Elemente de statistică

CURS 11 - 07.05.2014

Facultatea de Automatică și Calculatoare

Statistică și prelucrarea datelor

Statistică inferențială

- Statistica inferențială trage concluzii valabile pentru populație utilizând datele uneia sau mai multor selecții aleatoare și calcule probabilistice.

Statistică inferențială

- Statistica inferențială trage concluzii valabile pentru populație utilizând datele uneia sau mai multor selecții aleatoare și calcule probabilistice.
- Statistica inferențială se împarte în două mari domenii:

Statistică inferențială

- Statistica inferențială trage concluzii valabile pentru populație utilizând datele uneia sau mai multor selecții aleatoare și calcule probabilistice.
- Statistica inferențială se împarte în două mari domenii:
    1. estimarea parametrilor;

Statistică inferențială

- Statistica inferențială trage concluzii valabile pentru populație utilizând datele uneia sau mai multor selecții aleatoare și calcule probabilistice.
- Statistica inferențială se împarte în două mari domenii:
  1. estimarea parametrilor;
  2. ipoteze statistice.

Statistică inferențială

- Statistica inferențială trage concluzii valabile pentru populație utilizând datele uneia sau mai multor selecții aleatoare și calcule probabilistice.
- Statistica inferențială se împarte în două mari domenii:
  1. estimarea parametrilor;
  2. ipoteze statistice.

Fie X o v. a. cu densitatea $f(x, \theta), x \in \mathbb{R}, \theta \in \Theta \subset \mathbb{R}$.

- Mulțimea densităților de probabilitate $f(x, \theta)$ ce conțin parametrul necunoscut θ se numește **model probabilistic**.

Statistică inferențială

- Statistica inferențială trage concluzii valabile pentru populație utilizând datele uneia sau mai multor selecții aleatoare și calcule probabilistice.
- Statistica inferențială se împarte în două mari domenii:
  1. estimarea parametrilor;
  2. ipoteze statistice.

Fie X o v. a. cu densitatea $f(x,\theta), x \in \mathbb{R}, \theta \in \Theta \subset \mathbb{R}$.

- Mulțimea densităților de probabilitate $f(x,\theta)$ ce conțin parametrul necunoscut θ se numește **model probabilistic**.
- Exemple de familii de densități de probabilitate care pot fi alese ca modele probabilistice: normală, exponentială, Poisson etc. În momentul în care am ales o densitate de probabilitate de o anumită formă, incertitudinea legată de rezultatul particular al experimentului s-a transferat în incertitudinea legată de valorile parametrului (parametrilor).

Statistică inferențială

- **Repartiția selecției** $(X_1, X_2, ..., X_n)$ este definită ca repartiția comună a variabilelor de selecție $X_1, X_2, ..., X_n$. Dacă $x_1, x_2, ..., x_n$ sunt valorile luate de $X_1, X_2, ..., X_n$, atunci densitatea de probabilitate a selecției este notată prin

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

Statistică inferențială

- **Repartiția selecției** $(X_1, X_2, ..., X_n)$ este definită ca repartiția comună a variabilelor de selecție $X_1, X_2, ..., X_n$. Dacă $x_1, x_2, ..., x_n$ sunt valorile luate de $X_1, X_2, ..., X_n$, atunci densitatea de probabilitate a selecției este notată prin

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

Repartiția selecției depinde de θ și incorporează atât selecția cât și modelul probabilistic.

Statistică inferențială

- **Repartiția selecției** $(X_1, X_2, ..., X_n)$ este definită ca repartiția comună a variabilelor de selecție $X_1, X_2, ..., X_n$. Dacă $x_1, x_2, ..., x_n$ sunt valorile luate de $X_1, X_2, ..., X_n$, atunci densitatea de probabilitate a selecției este notată prin
$$L(x_1, x_2, ..., x_n; \theta), (x_1, x_2, ..., x_n) \in \mathbb{R}^n.$$
Repartiția selecției depinde de θ și incorporează atât selecția cât și modelul probabilistic.

Cea mai folosită formă de selecție este selecția aleatoare și este bazată pe ideea experimentului aleator.

- Vom spune că $X_1, X_2, ..., X_n$ este o **selecție aleatoare** asupra v. a. X care are densitatea de probabilitate $f(x, \theta)$ dacă $X_1, X_2, ..., X_n$ sunt independente și identic repartizate (i.i.r.) ca și X.
În cazul selecției aleatoare, densitatea de probabilitate comună a variabilelor de selecție este
$$L(x_1, x_2, ..., x_n; \theta) = \prod_{j=1}^{n} f(x_j, \theta).$$

Statistică inferențială

- O selecție aleatoare poate fi constituită prin repetarea unui experiment aleator de n ori. Realizarea selecției aleatoare (datele obținute în urma selecției) se notează cu $(x_1, x_2, ..., x_n)$ și mulțimea tuturor realizărilor definește **spațiul observațiilor**.

Statistică inferențială

- O selecție aleatoare poate fi constituită prin repetarea unui experiment aleator de n ori. Realizarea selecției aleatoare (datele obținute în urma selecției) se notează cu $(x_1, x_2, ..., x_n)$ și mulțimea tuturor realizărilor definește **spațiul observațiilor**.
  Modelul probabilist $f(x, \theta)$ împreună cu selecția $(X_1, X_2, ..., X_n)$ definesc **modelul statistic**.

Statistică inferențială

- O selecție aleatoare poate fi constituită prin repetarea unui experiment aleator de n ori. Realizarea selecției aleatoare (datele obținute în urma selecției) se notează cu $(x_1, x_2, ..., x_n)$ și mulțimea tuturor realizărilor definește **spațiul observațiilor**.
  Modelul probabilist $f(x, \theta)$ împreună cu selecția $(X_1, X_2, ..., X_n)$ definesc **modelul statistic**.
- Modelul statistic împreună cu datele observate $\longrightarrow$ întrebări:

Statistică inferențială

- O selecție aleatoare poate fi constituită prin repetarea unui experiment aleator de n ori. Realizarea selecției aleatoare (datele obținute în urma selecției) se notează cu $(x_1, x_2, ..., x_n)$ și mulțimea tuturor realizărilor definește **spațiul observațiilor**.
  Modelul probabilist $f(x, \theta)$ împreună cu selecția $(X_1, X_2, ..., X_n)$ definesc **modelul statistic**.
- Modelul statistic împreună cu datele observate $\longrightarrow$ întrebări:

1. datele observate sunt consistente cu modelul statistic postulat?

Statistică inferențială

- O selecție aleatoare poate fi constituită prin repetarea unui experiment aleator de n ori. Realizarea selecției aleatoare (datele obținute în urma selecției) se notează cu $(x_1, x_2, ..., x_n)$ și mulțimea tuturor realizărilor definește **spațiul observațiilor**.
  Modelul probabilist $f(x, \theta)$ împreună cu selecția $(X_1, X_2, ..., X_n)$ definesc **modelul statistic**.
- Modelul statistic împreună cu datele observate $\longrightarrow$ 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 necunoscuți $\theta \in \Theta$?

Statistică inferențială

- O selecție aleatoare poate fi constituită prin repetarea unui experiment aleator de n ori. Realizarea selecției aleatoare (datele obținute în urma selecției) se notează cu $(x_1, x_2, ..., x_n)$ și mulțimea tuturor realizărilor definește **spațiul observațiilor**.
  Modelul probabilist $f(x, \theta)$ împreună cu selecția $(X_1, X_2, ..., X_n)$ definesc **modelul statistic**.
- Modelul statistic împreună cu datele observate $\longrightarrow$ 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 necunoscuți $\theta \in \Theta$?
   a) putem descrește incertitudinea asupra lui θ prin deducerea spațiului parametrilor Θ la Θ_0 unde Θ_0 este o submulțime a lui Θ? (estimație prin intervale de încredere)

Statistică inferențială

- O selecție aleatoare poate fi constituită prin repetarea unui experiment aleator de n ori. Realizarea selecției aleatoare (datele obținute în urma selecției) se notează cu $(x_1, x_2, ..., x_n)$ și mulțimea tuturor realizărilor definește **spațiul observațiilor**.
  Modelul probabilist $f(x, \theta)$ împreună cu selecția $(X_1, X_2, ..., X_n)$ definesc **modelul statistic**.
- Modelul statistic împreună cu datele observate $\longrightarrow$ 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 necunoscuți $\theta \in \Theta$?
   a) putem descrește incertitudinea asupra lui θ prin deducerea spațiului parametrilor Θ la Θ_0 unde Θ_0 este o submulțime a lui Θ? (estimație prin intervale de încredere)
   b) putem descrește incertitudinea asupra lui θ prin alegerea unei valori particulare $\hat{\theta}$ din Θ ca având cea mai reprezentativă valoare a lui θ? (estimație punctuală)

Statistică inferențială

- O selecție aleatoare poate fi constituită prin repetarea unui experiment aleator de n ori. Realizarea selecției aleatoare (datele obținute în urma selecției) se notează cu $(x_1, x_2, ..., x_n)$ și mulțimea tuturor realizărilor definește **spațiul observațiilor**.
Modelul probabilist $f(x, \theta)$ împreună cu selecția $(X_1, X_2, ..., X_n)$ definesc **modelul statistic**.
- Modelul statistic împreună cu datele observate $\longrightarrow$ 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 necunoscuți $\theta \in \Theta$?
a) putem descrește incertitudinea asupra lui θ prin deducerea spațiului parametrilor Θ la Θ_0 unde Θ_0 este o submulțime a lui Θ? (estimație prin intervale de încredere)
b) putem descrește incertitudinea asupra lui θ prin alegerea unei valori particulare $\hat{\theta}$ din Θ ca având cea mai reprezentativă valoare a lui θ? (estimație punctuală)
c) putem considera întrebarea θ aparține unei submulțimi Θ_0 a lui Θ? (verificarea ipotezelor statistice)

Statistică inferențială

- Dată o populație de volum N se numește **statistică** a acelei populații orice mărime semnificativă calculată dintr-un eșantion aleator al acelei populații.

Statistică inferențială

- Dată o populație de volum N se numește **statistică** a acelei populații orice mărime semnificativă calculată dintr-un eșantion aleator al acelei populații. Pentru un parametru al repartiției, se numește **estimator** al acelui parametru orice statistică aproximând acel parametru.

Statistică inferențială

- Dată o populație de volum N se numește **statistică** a acelei populații orice mărime semnificativă calculată dintr-un eșantion aleator al acelei populații. Pentru un parametru al repartiției, se numește **estimator** al acelui parametru orice statistică aproximând acel parametru.
  Dacă parametrul are o valoare bine determinată (în cadrul analizei statistice), orice estimator al lui se numește **estimator punctual**.

Statistică inferențială

- Dată o populație de volum N se numește **statistică** a acelei populații orice mărime semnificativă calculată dintr-un eșantion aleator al acelei populații. Pentru un parametru al repartiției, se numește **estimator** al acelui parametru orice statistică aproximând acel parametru.
  Dacă parametrul are o valoare bine determinată (în cadrul analizei statistice), orice estimator al lui se numește **estimator punctual**.

Example 1

Considerăm populația diametrelor unor bile de rulmenți dintr-un lot. Ca statistică a acestei populații se poate lua media de selecție $\overline{x}$ a unui șir $(x_1, x_2, ..., x_n)$ format de diametrele unui eșantion de bile din acel lot. Alte statistici sunt: s^2, dispersia de selecție modificată, s, abaterea medie de selecție.

Statistică inferențială

Example 2

O societate de telefoane poate considera populația tuturor abonaților. Un eșantion se poate obține luând pe sărite din lista abonaților, de exemplu, din 25 în 25. Presupunem că societatea este interesată în estimarea unui parametru caracteristic; de exemplu, gradul de satisfacție g al abonaților relativ la serviciile oferite de socitate. Se cere o apreciere cu note de la 1 la 5. Este solicitat răspunsul abonaților din eșantionul ales și se obține un estimator al lui g.

Statistică inferențială

Example 2

O societate de telefoane poate considera populația tuturor abonaților. Un eșantion se poate obține luând pe sărite din lista abonaților, de exemplu, din 25 în 25. Presupunem că societatea este interesată în estimarea unui parametru caracteristic; de exemplu, gradul de satisfacție g al abonaților relativ la serviciile oferite de socitate. Se cere o apreciere cu note de la 1 la 5. Este solicitat răspunsul abonaților din eșantionul ales și se obține un estimator al lui g.

- De obicei se notează cu θ parametrul și cu $\hat{\theta}$ un estimator al său. De exemplu, θ reprezintă media atunci putem considera ca statistică a mediei,
$$\theta = \overline{X} = \frac{X_1 + X_2 + ... + X_n}{n}.$$
Dacă $x_1, x_2, ..., x_n$ sunt valorile luate de $X_1, X_2, ..., X_n$ atunci estimatorul lui θ este
$$\hat{\theta} = \overline{x} = \frac{1}{n}\left(x_1 + x_2 + ... + x_n\right).$$
Dacă măsurătorile sunt realizate cu acuratețe și precizie rezonabile, atunci $X_1, X_2, ..., X_n$ se pot considera repartizate la fel.

Exemple de modele statistice ce apar adesea în practică

Modelul Bernoulli
Presupunem că $X_1, X_2, ..., X_n$ este o selecție aleatoare a unei populații care urmează o distribuție Bernoulli cu parametrul $\theta \in [0,1]$ necunoscut.
Știm că distribuția 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 și $X_i = 0$ 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 știm valoarea lui θ.
Spațiul parametrului θ este $[0,1]$. Dacă $x_1, x_2, ..., x_n$ sunt valorile luate de $X_1, X_2, ..., X_n$ atunci densitatea de probabiltate pentru selecția de ordin i este

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

iar densitatea de probabilitate a eșantionului este dată de

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

Exemple de modele statistice ce apar adesea în practică

Modelul normal

Presupunem că $X_1, X_2, ..., X_n$ este o selecție aleatoare a unei populații care urmează o distribuție normală $N(m, \sigma)$ cu $\theta = (m, \sigma) \in \mathbb{R} \times (0, \infty)$ necunoscuți.

De exemplu putem avea observații asupra înălțimii în cm populației și intuim că este rezonabil să presupunem că distribuția înălțimii este normală cu media și dispersia necunoscute. Dacă $x_1, x_2, ..., x_n$ sunt valorile luate de $X_1, X_2, ..., X_n$ atunci densitatea de probabiltate a unui eșantion este dată de

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

Deoarece $\sum_{i=1}^{n}(x_i - m)^2 = n(\overline{x} - m)^2 + \sum_{i=1}^{n}(x_i - \overline{x})^2$, unde $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ și $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$, obținem

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

Exemple de modele statistice ce apar adesea în practică

Modelul normal

Presupunem că $X_1, X_2, ..., X_n$ este o selecție aleatoare a unei populații care urmează o distribuție normală $N(m, \sigma)$ cu $\theta = (m, \sigma) \in \mathbb{R} \times (0, \infty)$ necunoscuți.

De exemplu putem avea observații asupra înălțimii în cm populației și intuim că este rezonabil să presupunem că distribuția înălțimii este normală cu media și dispersia necunoscute. Dacă $x_1, x_2, ..., x_n$ sunt valorile luate de $X_1, X_2, ..., X_n$ atunci densitatea de probabiltate a unui eșantion este dată de

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

Deoarece $\sum_{i=1}^{n}(x_i - m)^2 = n(\overline{x} - m)^2 + \sum_{i=1}^{n}(x_i - \overline{x})^2$, unde $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ și

$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$, obținem

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

Estimatori punctuali

Construcția histogramei și încercarea de a intui modelul statistic sunt metode folosite de statistician în încercarea de a studia distribuția populației.

- ce model statistic trebuie folosit?

Estimatori punctuali

Construcția histogramei și încercarea de a intui modelul statistic sunt metode folosite de statistician în încercarea de a studia distribuția populației.

- ce model statistic trebuie folosit?
- valorile parametrilor care intra în definiția modelului statistic (de exemplu, știm că înălțimea urmează o distribuție normală, dar nu știm media și dispersia)

Estimatori punctuali

Construcția histogramei și încercarea de a intui modelul statistic sunt metode folosite de statistician în încercarea de a studia distribuția populației.

- ce model statistic trebuie folosit?
- valorile parametrilor care intra în definiția modelului statistic (de exemplu, știm că înălțimea urmează o distribuție normală, dar nu știm media și dispersia)

Pentru a justifica această alegere avem nevoie să studiem **estimatorii punctuali.**

Estimatori punctuali

- Să considerăm o variabilă aleatoare X a cărei lege de probabilitate conține un parametru θ. Fie $X_1, ..., X_n$ - n variabile aleatoare independente care au aceeași distribuție ca și X.
  Alegem o anumită funcție $\hat{\theta}(X_1, ..., X_n)$ pe care o vom utiliza ca estimator pentru θ. Cu alte cuvinte, dacă dispunem de valorile $x_1, x_2, ..., x_n$ obținute experimental, numărul $\hat{\theta}\left(x_1, x_2, ..., x_n\right)$ va fi considerat ca estimator punctual al parametrului θ.

Estimatori punctuali

- Să considerăm o variabilă aleatoare X a cărei lege de probabilitate conține un parametru θ. Fie $X_1, ..., X_n$ - n variabile aleatoare independente care au aceeași distribuție ca și X.
  Alegem o anumită funcție $\hat{\theta}(X_1, ..., X_n)$ pe care o vom utiliza ca estimator pentru θ. Cu alte cuvinte, dacă dispunem de valorile $x_1, x_2, ..., x_n$ obținute experimental, numărul $\hat{\theta}\left(x_1, x_2, ..., x_n\right)$ va fi considerat ca estimator punctual al parametrului θ.

Definition 3

Se consideră o populație de volum N și un parametru θ al acestei populații. Fie $x_1, x_2, ..., x_n$ un eșantion reprezentativ $(n << N)$ al populației; un estimator punctual $\hat{\theta} = \hat{\theta}\left(x_1, x_2, ..., x_n\right)$ al lui θ se numește **estimator nedeplasat** sau **absolut corect** dacă $M\left[\hat{\theta}(X_1, ..., X_n)\right] = \theta$.

Estimatori punctuali

- Să considerăm o variabilă aleatoare X a cărei lege de probabilitate conține un parametru θ. Fie $X_1, ..., X_n$ - n variabile aleatoare independente care au aceeași distribuție ca și X.
  Alegem o anumită funcție $\hat{\theta}(X_1, ..., X_n)$ pe care o vom utiliza ca estimator pentru θ. Cu alte cuvinte, dacă dispunem de valorile $x_1, x_2, ..., x_n$ obținute experimental, numărul $\hat{\theta}\,(x_1, x_2, ..., x_n)$ va fi considerat ca estimator punctual al parametrului θ.

Definition 3

Se consideră o populație de volum N și un parametru θ al acestei populații. Fie $x_1, x_2, ..., x_n$ un eșantion reprezentativ ($n << N$) al populației; un estimator punctual $\hat{\theta} = \hat{\theta}\,(x_1, x_2, ..., x_n)$ al lui θ se numește **estimator nedeplasat** sau **absolut corect** dacă $M\left[\hat{\theta}(X_1, ..., X_n)\right] = \theta$.

Dacă $\lim\limits_{n\to\infty} M\left[\hat{\theta}(X_1, ..., X_n)\right] = \theta$ și $\lim\limits_{n\to\infty} D\left[\hat{\theta}(X_1, ..., X_n)\right] = 0$ atunci $\hat{\theta}$ este un **estimator consistent** sau **corect** sau **deplasat** al lui θ.

Estimatori punctuali

Theorem 4

Fie $x_1, x_2, \dots, x_n$ $(n \geq 2)$ *o selecție de valori ale variabilelor* $X_1, X_2, ..., X_n$ *(v.a. independente la fel distribuite ca și* X*),* m *media teoretică și* σ^2 *dispersia teoretică. Notăm cu*
$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$*, media de selecție,* $\overline{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ *dispersia de selecție și*
$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ *dispersia de selecție modificată.*
Atunci media de selecție este un estimator punctual pentru media teoretică, iar dispersia de selecție și dispersia de selecție modificată sunt estimatori punctuali pentru dispersia teoretică. În plus,
$M\left[\overline{X}\right] = m, D\left[\overline{X}\right] = \frac{1}{n}\sigma^2 \Rightarrow$ *media de selecție este un estimator absolut corect al lui m.*
$M\left[\overline{\sigma^2}\right] = \frac{n-1}{n}\sigma^2 \Rightarrow$ *dispersia de selecție este un estimator corect al lui* σ^2*.*
$M\left[S^2\right] = \sigma^2 \Rightarrow$ *dispersia de selecție este un estimator absolut corect al lui* σ^2*.*

Estimatori punctuali

Dacă $X \in \mathrm{Pois}[\lambda]$, atunci $M[X] = D[X] = \lambda$ și atunci $\overline{X}$ și s^2 sunt estimatorii nedeplasați pentru λ. Se pune întrebarea pe care îl vom alege. Cum dispersia este o măsură a împrăștierii, intuiția sugerează să-l alegem pe acela care are cea mai mică dispersie și aceasta deoarece el are o repartiție mai concentrată în jurul lui λ.

Estimatori punctuali

Dacă $X \in \mathrm{Pois}[\lambda]$, atunci $M[X] = D[X] = \lambda$ și atunci $\overline{X}$ și s^2 sunt estimatorii nedeplasați pentru λ. Se pune întrebarea pe care îl vom alege. Cum dispersia este o măsură a împrăștierii, intuiția sugerează să-l alegem pe acela care are cea mai mică dispersie și aceasta deoarece el are o repartiție mai concentrată în jurul lui λ.

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

Theorem 5

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

Estimatori punctuali

Dacă $X \in \mathrm{Pois}[\lambda]$, atunci $M[X] = D[X] = \lambda$ și atunci $\overline{X}$ și s^2 sunt estimatorii nedeplasați pentru λ. Se pune întrebarea pe care îl vom alege. Cum dispersia este o măsură a împrăștierii, intuiția sugerează să-l alegem pe acela care are cea mai mică dispersie și aceasta deoarece el are o repartiție mai concentrată în jurul lui λ.
Dintre doi estimatori nedeplasați $\hat{\theta}$ și $\hat{\theta}_1$ pentru același parametru θ, se consideră că este mai precis cel care are dispersie mai mică.

Theorem 5

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

Demonstrație. Conform teoremei limită centrală avem că
$S_n = X_1 + X_2 + \ldots + X_n \Rightarrow S_n \in N\left(nm, \sqrt{n}\sigma\right)$.
Avem $M\left[\frac{1}{n}S_n\right] = \frac{1}{n}nm = m$, $D\left[\frac{1}{n}S_n\right] = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n} \Rightarrow \frac{1}{n}S_n \in N\left(m, \frac{\sigma}{\sqrt{n}}\right)$

□

Estimatori punctuali

Corollary 6

În condițiile teoremei 5 pentru orice $a < b$ avem

$$P\left(a \leq \overline{X} < b\right) \approx \Phi\left(\frac{b-m}{\frac{\sigma}{\sqrt{n}}}\right) - \Phi\left(\frac{a-m}{\frac{\sigma}{\sqrt{n}}}\right). \qquad (2)$$

Statisticienii recomandă folosirea formulei (2) pentru populații statistice de volum N și eșantioane de volum n unde $n \geq 30$ și $n \leq \frac{N}{3}$. Dacă $n < 30$ formula este utilă dacă populația inițială nu este departe de a fi normal distribuită. Pentru eșantioane mici statisticienii propun o corecție care să înlocuiască $\frac{\sigma}{\sqrt{n}}$ prin $\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$.

Estimatori punctuali

Problem 7

Considerăm greutățile populației de 350 studenți dintr-o facultate s-a constatat că media este $m = 70$ *și abaterea media pătratică este* $\sigma = 10$.
a) Să se determine probabilitatea ca un student luat la întâmplare să cântărească între 65 și 70 kg?
b)Să se determine probabilitatea ca media maselor să fie între 65 și 75 kg.
c) Să se determine probabilitatea ca extrăgând un eșantion de 36 studenți din acea populație, media eșantionului să fie cuprinsă între 65 și 75 kg.
d) Extrgînd un eșantion de 100 de studenți care este probabilitatea ca media maselor să fie sub 65 kg?

Estimatori punctuali

Soluție.
a) Pobabilitatea cerută la punctul a) nu poate fi calculată deoarece $X =$ greutatea populației nu este repartizată normal.

Estimatori punctuali

Soluție.
a) Pobabilitatea cerută la punctul a) nu poate fi calculată deoarece $X =$ greutatea populației nu este repartizată normal.
La fel și la b).

Estimatori punctuali

Soluție.
a) Pobabilitatea cerută la punctul a) nu poate fi calculată deoarece $X =$ greutatea populației nu este repartizată normal.
La fel și la b).
c) $n = 36, m = 70, \sigma_{\overline{X}} = \frac{10}{\sqrt{36}} = 1.6667$,
$P\left(65 \leq \overline{X} \leq 75\right) = \Phi\left(\frac{75-70}{1.6667}\right) - \Phi\left(\frac{65-70}{1.6667}\right) = 0.9986 - 0.0014 = .9972.$

Estimatori punctuali

Soluție.
a) Pobabilitatea cerută la punctul a) nu poate fi calculată deoarece $X =$ greutatea populației nu este repartizată normal.
La fel și la b).
c) $n = 36, m = 70, \sigma_{\overline{X}} = \frac{10}{\sqrt{36}} = 1.6667$,
$P\left(65 \leq \overline{X} \leq 75\right) = \Phi\left(\frac{75-70}{1.6667}\right) - \Phi\left(\frac{65-70}{1.6667}\right) = 0.9986 - 0.0014 = .9972.$
d) $n = 100, m = 70, \sigma_{\overline{X}} = \frac{10}{\sqrt{100}} = 1$,
$P\left(\overline{X} \leq 65\right) = \Phi\left(\frac{65-70}{1}\right) = 0.0000003.$▼

Metoda momentelor

Ideea metodei este de a egala momentele distribuției cu momentele statistice.și de.aici obținem estimarea parametrilor.
Știm că momentul inițial de ordin k este

$$M\left[X^k\right] = \sum_{i=1}^{n} x_i^k p_i$$

dacă distribuția este discretă și

$$M\left[X^k\right] = \int_{-\infty}^{\infty} x^k f(x)dx$$

dacă distribuția este continuă, iar momentul statistic inițial de ordin k este $\overline{X}^k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$.
De exemplu momentul inițial de ordin întâi este media iar media statistică este momentul statistic de ordin întâi. Egalăm acestea și obținem estimarea punctuală $\hat{\mu} = \overline{X}$.

Metoda momentelor

- Fie $x_1, x_2, \ldots, x_n$ $(n \geq 2)$ o selecție de valori ale variabilelor $X_1, X_2, ..., X_n$ care au o distribuție exponențială cu parametru λ. Avem un singur parametru de estimat. Știm 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 obținut cu metoda momentelor.

Metoda momentelor

- Fie $x_1, x_2, \ldots, x_n$ $(n \geq 2)$ o selecție de valori ale variabilelor $X_1, X_2, ..., X_n$ care au o distribuție exponențială cu parametru λ. Avem un singur parametru de estimat. Știm 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 obținut cu metoda momentelor.
  Ca exemplu, presupunem că este testat timpul de viață al unui modul electronic folosi în industria automobilelor. Timpul de viață urmează o distribuție exponențială. S-au făcut teste și s-au obținut 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 obținut cu metoda momentelor este $\hat{\lambda} = \frac{1}{21.65} = 0.046189$.

Metoda momentelor

- Fie $x_1, x_2, \ldots, x_n$ $(n \geq 2)$ o selecție de valori ale variabilelor $X_1, X_2, ..., X_n$ care au o distribuție exponențială cu parametru λ. Avem un singur parametru de estimat. Știm 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 obținut cu metoda momentelor.
  Ca exemplu, presupunem că este testat timpul de viață al unui modul electronic folosi în industria automobilelor. Timpul de viață urmează o distribuție exponențială. S-au făcut teste și s-au obținut 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 obținut cu metoda momentelor este $\hat{\lambda} = \frac{1}{21.65} = 0.04618\,9$.
- Fie $x_1, x_2, \ldots, x_n$ $(n \geq 2)$ o selecție de valori ale variabilelor $X_1, X_2, ..., X_n$ care au o distribuție normală de parametrii m și σ. În acest caz avem $M[X] = m$, $M\left[X^2\right] = m^2 + \sigma^2$.
  Rezultă că $m = \frac{1}{n}\sum\limits_{i=1} x_i$, $m^2 + \sigma^2 = \frac{1}{n}\sum\limits_{i=1} x_i^2 \Rightarrow \hat{m} = \frac{1}{n}\sum\limits_{i=1} x_i$,
  $\hat{\sigma}^2 = \frac{1}{n}\sum\limits_{i=1} x_i^2 - \left(\frac{1}{n}\sum\limits_{i=1} x_i\right)^2 \Rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum\limits_{i=1} (x_i - \overline{x})^2$. Reamintim că acest estimator pentru σ^2 este un estimator deplasat.

Metoda verosimilității maxime

- dezvoltată de Fisher și extinsă de Cramer, Rao și Wald

Metoda verosimilității maxime

- dezvoltată de Fisher și extinsă de Cramer, Rao și Wald
- cea mai folosită metodă de estimare și joacă un rol important în verificarea ipotezelor statistice

Metoda verosimilității maxime

- dezvoltată de Fisher și extinsă de Cramer, Rao și Wald
- cea mai folosită metodă de estimare și joacă un rol important în verificarea ipotezelor statistice
- Considerăm o selecție $X_1, X_2, ..., X_n$ i.i.d. cu v. a. X, ce are cu densitatea de probabilitate $f(x, \theta)$. Atunci funcția de verosimilitate maximă corespunzătoare selecției este

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

Metoda verosimilității maxime

- dezvoltată de Fisher și extinsă de Cramer, Rao și Wald
- cea mai folosită metodă de estimare și joacă un rol important în verificarea ipotezelor statistice
- Considerăm o selecție $X_1, X_2, ..., X_n$ i.i.d. cu v. a. X, ce are cu densitatea de probabilitate $f(x, \theta)$. Atunci funcția de verosimilitate maximă corespunzătoare selecției este

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

- Acceptăm următoarea **axiomă**: valoarea cea mai verosimilă a parametrului este cea care maximizează funcția $L(x_1, x_2, ..., x_n; \theta)$, ceea ce conduce la ecuația

$$\frac{\partial L(x_1, x_2, ..., x_n; \theta)}{\partial \theta} = 0. \tag{3}$$

Metoda verosimilității maxime

- Cum uneori este dficil de rezolvat ecuația (3), introducem funcțiile $\ln L(x_1, x_2, ..., x_n; \theta)$, **logaritmul funcției de verosimilitate** și

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

funcția scor.

Metoda verosimilității maxime

- Cum uneori este dficil de rezolvat ecuația (3), introducem funcțiile $\ln L(x_1, x_2, ..., x_n; \theta)$, **logaritmul funcției de verosimilitate** și

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

funcția scor.

Definition 8

$t_n\left(x_1, x_2, ..., x_n\right) : S \to \Theta$ se numește **estimație de verosimilitate maximă** pentru parametrul θ dacă este punct de maxim pentru funcția de verosimilitate, adică

$$\ln L(x_1, x_2, ..., x_n; t_n) \geq \ln L(x_1, x_2, ..., x_n; \theta), \quad \forall \theta \in \Theta.$$

Metoda verosimilității maxime

În cele ce urmează presupunem că $f(x_i, \theta)$ sunt derivabile până la ordinul doi inclusiv în raport cu θ. Atunci există $\dfrac{\partial L(x_1, x_2, ..., x_n; \theta)}{\partial \theta}$ și estimația de verosimilitate maximă este soluție a ecuației

$$\frac{\partial L(x_1, x_2, ..., x_n; \theta)}{\partial \theta} = 0 \tag{4}$$

Metoda verosimilității maxime

În cele ce urmează presupunem că $f(x_i, \theta)$ sunt derivabile până la ordinul doi inclusiv în raport cu θ. Atunci există $\dfrac{\partial L(x_1, x_2, ..., x_n; \theta)}{\partial \theta}$ și estimația de verosimilitate maximă este soluție a ecuației

$$\frac{\partial L(x_1, x_2, ..., x_n; \theta)}{\partial \theta} = 0 \qquad (4)$$

sau a ecuației

$$\frac{\partial \ln L(x_1, x_2, ..., x_n; \theta)}{\partial \theta} = 0 \qquad (5)$$

atunci când aceasta există.
Ecuațiile (4) și (5) se numesc **ecuații de verosimilitate.**

Metoda verosimilității maxime: media necunoscută, dispersia cunoscută

Fie $x_1, x_2, \ldots, x_n$ $(n \geq 2)$ o selecție de valori ale variabilelor $X_1, X_2, ..., X_n$ care au o distribuție normală, i. i. r. cu v. a. $X \in N\left(m, \sigma^2\right)$.

Metoda verosimilității maxime: media necunoscută, dispersia cunoscută

Fie $x_1, x_2, \ldots, x_n$ $(n \geq 2)$ o selecție de valori ale variabilelor $X_1, X_2, ..., X_n$ care au o distribuție normală, i. i. r. cu v. a. $X \in N\left(m, \sigma^2\right)$.

a) Dacă σ^2 este cunoscută și media m este necunoscută cu $\Theta = \mathbb{R}$, atunci funcția de verosimilitate maximă este

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

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

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

Metoda verosimilității maxime: media necunoscută, dispersia cunoscută

Dar

$$\frac{\partial \ln L(x_1, x_2, ..., x_n; m)}{\partial m} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - m), \ \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - m) = 0 \Rightarrow \hat{m} = \overline{x}.$$

Observăm că

$$\frac{\partial^2 \ln L(x_1, x_2, ..., x_n; m)}{\partial m^2} = -\frac{n}{\sigma^2} \leq 0, \forall m \in \mathbb{R},$$

deci $\hat{m} = \overline{x}$ este punct de maxim. Astfel $t_n(x_1, x_2, ..., x_n) = \overline{x}$ este punct de verosimilitate maximă.

Metoda verosimilității maxime: media cunoscută, dispersia necunoscută

Dacă m este cunoscută și σ^2, $\Theta = (0, \infty)$ este necunoscut folosim forma funcției de verosimilitate maximă dedusă în (1) și obținem

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

$$= \quad \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\overline{x} - m)^2\right\} \exp\left\{-\frac{n-1}{2\sigma^2}s^2\right\}.$$

Fixăm $m = \overline{x}$ și atunci

$$\begin{aligned} L(x_1, x_2, ..., x_n; \sigma^2) &= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{n-1}{2\sigma^2}s^2\right\} \Rightarrow \\ \ln L(x_1, x_2, ..., x_n; \sigma^2) &= -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{n-1}{2\sigma^2}s^2. \end{aligned}$$

Metoda verosimilității maxime: media cunoscută, dispersia necunoscută

$$\begin{aligned}
\frac{\partial \ln L(x_1, x_2, ..., x_n; \sigma^2)}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{n-1}{2\sigma^4} s^2, \frac{\partial \ln L(x_1, x_2, ..., x_n; \sigma^2)}{\partial \sigma^2} = 0 \Rightarrow \\
\hat{\sigma}^2 &= \frac{n-1}{n} s^2.
\end{aligned}$$

Dar

$$\frac{\partial^2 \ln L}{\partial \left(\sigma^2\right)^2}(x_1, x_2, \ldots, x_n; \hat{\sigma}^2) = \frac{n}{2\sigma^4} - \frac{n-1}{\sigma^6} s^2 \bigg|_{\sigma^2 = \frac{n-1}{n} s^2} = -\frac{1}{2} \frac{n^3}{s^4 (n-1)^2} \leq 0 \Rightarrow$$

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

Metoda verosimilității maxime: media și dispersia necunoscute

c) Dacă m și σ^2 sunt necunoscute, $\Theta = \left\{ \left(m, \sigma^2\right) \mid m \in \mathbb{R}, \sigma^2 \in (0, \infty) \right\}$, atunci valoarea maximă se găsește rezolvând sistemul

$$\begin{cases} \dfrac{\partial \ln L}{\partial m}(x_1, \ldots, x_n; m, \sigma^2) = 0 \\ \dfrac{\partial \ln L}{\partial \sigma^2}(x_1, \ldots, x_n; m, \sigma^2) = 0, \end{cases}$$

de unde

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

Obținem

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

Metoda verosimilității maxime: media și dispersia necunoscute

Diferențiala de ordinul doi este

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

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

Metoda verosimilității maxime

Example 9 (Modelul exponențial)

Presupunem că timpul de viață a unui aparat este distribuit $\mathrm{Exp}\,[\theta]$ unde $\theta \in (0, \infty)$ este necunoscut. Bazându-ne pe o selecție $(x_1, x_2, ..., x_n)$ de valori ale variabilelor $X_1, X_2, ..., X_n$ care au o distribuție exponențială obținem funcția de verosimilitate

$L(x_1, x_2, ..., x_n; \theta) = \theta^n \exp\{-n\overline{x}\theta\} \Rightarrow$

$\ln L(x_1, x_2, ..., x_n; \theta) = n \ln \theta - n\overline{x}\theta \Rightarrow \dfrac{\partial \ln L(x_1, x_2, ..., x_n; \theta)}{\partial \theta} = \dfrac{n}{\theta} - n\overline{x} \Rightarrow$

$\dfrac{n}{\theta} - n\overline{x} = 0 \Rightarrow \hat{\theta} = \frac{1}{\overline{x}}$.

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

$\hat{\theta}$ este estimatorul de verosimilitate maximă.▼

Metoda verosimilității maxime

Example 10

Fie X o variabila aleatoare cu densitatea de probabilitate

$$f(x) = \begin{cases} \frac{1}{\theta} x^{\frac{1-\theta}{\theta}}, 0 < x \leq 1, \theta > 0 \\ 0, \text{ în rest} \end{cases}$$

Se face o selecție $x_1, x_2, ..., x_n$. Să se determine estimatorul de verosimilitate maximă pentru parametrul θ al repartitiei.

Metoda verosimilității maxime

Example 11

Fie X o variabila aleatoare cu densitatea de probabilitate

$$f(x) = \begin{cases} (1+\lambda)x^{\lambda}, 0 \leq x \leq 1, \lambda > 0 \\ 0, \text{ în rest.} \end{cases}$$

Se face o selecție $x_1, x_2, ..., x_n$. Să se determine estimatorul de verosimilitate maximă pentru parametrul λ al repartitiei.