Metode numerice și statistică
Curs 11

lect. Ciprian Deliu
✉ cdeliu@tuiasi.ro
moodle.deliu.ro

Universitatea Tehnică "Gh. Asachi" Iași
Facultatea de Hidrotehnică, Geodezie și Ingineria Mediului

2019

Teoria selecției

- Cercetarea statistică a unei caracteristici pentru o populație (în general de volum mare) se face prin sondaje asupra unei părți finite, aleasă aleator a populației. Aceste părți, care se presupun a fi omogene din punct de vedere al caracteristicii studiate, se numesc **eșantioane**. Numărul de elemente dintr-un eșantion constituie *volumul eșantionului*.
- Procedeul de a obține un eșantion dintr-o populație se numește **selecție**. Dacă fiecare element al populației are șansă egală de a aparține unui eșantion, atunci avem o **selecție aleatoare simplă**.
- Selecțiile pot fi *cu repetiție* dacă elementul ales este reintrodus în populație înainte de extragerea următorului element (alegerile succesive sunt independente și echiprobabile) și *fără repetiție* în caz contrar.
- În cazul în care volumul populației N este foarte mare în raport cu volumul eșantionului n, nu se face nicio diferență între selecția cu repetiție și cea fără repetiție. Selecția fără repetiție prezintă interes numai atunci când volumul populației este mic.

- Procedeul prin care se obțin informații privind întreaga populație folosind rezultatele din studiul eșantioanelor se numește **inferență statistică**.
- Cercetarea statistică a unei caracteristici se poate face atât prin **estimarea parametrilor** (o caracteristică numerică, un parametru din funcția sau densitatea de repartiție), cât și prin **verificarea ipotezelor statistice**.
- Există două tipuri de estimări utilizate mai des în practică: **estimări punctuale** și **estimări prin intervale de încredere**.
- În cadrul estimării punctuale a unui parametru se folosește un procedeu de determinare, pe baza datelor unei selecții, a unui număr care aproximează valoarea reală a parametrului.
- În cadrul celuilalt tip de estimare a unui parametru λ, i se poate asocia un interval $(\underline{\lambda}, \bar{\lambda})$, numit *interval de încredere*, cu proprietatea că orice element din acesta reprezintă, cu o anumită probabilitate, o valoare aproximativă a parametrului:

$$P(\underline{\lambda} < \lambda < \bar{\lambda}) = \alpha,\ \alpha \in (0,1).$$

Intervalul $(\underline{\lambda}, \bar{\lambda})$ se numește $100\alpha\%$ interval de încredere, α este *nivel de încredere*, $1-\alpha$ este *pragul de încredere* (sau *nivel de semnificație*), iar $\underline{\lambda}, \bar{\lambda}$ sunt *limite de încredere* pentru parametrul λ.

Definiție

Fie X o variabilă aleatoare teoretică asociată unei caracteristici a unei populații Ω și $X_1, X_2, \ldots, X_n$ variabile aleatoare independente de selecție, asociate selecțiilor de volum n ale lui Ω.
O variabilă aleatoare $S : \Omega \to \mathbb{R}$ pentru care există o funcție $H : \mathbb{R}^n \to \mathbb{R}$ astfel încât

$$S(\omega_1, \omega_2, \ldots, \omega_n) = H\left(X_1(\omega_1), X_2(\omega_2), \ldots, X_n(\omega_n)\right), \forall \omega_i \in \Omega,\ i = 1, \ldots, n$$

se numește **funcție de selecție** *sau* **statistică**.

- Pentru comoditate vom nota

$$S = H(X_1, X_2, \ldots, X_n) = H(X, n).$$

- Informațiile obținute asupra variabilelor aleatoare de selecție $X_i,\ i = 1, \ldots, n$ ne vor permite estimarea unor parametri asociați unei distribuții.

Definiție

Fie $X : \Omega \to \mathbb{R}$ o variabilă aleatoare a cărei repartiție depinde de un parametru real θ, asociată unei caracteristici numerice a unei populații statistice Ω, $\{\omega_1, \omega_2, \ldots, \omega_n\} \subset \Omega$ un eșantion de volum n și $\{x_1, x_2, \ldots, x_n\}$ valorile lui X pe eșantionul respectiv, adică $X(\omega_i) = x_i,\ i = 1, \ldots, n$.

- *Se numește* **estimator** *pentru parametrul θ orice funcție de selecție $H(X, n)$ care aproximează acest parametru;*
- *Valoarea $H(x_1, x_2, \ldots, x_n) = \hat{\theta}$ se numește* **estimație punctuală** *a parametrului θ.*

Exemplu:
Fie o variabilă aleatoare X repartizată $N(\mu, \sigma)$. Pe un eșantion de volum 4 se obțin valorile $x_1 = 25, x_2 = 30, x_3 = 29, x_4 = 31$. Media de selecție a acestor valori $\bar{x} = \frac{25+30+29+31}{4} = 28.75$ este o estimație punctuală a parametrului μ.

Definiție

Estimatorul $H(X, n)$ se numește:

- **consistent** *pentru parametrul θ dacă*

$$\forall \varepsilon > 0,\ \lim_{n\to\infty} P\left(|H(X,n) - \theta| < \varepsilon\right) = 1$$

- **corect** *pentru parametrul θ dacă*

$$\lim_{n\to\infty} E(H(X,n)) = \theta, \lim_{n\to\infty} Var(H(X,n)) = 0$$

- **absolut corect** *pentru parametrul θ dacă*

$$E(H(X,n)) = \theta, \lim_{n\to\infty} Var(H(X,n)) = 0$$

- **nedeplasat** *pentru parametrul θ dacă*

$$E(H(X,n)) = \theta$$

Estimatori pentru medie și dispersie

Estimatorul

$$m(X,n) = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

se numește **medie de selecție** și este un estimator absolut corect și nedeplasat pentru $E(X)$:

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} nE(X) = E(X)$$

$$Var(\bar{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2} nVar(X)$$

$$\lim_{n\to\infty} Var(\bar{X}) = \lim_{n\to\infty} \frac{1}{n} Var(X) = 0$$

Estimatori pentru medie și dispersie

Estimatorul

$$D^2(X,n) = S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2$$

se numește **dispersie de selecție** și este un estimator corect pentru $Var(X)$. Avem

$$E(S^2) = \frac{n-1}{n} Var(X)$$

deci S^2 nu este un estimator nedeplasat pentru $Var(X)$.
Dacă se alege

$$s^2 = \frac{n}{n-1} S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

obținem un estimator absolut corect și nedeplasat pentru $Var(X)$.

Verosimilitate maximă

Fie X o variabilă aleatoare discretă care ia valorile $\{x_i | i = 1, \ldots, n\}$ cu probabilitățile depinzând de un parametru θ și $X_1, X_2, \ldots, X_n$ variabilele aleatoare de selecție independente corespunzătoare selecțiilor de volum n. Probabilitatea ca vectorul aleator $(X_1, X_2, \ldots, X_n)$ să ia valoarea $(x_1, x_2, \ldots, x_n)$ este $\prod_{i=1}^{n} P(x_i, \theta)$. Funcția

$$V(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} P(x_i, \theta)$$

se numește **funcție de verosimilitate** a variabilei aleatoare X.
Dacă X este o variabilă aleatoare continuă cu densitatea de repartiție $f(x, \theta)$, unde θ este un parametru care trebuie determinat, atunci funcția de verosimilitate se definește ca fiind densitatea de repartiție a vectorului aleator $(X_1, X_2, \ldots, X_n)$, adică

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

- *Principiul verosimilității maxime* constă în determinarea parametrului θ din condiția ca $V(x_1, x_2, \ldots, x_n; \theta)$, considerată ca o funcție diferențiabilă de θ pentru un eșantion dat, să admită un maxim.
- Deoarece $f(x) = \ln x$ este o funcție monoton crescătoare, rezultă că funcțiile $\ln V(x_1, x_2, \ldots, x_n; \theta)$ și $V(x_1, x_2, \ldots, x_n; \theta)$ își ating valoarea maximă pentru aceeași valoare a lui θ.
- Valoarea $\hat{\theta}$ a parametrului θ pentru care V admite un maxim se numește *estimație de verosimilitate maximă* și este soluția ecuației

$$\frac{\partial(\ln V)}{\partial \theta} = 0 \Leftrightarrow \sum_{i=1}^{n} \frac{1}{f(x_i, \theta)} \frac{\partial f(x_i, \theta)}{\partial \theta} = 0.$$

numită *ecuația de verosimilitate maximă*.

Exemplu: Pentru o variabilă aleatoare repartizată exponențial de parametru θ se obține funcția de verosimilitate

$$V = (\theta e^{-\theta x_1})(\theta e^{-\theta x_2}) \ldots (\theta e^{-\theta x_n}) = \theta^n e^{-\theta(x_1 + \cdots + x_n)}$$

care își atinge maximumul pentru $\hat{\theta} = \frac{n}{x_1 + \cdots + x_n}$, așadar estimatorul de verosimilitate maximă pentru parametrul θ este

$$\theta(X, n) = \frac{n}{X_1 + X_2 + \cdots + X_n}.$$

Interval de încredere pentru medie (σ cunoscut)

Intervalul de încredere pentru media unei variabile aleatoare X având repartiția $N(m, \sigma^2)$ cu $m \in \mathbb{R}$ necunoscut și $\sigma^2 > 0$ cunoscut este de forma $\left(\bar{X} - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right)$, cu

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

unde $\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$ este media de selecție, $\alpha \in (0, 1)$ este *nivelul de semnificație* iar $z_{1-\frac{\alpha}{2}}$ este cuantila de ordin $1 - \frac{\alpha}{2}$ a repartiției normale standard, mai exact:

$$z_{1-\frac{\alpha}{2}} = F^{-1}\left(1 - \frac{\alpha}{2}\right) = \Phi^{-1}\left(\frac{1-\alpha}{2}\right)$$

unde F este funcția de repartiție $N(0, 1)$ iar Φ este funcția lui Laplace.

- Valoarea $\Delta\bar{X} = z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ se numește *marjă de eroare* și cu ajutorul acesteia putem rescrie intervalul de încredere sub forma

$$\left(\bar{X} - \Delta\bar{X}, \bar{X} + \Delta\bar{X}\right)$$

- Cu cât nivelul de încredere $1 - \alpha$ este mai mare, cu atât marja de eroare $\Delta\bar{X}$ este mai mare și lungimea intervalului de încredere este mai mare.
- Pentru un nivel de încredere dat, volumul minim necesar al unui eșantion pentru a obține un interval de încredere cu marja de eroare $\Delta\bar{X}$ este

$$n = \left(\frac{z_{1-\frac{\alpha}{2}} \cdot \sigma}{\Delta\bar{X}}\right)^2$$

- Pentru selecții de volum mare, intervalul de încredere pentru medie este valabil și pentru cazul în care variabila X are o repartiție oarecare datorită teoremei limită centrală.

Exemplu

De la o mașină de îmbuteliat băuturi răcoritoare s-au testat 36 sticle și s-a obținut volumul mediu de 2.25 l. Presupunând că volumul este normal distribuit cu abaterea medie pătratică de 0.15 l, să se afle un interval de încredere 90% pentru volumul mediu. Cât de mare trebuie să fie volumul eșantionului pentru a avea o marjă de eroare de $0,01$?
Rezolvare:

- Avem $n = 36$, $\bar{X} = 2.25$ și $\sigma = 0.15$.
- Cum nivelul de încredere este $1 - \alpha = 0.9$, găsim nivelul de semnificație $\alpha = 0.1$, de unde $1 - \frac{\alpha}{2} = 0.95$.
- Cuantila corespunzătoare acestei valori din repartiția normală standard este $z_{1-\frac{\alpha}{2}} = 1.645$
- Marja de eroare $\Delta\bar{X} = z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} = 0.0411$
- Intervalul de încredere 90% este $\left(\bar{X} - \Delta\bar{X}, \bar{X} + \Delta\bar{X}\right) = (2.2089, 2.2911)$
- Volumul eșantionului necesar pentru a avea o marjă de eroare de $0,01$:

$$n = \left(\frac{z_{1-\frac{\alpha}{2}} \cdot \sigma}{\Delta\bar{X}}\right)^2 = 608.7473 \simeq 609$$

Exerciții

1. S-a efectuat un studiu asupra înălțimii sportivilor pe un eșantion de 50 persoane, în urma căruia a rezultat că înălțimea medie este 1.745 m, cu abaterea medie pătratică de 0.069 m. Să se găsească un interval de încredere 98% pentru înălțimea medie a sportivilor. Cât de mare trebuie să fie eșantionul pentru a avea o marjă de eroare de 0.01?
   R: $(1.722, 1.767)$; $n = 258$.
2. În urma unui studiu făcut pe un eșantion de 100 mașini, s-a obținut un număr mediu de kilometri parcurși anual de 23500, cu abaterea medie pătratică de 3900 km.
   Să se afle un interval de încredere 99% pentru numărul mediu de kilometri parcurși anual de o mașină. Cât de mare trebuie să fie volumul eșantionului pentru a aproxima acest număr mediu cu o marjă de eroare de 100km?
   R: $(22495.4266, 24504.5734)$; $n = 10092$.