$$\ln L(x_1, x_2, \ldots, x_n; \sigma^2) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{n-1}{2\sigma^2}s^2.$$

$$\frac{\partial \ln L(x_1, x_2, \ldots, x_n; \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{n-1}{2\sigma^4}s^2,$$

$$\frac{\partial \ln L(x_1, x_2, \ldots, x_n; \sigma^2)}{\partial \sigma^2} = 0 \Rightarrow \hat{\sigma}^2 = \frac{n-1}{n}s^2.$$

Dar

$$\begin{aligned} \frac{d^2 \ln L}{d\left(\sigma^2\right)^2}(x_1, x_2, \ldots, x_n; \hat{\sigma}^2) = \frac{n}{2\sigma^4} - \frac{n-1}{\sigma^6}s^2\Big|_{\sigma^2 = \frac{n-1}{n}s^2} = \\ = -\frac{1}{2}\frac{n^3}{s^4(n-1)^2} \leq 0. \end{aligned}$$

Rezultă că $\hat{\sigma}^2$ este estimatorul de verosimilitate maximă.

Dacă avem de estimat mai mulţi parametri $1, 2, ..., p$, cu X cu densitatea de probabilitate $f(x, \theta_1, \theta_2, ..., \theta_p)$, atunci în mod analog cu cazul unui singur parametru, aplicăm principiul verosimilităţii maxime. Considerăm o selecţie $X_1, X_2, \ldots, X_n$ i.i.d cu v. a. X şi fie $x_1, x_2, \ldots, x_n$ valorile variabilelor de selecţie. Atunci *funcţia de verosimilitate maximă* corespunzătoare selecţiei este

$$L(x_1, x_2, \ldots, x_n; \theta_1, \theta_2, ..., \theta_p) = \prod_{j=1}^{n} f(x_j, \theta_1, \theta_2, ..., \theta_p).$$

Se aleg pentru parametri acele valori care maximizează funcţia de verosimilitate

$$L(x_1, x_2, \ldots, x_n; \theta_1, \theta_2, ..., \theta_p)$$

sau ceea ce este acelaşi lucru acele valori care maximizează $\ln L(x_1, x_2, \ldots, x_n; \theta_1, \theta_2, ..., \theta_p)$.

Aceasta implică:

$$\begin{cases} \dfrac{\partial \ln L(x_1, x_2, \ldots, x_n; \theta_1, \theta_2, ..., \theta_p)}{\partial \theta_1} = 0 \\ \dfrac{\partial \ln L(x_1, x_2, \ldots, x_n; \theta_1, \theta_2, ..., \theta_p)}{\partial \theta_2} = 0 \\ \cdots \\ \dfrac{\partial \ln L(x_1, x_2, \ldots, x_n; \theta_1, \theta_2, ..., \theta_p)}{\partial \theta_p} = 0 \end{cases} \tag{11.7}$$

şi diferenţiala de ordin doi, care este o formă pătratică, calculată în punctul $(\theta_1, \theta_2, ..., \theta_p)$, soluţie a sistemului (11.7), negativ definită.

Exemplul 11.4.5 Fie $x_1, x_2, \ldots, x_n$ $(n \geq 2)$ o selecţie de valori ale variabilelor de selecţie $X_1, X_2, \ldots, X_n$ care au o distribuţie normală, i. i. d. cu v. a. $X \in N(m, \sigma)$. Dacă m şi σ^2 sunt necunoscute, $\Theta = \{(m, \sigma^2) \,|m \in \mathbb{R}, \sigma^2 \in (0, \infty)\}$, atunci estimatorul de verosimilitate maximă $(\hat{m}, \hat{\sigma}^2)$ se obţine rezolvând sistemul

$$\begin{cases} \dfrac{\partial \ln L(x_1, x_2, \ldots, x_n; m, \sigma^2)}{\partial m} = \dfrac{1}{\sigma^2} \displaystyle\sum_{i=1}^{n} (x_i - m) = 0 \\ \dfrac{\partial \ln L(x_1, x_2, \ldots, x_n; m, \sigma^2)}{\partial \sigma^2} = 0 \end{cases}$$

adică

$$\begin{cases} \dfrac{1}{\sigma^2} \displaystyle\sum_{i=1}^{n} (x_i - m) = 0 \\ -\dfrac{n}{2\sigma^2} + \dfrac{1}{2(\sigma^2)^2} \displaystyle\sum_{i=1}^{n} (\overline{x} - m)^2 + \dfrac{n-1}{2(\sigma^2)^2} s^2 = 0 \end{cases}.$$

Obţinem

$$\begin{cases} \hat{m} = \overline{x} \\ \hat{\sigma}^2 = \dfrac{n-1}{n} s^2 \end{cases}$$

Calculăm

$$d^2 \ln L(x_1, x_2, \ldots, x_n; m, \sigma^2) = \frac{\partial^2 \ln L(x_1, \ldots, x_n; m, \sigma^2)}{\partial m^2} dm^2 +$$
$$+2 \frac{\partial^2 \ln L(x_1, \ldots, x_n; m, \sigma^2)}{\partial m \partial \sigma^2} dm d\sigma^2 + \frac{\partial^2 \ln L(x_1, \ldots, x_n; m, \sigma^2)}{(\partial \sigma^2)^2} \left(d\sigma^2\right)^2,$$

$$d^2 \ln L(x_1, x_2, \ldots, x_n; \hat{m}, \hat{\sigma}^2) = -\frac{1}{\sigma^2} dm^2 - \frac{n^3}{2(n-1)^2 (s^2)^2} \left(d\sigma^2\right)^2$$

care este o formă pătratică, evident, negativ definită. De aici rezultă că $\left(\overline{x}, \frac{n-1}{n}s^2\right)$ este unicul estimator de verosimilitate maximă pentru (m, σ^2). $\Diamond$

Exemplul 11.4.6 Presupunem că timpul de viaţă a unui aparat este distribuit $Exp\,[\theta]$ unde $\theta \in (0, \infty)$ este necunoscut. Bazându-ne pe o selecţie $(x_1, x_2, \ldots, x_n)$ de valori ale variabilelor $X_1, X_2, \ldots, X_n$ care au o distribuţie exponenţială obţinem funcţia de verosimilitate

$$L(x_1, x_2, \ldots, x_n; \theta) = \theta^n \exp\{-n\overline{x}\theta\} \Rightarrow$$
$$\ln L(x_1, x_2, \ldots, x_n; \theta) = n \ln \theta - n\overline{x}\theta \Rightarrow$$
$$\frac{\partial \ln L(x_1, x_2, \ldots, x_n; \theta)}{\partial \theta} = \frac{n}{\theta} - n\overline{x} \Rightarrow \frac{n}{\theta} - n\overline{x} = 0 \Rightarrow \hat{\theta} = \frac{1}{\overline{x}}.$$

$$\frac{\partial^2 \ln L}{\partial \theta^2}(x_1, x_2, \ldots, x_n; \theta) = -\frac{n}{\theta^2} \Rightarrow$$
$$\frac{\partial^2 \ln L}{\partial \theta^2}(x_1, x_2, \ldots, x_n; \frac{1}{\overline{x}}) = -n\left(\overline{x}\right)^2 \leq 0 \Rightarrow$$

$\hat{\theta} = \frac{1}{\overline{x}}$ este estimatorul de verosimilitate maximă a lui θ. $\Diamond$