STATISTICĂ - LABORATOR 7

Teste pentru compararea dispersiilor şi mediilor

1. Testul F pentru compararea dispersiilor a două legi normale

(doar test bilateral)

Se consideră două caracteristici independente X_1 şi X_2. Se presupune că X_1 şi X_2 urmează fiecare legea normală:

$$X_1 \sim N(m_1, \sigma_1) \;\text{ şi }\; X_2 \sim N(m_2, \sigma_2).$$

Pentru fiecare caracteristică, se consideră câte o selecţie repetată de volum n_1, respectiv n_2. Fie

- $X_{11}, X_{12}, \ldots, X_{1n_1}$ variabilele de selecţie pentru selecţia 1;
- $X_{21}, X_{22}, \ldots, X_{2n_2}$ variabilele de selecţie pentru selecţia 2.

Se definesc:

- mediile de selecţie ($\boxed{\text{mean}}$):

$$\overline{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1i}, \qquad \overline{X}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} X_{2j} \tag{1}$$

- dispersiile de selecţie ($\boxed{\text{var}}$):

$$s_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_{1i}-\overline{X}_1)^2, \qquad s_2^2 = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(X_{2j}-\overline{X}_2)^2 \tag{2}$$

- abaterile standard de selecţie ($\boxed{\text{std}}$):

$$s_1 = \sqrt{s_1^2}, \qquad s_2 = \sqrt{s_2^2} \tag{3}$$

Relativ la dispersiile teoretice ale celor două caracteristici, verificăm ipoteza nulă:

$H_0 : \sigma_1^2 = \sigma_2^2$ cu alternativa:

$H_1 : \sigma_1^2 \neq \sigma_2^2$ **test F bilateral**.

Statistica de test care se utilizează este

$$F = \frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}}, \tag{4}$$

care urmează legea $F(n_1-1, n_2-1)$.

Se dă un nivel de semnificaţie $\alpha \in (0,1)$.

Definim regiunea U de respingere a ipotezei nule H_0, prin:

$$(-\infty, f_{\frac{\alpha}{2}}\Big] \cup \Big[f_{1-\frac{\alpha}{2}}, +\infty) \tag{5}$$

unde cuantilele $f_{\frac{\alpha}{2}}$ şi $f_{1-\frac{\alpha}{2}}$ se referă la legea $F(n_1-1, n_2-1)$. ($\boxed{\text{finv}}$)

Se calculează valoarea f_0 a statisticii de test F, atunci când ipoteza nulă H_0 este adevărată:

$$f_0 = \frac{s_1^2}{s_2^2} \tag{6}$$

<u>Luarea deciziei</u>:
dacă $f_0 \in U$ se respinge ipoteza nulă H_0 (în acest caz se acceptă alternativa H_1)
dacă $f_0 \notin U$ se acceptă ipoteza nulă H_0

<u>Funcţia Matlab `vartest2`</u>

Sistemul Matlab, prin *Statistics toolbox*, dispune de funcţia `vartest2`, cu aplicabilitate la testul F pentru compararea dispersiilor a două legi normale. Apelarea acestei funcţii se face prin:

```
[h,P,ci,stats] = vartest2(x,y,alpha,tail);
```

În urma executării acestei instrucţiuni, se efectuează testul F asupra datelor conţinute în vectorii `x` şi `y`, folosind nivelul de semnificaţie `alpha`.

Parametrul `tail` specifică una din cele 3 alternative, care conduc la testul unilateral stânga (`tail=-1`), unilateral dreapta (`tail=1`) şi bilateral (`tail=0`).

Dacă `h=1`, atunci ipoteza nulă H_0 va fi respinsă, iar dacă `h=0`, ipoteza nulă H_0 va fi acceptată.

Apelul permite obţinerea valorii critice `P`, precum şi a intervalului de încredere pentru raportul dispersiilor teoretice, corespunzător probabilităţii de încredere `1-alpha`, obţinut în vectorul cu două componente `ci`.

Parametrul `stats` are următoarele câmpuri:
`fstat` – valoarea statisticii de test
`df1, df2` – numărul gradelor de libertate ale testului.

De exemplu, pentru returnarea valorii statisticii de test, se utilizează instrucţiunea:
`f0= stats.fstat`

2. Teste pentru compararea mediilor

Se consideră două caracteristici <u>independente</u> X_1 şi X_2. Se presupune că:

$$X_1 \sim N(m_1, \sigma_1) \quad \text{şi} \quad X_2 \sim N(m_2, \sigma_2).$$

Pentru fiecare caracteristică, se consideră câte o selecţie repetată de volum n_1, respectiv n_2.

Relativ la mediile teoretice ale celor două caracteristici, verificăm ipoteza nulă:

$H_0 : m_1 = m_2$ cu una din alternativele:

$H_1 : m_1 < m_2$ **test la stânga**

$H_1 : m_1 > m_2$ **test la dreapta**

$H_1 : m_1 \neq m_2$ **test bilateral**

Distingem următoarele cazuri:

A) dispersiile σ_1^2 şi σ_2^2 sunt cunoscute

Se aplică testul Z pentru compararea mediilor.

Statistica de test care se utilizează este

$$Z = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N\left(0, 1\right).$$

Se dă un nivel de semnificaţie $\alpha \in (0, 1)$.

Se defineşte regiunea U de respingere a ipotezei nule H_0, astfel:

- $U = (-\infty, z_\alpha]$ pentru testul Z la stânga
- $U = [z_{1-\alpha}, \infty)$ pentru testul Z la dreapta
- $U = \left(-\infty, -z_{1-\frac{\alpha}{2}}\right] \cup \left[z_{1-\frac{\alpha}{2}}, +\infty\right)$ pentru testul Z bilateral

unde cuantilele z_α, $z_{1-\alpha}$ şi $z_{1-\frac{\alpha}{2}}$ se referă la legea $N(0, 1)$. (norminv)

Se calculează valoarea z_0 a statisticii de test Z, atunci când ipoteza H_0 este adevărată:

$$z_0 = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Luarea deciziei:

dacă $z_0 \in U$, se respinge ipoteza nulă H_0 (în acest caz se acceptă alternativa H_1)

dacă $z_0 \notin U$, se acceptă ipoteza nulă H_0.

B) dispersiile σ_1^2 şi σ_2^2 sunt necunoscute şi egale $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Se aplică testul T pentru compararea mediilor.

Statistica de test care se utilizează este

$$T = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (m_1 - m_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim T(n_1 + n_2 - 2) \tag{7}$$

unde s_p este abaterea standard de selecţie combinată, definită astfel:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{8}$$

Se dă un nivel de semnificaţie $\alpha \in (0, 1)$.

Corespunzător celor trei alternative, definim regiunea U de respingere a ipotezei nule H_0, prin:

- $(-\infty, t_\alpha]$ (test T la stânga)
- $[t_{1-\alpha}, \infty)$ (test T la dreapta)
- $\left(-\infty, -t_{1-\frac{\alpha}{2}}\right] \cup \left[t_{1-\frac{\alpha}{2}}, +\infty\right)$ (test T bilateral)

unde cuantilele t_α, $t_{1-\alpha}$ şi $t_{1-\frac{\alpha}{2}}$ se referă la legea $T(n_1 + n_2 - 2)$. ($\boxed{\text{tinv}}$)

Se calculează valoarea t_0 a statisticii de test T, atunci când ipoteza nulă H_0 este adevărată:

$$t_0 = \frac{(\overline{x}_1 - \overline{x}_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{9}$$

Luarea deciziei

dacă $t_0 \in U$ se respinge ipoteza nulă H_0 (în acest caz se acceptă alternativa H_1)
dacă $t_0 \notin U$ se acceptă ipoteza nulă H_0

C) dispersiile σ_1^2 şi σ_2^2 sunt necunoscute şi diferite

Se aplică testul T pentru compararea mediilor.

Statistica de test care se utilizează este

$$T = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (m_1 - m_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \tag{10}$$

care urmează legea $T(n)$, unde numărul n al gradelor de libertate se calculează din formula:

$$\frac{1}{n} = \frac{c^2}{n_1 - 1} + \frac{(1-c)^2}{n_2 - 1}, \quad \text{iar} \quad c = \frac{\frac{s_1^2}{n_1}}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}. \tag{11}$$

Se dă un nivel de semnificaţie $\alpha \in (0, 1)$.

Corespunzător celor trei alternative, definim regiunea U de respingere a ipotezei nule H_0, prin:

- $(-\infty, t_\alpha]$ (test T la stânga)
- $[t_{1-\alpha}, \infty)$ (test T la dreapta)
- $\left(-\infty, -t_{1-\frac{\alpha}{2}}\right] \cup \left[t_{1-\frac{\alpha}{2}}, +\infty\right)$ (test T bilateral)

unde cuantilele t_α, $t_{1-\alpha}$ şi $t_{1-\frac{\alpha}{2}}$ se referă la legea $T(n)$, cu n obţinut din formula (11).

Se calculează valoarea t_0 a statisticii de test T, atunci când ipoteza nulă H_0 este adevărată:

$$t_0 = \frac{(\overline{x}_1 - \overline{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{12}$$

Luarea deciziei:

dacă $t_0 \in U$ se respinge ipoteza nulă H_0 (în acest caz se acceptă alternativa H_1)
dacă $t_0 \notin U$ se acceptă ipoteza nulă H_0

Funcţia Matlab `ttest2` (cu opţiunea 'equal')

Sistemul Matlab, prin *Statistics toolbox*, dispune de funcţia `ttest2`, cu aplicabilitate la testul T pentru compararea a două medii, când dispersiile sunt necunoscute şi egale. Apelarea acestei funcţii se face prin:

```
[h,P,ci,stats] = ttest2(x,y,alpha,tail,'equal');
```

Funcţia Matlab `ttest2` (cu opţiunea 'unequal')

În Matlab, testul T pentru compararea mediilor când dispersiile sunt necunoscute şi diferite este implementat prin apelarea funcţiei `ttest2`, astfel:

```
[h,P,ci,stats] = ttest2(x,y,alpha,tail,'unequal')
```

În urma executării unei astfel de instrucţiuni, se efectuează testul T asupra datelor conţinute în vectorii `x` şi `y`, folosind nivelul de semnificaţie `alpha`.

Parametrul `tail` specifică una din cele 3 alternative, care conduc la testul unilateral stânga (`tail=-1`), unilateral dreapta (`tail=1`) şi bilateral (`tail=0`).

Dacă `h=1`, atunci ipoteza nulă H_0 va fi respinsă, iar dacă `h=0`, ipoteza nulă H_0 va fi acceptată.

Apelul permite obţinerea valorii critice `P`, precum şi a intervalului de încredere pentru diferenţa mediilor teoretice, corespunzător probabilităţii de încredere `1-alpha`, obţinut în vectorul cu două componente `ci`.

Parametrul `stats` are următoarele câmpuri:

`tstat` – valoarea statisticii de test

`df` – numărul gradelor de libertate ale testului

`sd` – abaterea standard de selecţie.

De exemplu, pentru returnarea valorii statisticii de test, se utilizează instrucţiunea:

```
t0= stats.tstat
```

Observaţie: Rezultatele obţinute la testele pentru compararea mediilor (cazurile A, B, C) pot fi aplicate şi în cazul a două caracteristici independente X_1 şi X_2 care nu urmează legea normală, dacă volumul selecţiilor este mare, adică $n_1, n_2 \geq 20$.