Laborator 6 - Statistică inferenţială

I. Inferenţă asupra dispersiilor a două populaţii - Testul F (simetric) asupra dispersiilor a două populaţii

Se consideră o două populaţii statistice cărora nu li se cunosc dispersiile σ_1^2 şi σ_2^2. Din cele două populaţii se extrag două eşantioane aleatoare simple (şi independente între ele) cărora li se calculează dispersiile s_1^2 şi s_2^2. Scorul $F = \frac{s_1^2}{s_2^2}$ este distribuit $F(n_1 - 1, n_2 - 1)$.

Testul F decurge astfel:

1. se formulează ipoteza nulă, care susţine că dispersiile celor două populaţii sunt egale:

$$\boxed{H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1}$$

2. se formulează ipoteza alternativă care susţine că dispersiile sunt diferite:

$$\boxed{H_a : \frac{\sigma_1^2}{\sigma_2^2} \neq 1}$$

3. se fixează nivelul de semnificaţie: α (care uzual poate fi 1% sau 5%);
4. se calculează scorul testului:

$$\boxed{F = \frac{s_1^2}{s_2^2}}$$

5. se determină valoarile critice F_s^* şi F_d^*:

$$\boxed{F_s^* = qf(\alpha/2, n_1 - 1, n_2 - 1)} \quad \boxed{F_d^* = qf(1 - \alpha/2, n_2 - 1, n_1 - 1) = 1/F_s^*}$$

6. **ipoteza nulă H_0 este respinsă** şi se acceptă că dispersiile sunt diferite dacă

$$\boxed{F \notin (F_s^*, F_d^*)}$$

altfel nu există suficiente dovezi pentru a accepta că dispersiile sunt diferite.

Exerciţiu rezolvat. Rezultatele unui test psihologic efectuat pe două eşantioane, unul de femei şi unul de bărbaţi sunt următoarele:

bărbaţi: $n_1 = 120, s_1 = 5.05$

femei: $n_2 = 135, s_2 = 5.44$

Se poate trage concluzia că dispersiile celor două populaţii diferă semnificativ (1%)?

```
> alfa = 0.01
> n1= 120
> n2 = 135
> s1 = 5.05
> s2 = 5.44
> critical_F_s= qf(alfa/2, n1 - 1, n2 - 1)
> critical_F_d= qf(1 - alfa/2, n2 - 1, n1 - 1)
> critical_F_s
> critical_F_d
> F_score
```

Rezultatul va fi $F = 0.86175$, $F_s^* = 0.62843$ şi $F_d^* = 1.59125$; deoarece $F \in [F_s^*, F_d^*]$ ipoteza nulă nu poate fi respinsă. în acest caz putem considera că nu există dovezi semnificative pentru a afirma că dispersiile sunt diferite.

Exerciţii propuse

I.1 Scrieţi o funcţie (numită **F_test**) care să calculeze şi să returneze valoarile critice şi scorul testului F (parametrii funcţiei vor fi: tipul ipotezei alternative, α, n_1, n_2, σ_1 etc., eşantioanele se pot extrage din fişier ca mai jos). Funcţia aceasta va fi utilizată apoi la rezolvarea exerciţiilor care urmează.

```
> x1 = read.table("program.txt", header = TRUE)[['A']]
> x2 = read.table("program.txt", header = TRUE)[['B']]
> n1 = length(x1)
> s1 = sd(x1)
> ...
```

I.2 Un profesor crede că un anumit program de lectură imbunătăţeşte abilităţile şi dorinţa copiilor de a citi. Pentru aceasta el alege două grupuri de elevi: unul de 22 de elevi care urmează programul prescris (A) şi unul de 24 de elevi care nu urmează acest program (B). Rezultatele sunt date în fişierul *program.txt*.

Să se decidă cu 1% şi 5% nivel de semnificaţie dacă dispersiile celor două populaţii sunt diferite.

I.3 Cercetătorii studiază amplitudinea mişcării obţinută prin stimularea nervoasă a şoarecilor. Pentru şoarecii drogaţi se obţin următoarele date:

12.512 12.869 19.098 15.350 13.297 15.589

Pentru şoarecii normali se obţin următoarele date:

11.074 9.686 12.164 8.351 12.182 11.489

Influenţa drogurilor este semnificativă în ceea ce priveşte cele doua dispersii (5% nivel de semnificaţie)?

II. Inferenţă asupra mediilor a două populaţii - Testul t pentru diferenţa mediilor unor populaţii cu dispersii necunoscute

Se consideră o două populaţii statistice cărora nu li se cunosc dispersiile. Se aleg două eşantioane aleatoare simple şi independente între ele cu mediile de selecţie $\overline{x}_{n_1}$ şi $\overline{x}_{n_2}$ şi dispersiile σ_1^2 şi σ_2^2. Înaintea efectuării testului t se foloseşte testul F pentru a decide dacă dispersiile celor două populaţii sunt diferite. Dacă dispersiile nu sunt diferite $\boxed{t = \dfrac{(\overline{x}_{n_1} - \overline{x}_{n_2}) - m_0}{\sqrt{\dfrac{s^2}{n_1} + \dfrac{s^2}{n_2}}}}$ urmează o distribuţie Student[1]: $t(n_1 + n_2 - 2)$. Dacă dispersiile sunt diferite, atunci $\boxed{t = \dfrac{(\overline{x}_{n_1} - \overline{x}_{n_2}) - m_0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}}$ urmează o distribuţie Student: $t(\min(n_1 - 1, n_2 - 1))$.

Testul t decurge astfel:

[1] În acest caz $s = \sqrt{\dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$

1. se formulează ipoteza nulă, care susţine că diferenţa mediilor celor două populaţii ia o valoare particulară (de cele mai multe ori zero):

$$\boxed{H_0 : \mu_1 - \mu_2 = m_0}$$

2. se formulează o ipoteză alternativă care poate fi de trei feluri:

$\boxed{H_a : \mu_1 - \mu_2 < m_0}$ (ipoteză asimetrică la stânga) sau

$\boxed{H_a : \mu_1 - \mu_2 > m_0}$ (ipoteză asimetrică la dreapta) sau

$\boxed{H_a : \mu_1 - \mu_2 \neq m_0}$ (ipoteză simetrică)

3. se fixează nivelul de semnificaţie: α (care uzual poate fi 1% sau 5%);

4. se calculează scorul testului:

a) dacă dispersiile sunt diferite:

$$\boxed{t = \frac{(\overline{x}_{n_1} - \overline{x}_{n_2}) - m_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}}$$

iar numărul de grade de libertate este $df = \min(n_1 - 1, n_2 - 1)$.

b) dacă dispersiile sunt "egale"

$$\boxed{t = \frac{(\overline{x}_{n_1} - \overline{x}_{n_2}) - m_0}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}}$$

unde $s^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$, iar numărul de grade de libertate este $df = n_1 + n_2 - 2$.

5. se determină valoarea critică t^*:

$\boxed{t^* = qt(2\alpha, df)}$ pentru ipoteză H_a asimetrică la stânga ($t^* < 0$),

$\boxed{t^* = qt(1 - \alpha, df)}$ pentru ipoteză H_a asimetrică la dreapta ($t^* > 0$),

$\boxed{t^* = -qt(\alpha/2, df) = qt(1 - \alpha/2, df)}$ pentru ipoteză H_a simetrică ($t^* > 0$).

6. **ipoteza nulă H_0 este respinsă dacă**

$\boxed{t < t^*}$ pentru ipoteză H_a asimetrică la stânga sau

$\boxed{t > t^*}$ pentru ipoteză H_a asimetrică la dreapta sau

$\boxed{|t| > |t^*|}$ pentru ipoteză H_a imetrică,

altfel se spune că **nu există suficiente dovezi pentru a respinge ipoteza nulă H_0 şi a accepta ipoteza alternativă H_a.**

Exerciţiu rezolvat. Rezultatele unui test psihologic efectuat pe două eşantioane, unul de femei şi unul de bărbaţi sunt următoarele:

$$\text{bărbaţi: } n_1 = 110, \overline{x}_1 = 25.84, s_1 = 4.25$$

$$\text{femei: } n_2 = 105, \overline{x}_1 = 21.53, s_2 = 3.85$$

Se poate trage concluzia că mediile celor două populaţii diferă semnificativ (1%)?

Observaţie. Testarea diferenţei mediilor echivalează cu o ipoteză alternativă de tipul $\mu_1 - \mu_2 \neq 0 = m_0$.

```
> alfa = 0.01
> m0 = 0
> sample1_mean = 25.84
> sample2_mean = 21.53
> n1 = 110
> n2 = 105
> s1 = 4.25
> s2 = 3.85
> critical_F_s= qf(alfa/2, n1 - 1, n2 - 1) # testul F
> critical_F_d= qf(1 - alfa/2, n2 - 1, n1 - 1)
> F_score = s1^2/ s2^2
> if(F_score < critical_F_s | F_score > critical_F_d) {
+ df = min(n1 - 1, n2 - 1)
+ combined_s = sqrt(s1^2/n1 + s2^2/n2)
+ } else {
+ df = n1 + n2 - 2
+ combined_s = sqrt(((n1 - 1)*s1^2 + (n2 - 1)*s2^2)/df)*sqrt(1/n1+1/n2)
+ }
> critical_t = qt(1 - alfa/2, df)
> t_score = (sample1_mean - sample2_mean - m0)/combined_sigma
> critical_t
> t_score
```

Rezultatul va fi $|t^*| = 2.59910 > |t| = 1.87248$, ipoteza nulă nu poate fi respinsă.

Exerciţii propuse

II.1 Scrieţi o funcţie (numită, de exemplu, **t_test _means**) care să calculeze şi să returneze valoarea critică şi scorul testului t pentru diferenţa mediilor; această funcţie va include şi testul F simetric (parametrii vor fi α, n_1, n_2, s_1 etc.). Funcţia aceasta va fi utilizată, pentru rezolvarea exerciţiilor care urmează.

I.2 Un profesor crede că un anumit program de lectură imbunătăţeşte abilităţile şi dorinţa copiilor de a citi. Pentru aceasta el alege două grupuri de elevi: unul de 22 de elevi care urmează programul prescris (A) şi unul de 24 de elevi care nu urmează acest program (B). Rezultatele sunt date în fişierul *program.txt*.

Să se decidă cu 1% şi 5% nivel de semnificaţie dacă mediile celor două populaţii diferă semnificativ.

I.3 Cercetătorii studiază amplitudinea mişcării obţinută prin stimularea nervoasă a şoarecilor. Pentru şoarecii drogaţi se obţin următoarele date:

12.512 12.869 19.098 15.350 13.297 15.589

Pentru şoarecii normali se obţin următoarele date:

11.074 9.686 12.164 8.351 12.182 11.489

Influenţa drogurilor este semnificativă în ceea ce priveşte cele două medii (1% nivel de semnificaţie)?

III. Inferenţă asupra mediei - Testul Z pentru diferenţa mediilor unor populaţii cu dispersii cunoscute

Se consideră o două populaţii statistice cărora li se cunoasc dispersiile σ_1^2 şi σ_2^2. Se aleg două eşantioane aleatoare simple şi independente între ele cu mediile de selecţie $\overline{x}_{n_1}$ şi $\overline{x}_{n_2}$. Dacă populaţiile urmează o lege normală sau dimensiunea eşantioanelor este suficient de mare, scorul

$$z = \frac{(\overline{x}_{n_1} - \overline{x}_{n_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

este distribuit (eventual cu aproximaţie) normal standard: $N(0,1)$.

Testul Z decurge astfel:

1. se formulează ipoteza nulă, care susţine că diferenţa mediilor celor două populaţii ia o valoare particulară:
$$\boxed{H_0 : \ \mu_1 - \mu_2 = m_0}$$
2. se formulează o ipoteză alternativă care poate fi de trei feluri:
$$\boxed{H_a : \ \mu_1 - \mu_2 < m_0} \quad \text{(ipoteză asimetrică la stânga) sau}$$
$$\boxed{H_a : \ \mu_1 - \mu_2 > m_0} \quad \text{(ipoteză asimetrică la dreapta) sau}$$
$$\boxed{H_a : \ \mu_1 - \mu_2 \neq m_0} \quad \text{(ipoteză simetrică)}$$
3. se fixează nivelul de semnificaţie: α (care uzual poate fi 1% sau 5%);
4. se calculează scorul testului:
$$\boxed{\frac{(\overline{x}_{n_1} - \overline{x}_{n_2}) - m_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}}$$
5. se determină valoarea critică z^*:
$$\boxed{z^* = qnorm(\alpha, 0, 1)} \quad \text{pentru ipoteză } H_a \text{ asimetrică la stânga,}$$
$$\boxed{z^* = qnorm(1-\alpha, 0, 1)} \quad \text{pentru ipoteză } H_a \text{ asimetrică la dreapta,}$$
$$\boxed{z^* = qnorm(1-\alpha/2, 0, 1)} \quad \text{pentru ipoteză } H_a \text{ simetrică.}$$
6. **ipoteza nulă H_0 este respinsă dacă**
$$\boxed{z < z^*} \quad \text{pentru ipoteză } H_a \text{ asimetrică la stânga sau}$$
$$\boxed{z > z^*} \quad \text{pentru ipoteză } H_a \text{ asimetrică la dreapta sau}$$
$$\boxed{|z| > |z^*|} \quad \text{pentru ipoteză } H_a \text{ simetrică,}$$

altfel se spune că **nu există suficiente dovezi pentru a respinge ipoteza nulă H_0 şi a accepta ipoteza alternativă H_a.**

Exerciţiu rezolvat. Se compara durata de viata a doua tipuri de baterii. Primul tip are o deviatie standard de 4 ore, al doilea tip are o deviatie standard de 3 ore. Se aleg doua esantioane fiecare de dimensiune de 100 de baterii. Pentru primul esantion media de viata este de 48 de ore, iar pentru cel de-al doilea de 47 de ore.

Sa se testeze diferenta mediilor de viata cu 5% nivel de semnificaţie.

Observaţie. Testarea diferenţei mediilor echivalează cu o ipoteză alternativă de tipul $\mu_1 - \mu_2 \neq 0 = m_0$.

```
> alfa = 0.05
> m0 = 0
> sample1_mean = 48
> sample2_mean = 47
> n1 = 100
> n2 = 100
> sigma1 = 4
> sigma2 = 3
> combined_sigma = sqrt(sigma1^2/n1 + sigma2^2/n2)
> critical_z = qnorm(1 - alfa/2
> z_score = (sample1_mean - sample2_mean - m0)/combined_sigma
> critical_z
> z_score
```

Rezultatul va fi $|z^*| = 1.95996 < |z| = 2.00$, ipoteza nulă va fi respinsă şi se acceptă că mediile celor două populaţii sunt diferite.

Exerciţii propuse

III.1 Scrieţi o funcţie (numită, de exemplu, **z_test _means**) care să calculeze şi să returneze valoarea critică şi scorul testului z pentru diferenţa mediilor (parametrii vor fi α, n_1, n_2, σ_1 etc.). Funcţia aceasta va fi utilizată, pentru rezolvarea exerciţiilor de mai jos.

III.2 80 dintre angajaţii unei firme au un salariu mediu săptămânal de 160$ (deviaţia standard a întregii populaţii fiind 3.24$). 70 dintre angajaţii unei alte firme au în medie 155$ salariu pe săptămână (deviaţia standard a întregii populaţii fiind 2.25$). Să se testeze dacă salariul mediu săptămânal la cele doua firme diferă semnificativ (1% nivel de semnificaţie).

III.3 Un raport recent arată ca absolvenţii de universitate fără diplomă se căsătoresc mai repede decât cei cu diplomă. Sunt aleşi câte 100 de indivizi din cele două populaţii; pentru aceste două eşantioaneabsolvenţii fără diplomă se căsătoresc în medie la 22.8 ani (deviaţia standerd cunoscută populaţiei fiind $\sigma_1 = 1.3$ ani) iar cei cu diplomă la 23.3 ani (deviaţia cunoscută a populaţiei este $\sigma_2 = 1.9$ ani).

Cu 1% nivel de semnificaţie se poate trage concluzia că raportul este corect?

III.4 O firma producătoare de pastă de dinţi doreşte să ştie dacă introducerea unui nou compus chimic măreşte performanţele produsului său. Sunt alese doua eşantioane fiecare a câte 100 de indivizi. Pentru primul grup (care foloseste noul produs) se numără în medie 3 carii (deviaţia standard cunoscută este 0.6), pentru cel de-al doilea grup se numără în medie 3.5 carii (deviaţia standard este de 0.4).

Cu 1% nivel de semnificaţie se poate trage concluzia că prin introducerea noului compus pasta de dinţi îşi îmbunătăţeşte performanţele? Dar cu 5%?

III.5 Pentru două eşantioane provenind din două populaţii statistice distincte A şi B se determină următoarele valori

$$(A) \quad \sigma_1 = 0.75, n_1 = 155, \overline{x}_{n_1} = 15$$

(B) $\sigma_2 = 0.78, n_2 = 150, \overline{x}_{n_2} = 14.5$

Să se testeze diferenţa mediilor celor două populaţii cu 1% si cu 5% nivel de semnificaţie.