Metode numerice și statistică
Curs 12

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

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

2019

Interval de încredere pentru medie (σ necunoscut)

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

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

unde $\bar{X} = \frac{X_1+X_2+\cdots+X_n}{n}$ este media de selecție, $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ este estimația nedeplasată a lui σ^2, $\alpha \in (0,1)$ este nivelul de semnificație iar $t_{n-1,1-\frac{\alpha}{2}}$ este cuantila de ordin $1-\frac{\alpha}{2}$ a repartiției Student cu $n-1$ grade de libertate:

$$t_{n-1,1-\frac{\alpha}{2}} = F_{n-1}^{-1}\left(1-\frac{\alpha}{2}\right)$$

unde F_{n-1} este funcția de repartiție Student cu $n-1$ grade de libertate.

Exemplu

S-a efectuat un studiu privind cantitatea de apă minerală dintr-un număr de 8 butelii și s-au obținut următoarele rezultate (în litri): 4.96, 4.90, 4.98, 5, 4.94, 5, 5.02, 4.92. Să se determine un interval de încredere 95% pentru volumul mediu al buteliilor de apă presupunând că volumul are o distribuție normală.
Rezolvare:

- Avem $n = 8$, $\bar{X} = \frac{4.96+4.90+4.98+5+4.94+5+5.02+4.92}{8} = 4.965$
- Estimația nedeplasată a dispersiei $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2 = 0.0018$
- Cum nivelul de încredere este $1-\alpha = 0.95$, găsim nivelul de semnificație $\alpha = 0.05$, de unde $1-\frac{\alpha}{2} = 0.975$.
- Cuantila corespunzătoare acestei valori din repartiția Student cu $n-1=7$ grade de libertate este $t_{n-1,1-\frac{\alpha}{2}} = 2.3646$
- Marja de eroare
$$t_{n-1,1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} = 0.0355$$
- Intervalul de încredere 95% este $(4.9295, 5.0005)$.

Interval de încredere pentru dispersie

Intervalul de încredere pentru dispersia unei variabile aleatoare X având repartiția $N(m,\sigma^2)$ cu $m \in \mathbb{R}$ necunoscut și $\sigma^2 > 0$ necunoscut este de forma

$$\left(\frac{(n-1)\cdot s^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \frac{(n-1)\cdot s^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right)$$

unde $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ este estimația nedeplasată a lui σ^2, $\alpha \in (0,1)$ este nivelul de semnificație, iar $\chi^2_{n-1,1-\frac{\alpha}{2}}$ și $\chi^2_{n-1,\frac{\alpha}{2}}$ sunt cuantilele de ordin $1-\frac{\alpha}{2}$ și $\frac{\alpha}{2}$ ale repartiției χ^2 cu $n-1$ grade de libertate.
Exemplu: Grosimea unui tip de sticlă are o repartiție normală. S-a efectuat un studiu pe 5 bucăți și s-au obținut următoarele rezultate (în mm): 4.8, 5, 4.6, 4.4, 5.2. Să se determine un interval de încredere 90% pentru abaterea medie pătratică de la grosimea medie. **R**: $(0.2053, 0.7502)$

Exerciții

1. S-a făcut o analiză chimică a unei substanțe din 7 containere obținându-se următoarele cantități (în litri): 9.7, 9.8, 10, 10.1, 10.2, 10.4, 9.6. Să se afle un interval de încredere 95% pentru conținutul mediu de substanță, respectiv pentru abaterea medie pătratică. Cât trebuie să fie volumul eșantionului pentru a avea o marjă de eroare a conținutului mediu de 0.1?
R: $\bar{X} = 9.9714$, $I = (9.706, 10.237)$;
$s = 0.287$, $I = (0.185, 0.632)$; $N = 50$
2. S-a efectuat un studiu asupra greutății medii a 10 ciocolate și s-au obținut următoarele rezultate (în grame): 101, 104, 98, 105, 97, 96, 102, 101, 99, 103. Să se afle un interval de încredere 96% pentru greutatea medie și abaterea medie pătratică, presupunând că greutatea este normal distribuită. Cât trebuie să fie volumul eșantionului pentru a avea o marjă de eroare a greutății medii de 1?
R: $\bar{X} = 100.6$, $I = (98.305, 102.895)$;
$s = 3.026$, $I = (2.046, 5.704)$; $N = 53$

Noțiuni generale

- Fie X o variabilă aleatoare teoretică corespunzătoare unei caracteristici a unei populații. Pentru a stabili tipul repartiției, valorile parametrilor necunoscuți sau unele caracteristici numerice (media, dispersia) ale variabilei aleatoare X, se fac anumite ipoteze care trebuie verificate pe baza datelor obținute în urma unor selecții.
- Prin **ipoteză statistică** înțelegem o afirmație privind fie tipul repartiției, fie valorile caracteristicilor numerice, fie valorile parametrilor necunoscuți ai repartiției, afirmație care poate fi *acceptată* sau *respinsă*.
- Dacă variabila aleatoare X are o repartiție cunoscută, iar ipoteza se referă la parametrii acestei repartiții, se numește **ipoteză parametrică**.
- Dacă variabila aleatoare X are o repartiție necunoscută, iar ipoteza se referă la tipul acestei repartiții, se numește **ipoteză neparametrică**.

- Ipotezele parametrice sunt numite *parametrice simple* atunci când se referă la toți parametrii funcției de repartiție ai variabilei aleatoare X, atribuindu-i fiecăruia anumite valori.
- Ipotezele parametrice sunt numite *parametrice compuse* când se referă numai la unii dintre parametri sau presupune apartenența a cel puțin unuia dintre ei la o mulțime de valori.
- Ipoteza care urmează să fie testată se numește **ipoteză nulă** și se notează cu H_0.
- Orice altă ipoteză care se poate accepta când se respinge H_0 se numește **ipoteză alternativă** și se notează cu H_a.
- Verificarea ipotezelor statistice se face prin *teste statistice*, care sunt metode decizionale de acceptare sau respingere a ipotezelor (pe baza prelucrării datelor obținute în urma unor selecții) și au la bază criterii de testare.

Definiție

O funcție de selecție $C(X,n)$ se numește **criteriu de testare** *al ipotezei H_0 dacă îndeplinește condițiile:*

1. *repartiția variabilei aleatoare $C(X,n)$ depinde de valoarea de adevăr a ipotezei H_0;*
2. *dacă ipoteza H_0 este adevărată, atunci variabila aleatoare $C(X,n)$ are repartiție complet specificată.*

Deoarece eșantionul ales nu este întotdeauna reprezentativ pentru întreaga populație, pot apărea două tipuri de erori:

- *eroare de tipul I*: ipoteza H_0 este respinsă (se acceptă ipoteza H_a) deși în realitate este adevărată:
$$\alpha = P(H_0 \text{ este respinsă} | H_0 \text{ este adevărată})$$
- *eroare de tipul II*: ipoteza H_0 este acceptată (se respinge ipoteza H_a) deși în realitate este falsă:
$$\beta = P(H_0 \text{ este acceptată} | H_0 \text{ este falsă})$$

Probabilitatea de a face o eroare de tipul II poate fi rescrisă astfel:

$$\beta = P(H_0 \text{ este acceptată} | H_a \text{ este adevărată})$$

sau

$$1-\beta = P(H_0 \text{ este respinsă} | H_a \text{ este adevărată}).$$

Probabilitatea maximă acceptată pentru o eroare de tipul I se numește **nivel de semnificație**. Pentru un nivel de specificație dat α se poate determina un interval sau o reuniune de intervale $W \subset \mathbb{R}$ astfel încât :

1. $P(C(X,n) \in W | H_0 \text{ este adevărată}) = \alpha$ dacă $C(X,n)$ este variabilă aleatoare continuă;
2. $P(C(X,n) \in W | H_0 \text{ este adevărată}) \leq \alpha$ dacă $C(X,n)$ este variabilă aleatoare discretă;

Mulțimea W se numește **regiune de respingere** (sau **regiune critică**) pentru ipoteza H_0, iar complementara $\bar{W}$ se numește **regiune de acceptare** a ipotezei H_0.

Pentru testarea unei ipoteze statistice se parcurg următoarele etape:

1. Se alege ipoteza nulă H_0 și ipoteza alternativă H_a
2. Se alege un criteriu de testare $C(X, n)$
3. Se alege un nivel de semnificație α
4. Se determină regiunea de respingere W cu proprietatea că

$$P(C(X,n) \in W | H_0 \text{ este adevărată}) = \alpha \text{ (sau } \leq \alpha)$$

5. Se face o selecție de volum n, obținându-se valorile $x_1, x_2, \ldots, x_n$ ale variabilelor aleatoare de selecție $X_1, X_2, \ldots, X_n$ corespunzătoare caracteristicii studiate
6. Se calculează $C(x_1, x_2, \ldots, x_n)$:
   - Dacă $C(x_1, x_2, \ldots, x_n) \in W$ atunci H_0 este acceptată
   - Dacă $C(x_1, x_2, \ldots, x_n) \in W$ atunci H_0 este respinsă (H_a este acceptată)

Test bilateral pentru medie (σ cunoscut)

Fie X o variabilă aleatoare teoretică cu repartiție normală $N(m, \sigma^2)$ cu media necunoscută și m_0 o estimație a acestei medii. Se face o selecție de volum n, obținându-se variabilele aleatoare de selecție $X_1, X_2, \ldots, X_n$.

1. Dorim să verificăm ipoteza

$$H_0 : \ m = m_0$$

   față de alternativa

$$H_a : \ m \neq m_0.$$

2. Alegem criteriul de testare

$$C(X,n) = \frac{\bar{X} - m_0}{\frac{\sigma}{\sqrt{n}}}$$

   deoarece dacă H_0 este adevărată, atunci $C(X, n)$ are repartiția complet specificată $N(0, 1)$.
3. Alegem un nivel de semnificație α (de obicei 0.1, 0.05 sau 0.01).

4. Se determină z = cuantila de ordin $1 - \frac{\alpha}{2}$ a repartiției normale standard, corespunzătoare nivelului de semnificație ales α. Avem

$$P(|C(X,n)| > z | m = m_0) = \alpha$$

   deci regiunea de respingere este

$$W = (-\infty, -z) \cup (z, \infty).$$

5. Se face o selecție de volum n, obținându-se valorile $x_1, x_2, \ldots, x_n$ și se calculează

$$\bar{X} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

6. Dacă $\frac{\bar{X} - m_0}{\frac{\sigma}{\sqrt{n}}} \in [-z, z]$ atunci H_0 este acceptată.

   Dacă $\frac{\bar{X} - m_0}{\frac{\sigma}{\sqrt{n}}} \in (-\infty, -z) \cup (z, \infty)$ atunci H_0 este respinsă și se acceptă H_a.

Exemplu

La o fabrică patronul asigură că durata medie de funcționare a unui produs este de 800 ore cu abaterea medie pătratică de 50 ore. Pe un eșantion de 20 bucăți s-a obținut că durata medie de funcționare este 790 ore. Presupunând că durata de funcționare este normal distribuită să se verifice cu nivelul de semnificație de 4% dacă afirmația patronului este adevărată.

1. ipotezele $H_0 : \ m = 800, \qquad H_a : m \neq 800$
2. criteriul de testare $\frac{\bar{X} - 800}{\frac{50}{\sqrt{20}}}$
3. nivelul de semnificație $\alpha = 0.04$
4. cuantila de ordin $1 - \frac{\alpha}{2} = 0.98$ este $z = 2.05$
5. pentru $\bar{X} = 790$ avem $\frac{\bar{X} - 800}{\frac{50}{\sqrt{20}}} = -0.894$
6. ipoteza H_0 se acceptă deoarece $-0.894 \in [-2.05, 2.05]$

Teste unilaterale pentru medie (σ cunoscut)

Test unilateral dreapta:

- $H_0 : \ m = m_0, \qquad H_a : \ m > m_0$
- criteriul de testare $C(X,n) = \frac{\bar{X} - m_0}{\frac{\sigma}{\sqrt{n}}}$
- z = cuantila de ordin $1 - \alpha$ a repartiției normale standard, corespunzătoare nivelului de semnificație ales α.
- regiunea de respingere este $W = (z, \infty)$.

Test unilateral stânga:

- $H_0 : \ m = m_0, \qquad H_a : \ m < m_0$
- criteriul de testare $C(X,n) = \frac{\bar{X} - m_0}{\frac{\sigma}{\sqrt{n}}}$
- z = cuantila de ordin α a repartiției normale standard, corespunzătoare nivelului de semnificație ales α.
- regiunea de respingere este $W = (-\infty, z)$.

Teste pentru proporție

Considerăm problema testării ipotezei că proporția de indivizi dintr-o populație care au o anumită caracteristică are o valoare specificată. Considerăm variabila aleatoare teoretică Bernoulli $X : \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$, unde 1 înseamnă că individul are caracteristica respectivă. Media de selecție corespunzătoare $\hat{p} = \frac{X_1 + X_2 + \cdots + X_n}{n}$ este un estimator nedeplasat pentru proporția p și pentru valori mari ale lui n are repartiția $N\left(p, \frac{pq}{n}\right)$ unde $q = 1 - p$.

1. Se formulează ipotezele $H_0 : \ p = p_0, \qquad H_a : \ p \neq p_0$
2. Se alege criteriul de testare $C(X,n) = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
3. Se alege un nivel de semnificație α
4. Se determină z = cuantila de ordin $1 - \frac{\alpha}{2}$ a repartiției normale standard, corespunzătoare nivelului de semnificație ales α
5. Se află estimația punctuală a proporției $\hat{p}$ pe un eșantion de volum n
6. Dacă $\frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \in [-z, z]$ atunci H_0 este acceptată.

   Dacă $\frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \in (-\infty, -z) \cup (z, \infty)$ atunci H_0 este respinsă.