

## Laborator 4 - Statistică inferențială

### I. Intervale de încredere pentru media unei populații cu dispersia necunoscută

Se consideră o populație căreia nu i se cunoaște dispersia. În acest caz se folosește drept estimator al deviației standard  $\sigma$ , deviația standard a eșantionului  $s$ . În acest caz, scorul  $t = \frac{\bar{x}_n - \mu}{s/\sqrt{n}}$  este distribuit Student cu  $n - 1$  grade de libertate:  $t(n - 1)$ .

Se caută un interval în care media populației  $\mu$ , necunoscută și ea, să se găsească cu probabilitate prescrisă (0.9, 0.95 sau 0.99). Un astfel de interval este următorul:

$$\left( \bar{x}_n - t^* \cdot \frac{s}{\sqrt{n}}, \bar{x}_n + t^* \cdot \frac{s}{\sqrt{n}} \right)$$

unde  $t^*$ , numit valoarea critică, se determină astfel

$$t^* = -qt(\alpha/2, n - 1) = qt(1 - \alpha/2, n - 1)$$

$\alpha$  este egal cu  $1 -$  nivelul de încredere, iar  $s$  este deviația standard a eșantionului. În cazul în care sunt cunoscute valorile din eșantion,  $\bar{x}_n$  și  $s$  se calculează astfel:

$$\bar{x}_n = \text{mean}(\text{date-eșantion}), s = \text{sd}(\text{date-eșantion})$$

În calculele de mai jos vom folosi un estimator pentru eroarea standard a mediei, anume  $se = \frac{s}{\sqrt{n}}$ .

**Exercițiu rezolvat.** O companie ce produce jucării dorește să afle cât de interesante sunt produsele sale. 60 de copii dintr-un eșantion sunt rugați să răspundă cu o valoare între 0 și 5 și se determină o medie egală cu 3.3, cu o deviație standard  $s = 0.4$ . Cât de interesante, în medie, sunt jucăriile companiei (95% nivel de încredere)?

```
> alfa = 0.05
> sample_mean = 3.3
> n = 60
> s = 0.4
> se = s/sqrt(n)
> critical.t = qt(1 - alfa/2, n - 1)
> a = sample_mean - critical.t*sigma/sqrt(n)
> b = sample_mean + critical.t*sigma/sqrt(n)
> interval = c(a, b)
> interval
```

Rezultatul este intervalul [3.19667, 3.40333].

### Exerciții propuse

- I.1 Scrieți într-un script o funcție (numită **t\_conf\_interval**) care să calculeze intervalul de încredere ca mai sus (parametrii funcției vor fi:  $n$ ,  $\bar{x}_n$ ,  $\alpha$  etc). Funcția aceasta va fi utilizată la rezolvarea exercițiilor de mai jos.
- I.2 196 de studenți aleși aleator au fost întrebați cât de mulți bani au investit în cumpărături online săptămâna trecută. Media a fost calculată la 44.65\$, cu o dispersie (a eșantionului) egală cu  $s^2 = 2.25$ . Calculați un interval de încredere de 99% pentru media populației (despre care se presupune că urmează o lege normală).
- I.3 O companie de dulciuri consideră că nivelul de zahăr în produsele sale poate avea valori între 1 și 20, urmând o lege normală. Se consideră un eșantion de 49 de produse. Media nivelului de zahăr este 12 iar deviația standard a eșantionului este de 1.75.

- (a) Determinați intervalele de încredere de 99% și 95% pentru media nivelului de zahăr.
- (b) După modificarea rețetei, s-au testat 49 produse și s-a găsit că media nivelului de zahăr este de 13.5 cu o deviație standard de 1.25. Determinați un interval de încredere de 95% pentru media nivelului de zahăr.

- I.4 Modificați funcția de mai sus pentru cazul când eșantionul este dat într-un fișier (trebuie calculată media de selecție, deviația standard și dimensiunea eșantionului). Aplicați funcția astfel modificată pentru a rezolva și următorul exercițiu.
- I.5 Pentru un eșantion aleator simplu dintr-o populație normală cu dispersia necunoscută se măsoară următoarele valori:

12 11 12 10 11 12 13 12 11 11 13 14 10

Să se determine, utilizând aceste date, intervalele de încredere de 90%, 95% și 99% pentru media populației.

## II. Testarea ipotezelor statistice

Avem o populație statistică a căreia nu i se cunoaște complet distribuția. Un test statistic asupra unor aspecte ale distribuției<sup>1</sup> populației urmează următoarea procedură generală:

- se formulează o ipoteză, numită ipoteza nulă  $H_0$ , care precizează complet distribuția populației.
- ipoteza nulă este "atacată" de o ipoteză alternativă  $H_a$ , care susține o presupunere diferită asupra distribuției populației.
- în cazul în care există dovezi suficiente (statistic semnificative) **ipoteza nulă,  $H_0$ , este respinsă și se acceptă ipoteza alternativă  $H_a$ .**
- dacă dovezile împotriva ipotezei nule nu sunt statistic semnificative, atunci **ipoteza nulă  $H_0$  nu poate fi respinsă, (un test statistic nu se termină prin acceptarea ipotezei nule).**

La efectuarea unui test statistic se pot face două tipuri de erori:

- **eroare de tipul I:** rezultatul testului impune respingerea ipotezei nule  $H_0$ , deși, în realitate, ea este adevărată - această eroare este cauzată de o încredere excesivă.
- **eroare de tipul II:** rezultatul testului nu cere respingerea ipotezei nule  $H_0$ , deși, în realitate, ea este nu adevărată - această eroare este cauzată de un scepticism accentuat.

	$H_0$ nu este respinsă	$H_0$ este respinsă
$H_0$ este adevărată	corect	eroare de tip I
$H_0$ este falsă	eroare de tip II	corect

## II. Testul $z$ asupra proporțiilor

---

<sup>1</sup>De exemplu asupra mediei sau dispersiei.

Se consideră o variabilă  $X$  ce numără succesele din  $n$  încercări.  $X$  este distribuită binomial -  $X : B(n, p)$ . Testul proporțiilor inferează asupra probabilității  $p$ . Se notează cu  $p' = \frac{X}{n}$  frecvența dată de eșantion. Deoarece  $M(X) = np$  și  $D^2(X) = np(1 - p)$ , vom avea

$$M(p') = p \text{ și } D^2(p') = \frac{p(1 - p)}{n}.$$

Pentru  $n$  suficient de mare ( $n \geq 20$  și  $np \geq 5$ )  $p'$  urmează aproximativ o distribuție normală. Statistica  $z = \frac{p' - p}{\sqrt{p(1 - p)/n}}$  este distribuită normal standard:  $N(0, 1)$ .

Testul asupra proporțiilor decurge astfel:

1. se formulează ipoteza nulă, care susține că probabilitatea  $p$  ia o valoare particulară:

$$H_0 : p = p_0$$

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

$$H_a : p < p_0 \quad (\text{ipoteză asimetrică la stânga}) \text{ sau}$$

$$H_a : p > p_0 \quad (\text{ipoteză asimetrică la dreapta}) \text{ sau}$$

$$H_a : p \neq p_0 \quad (\text{ipoteză simetrică})$$

3. se fixează nivelul de semnificație:  $\alpha$  (care uzual poate fi 1% sau 5%);

4. se calculează scorul testului:

$$z = \frac{p' - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

5. se determină valoarea critică  $z^*$ :

$$z^* = qnorm(\alpha, 0, 1) \quad \text{pentru ipoteză } H_a \text{ asimetrică la stânga } (z^* < 0),$$

$$z^* = qnorm(1 - \alpha, 0, 1) \quad \text{pentru ipoteză } H_a \text{ asimetrică la dreapta } (z^* > 0),$$

$$z^* = -qnorm(\alpha/2, 0, 1) = qnorm(1 - \alpha/2, 0, 1) \quad \text{pentru ipoteză } H_a \text{ simetrică } (z^* > 0).$$

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

$$z < z^* \quad \text{pentru ipoteză } H_a \text{ asimetrică la stânga sau}$$

$$z > z^* \quad \text{pentru ipoteză } H_a \text{ asimetrică la dreapta sau}$$

$$|z| > |z^*| \quad \text{pentru ipoteză } H_a \text{ simetrică,}$$

altfel vom spune că **nu există suficiente dovezi pentru a respinge ipoteza nulă  $H_0$  și a accepta ipoteza alternativă  $H_a$ .**

**Exercițiu rezolvat.** Un politician susține ca va primi mai puțin de 60% dintre voturi în colegiul său. Un eșantion dintr-o 100 de alegători arată că 63 dintre ei au votat pentru acest politician. Putem respinge afirmația politicianului? (1% nivel de semnificație)

```
> alfa = 0.01
> n = 100
> succese = 63
> p_prim = succese/n
> p0 = 0.6
> z_score = (p_prim - p0)/sqrt(p0(1 - p0)/n)
> critical_z = qnorm(1 - alfa, 0, 1)
> z_score
> critical_z
```

Rezultatul este  $z = 0.61237 < z^* = 2.32634$ , deci ipoteza nulă nu se poate respinge.

### Exerciții propuse

- II.1 Scrieți într-un script o funcție (numită **test\_proportion**) care să calculeze și să returneze valoarea critică și scorul testului proporțiilor (parametrii vor fi  $\alpha$ ,  $n$ , numărul de succese,  $p_0$ ). Funcția aceasta va fi utilizată, pentru rezolvarea exercițiilor de mai jos.
- II.2 Se presupune că dintr-un număr mare de componente, 10% sunt defecte. Se testează dacă procentul defectelor a crescut. Se testează în acest sens 150 de componente și se determină că 20 dintre ele sunt defecte. Se poate afirma cu nivel de semnificație de 5% că procentul componentelor defecte este mai mare decât 10%?
- II.3 O agenție imobiliară își schimbă managementul deoarece 10% dintre clienți declară ca sunt nemulțumiți de serviciile oferite. După schimbarea managementului, dintr-un eșantion de 110 clienți, 15 sunt nemulțumiți. Se poate trage concluzia că schimbarea a fost inutilă?
- II.4 Datele istorice arată că 2.5% dintre pacienții care supraviețuiesc unei proceduri pe cord au un IQ sub 70 de puncte. Dintr-un eșantion de 128 de pacienți care au trecut prin această procedură, 10 au un IQ sub 70. Se poate trage concluzia că procentul de mai sus a crescut în timp?
- II.5 Să se testeze o ipoteza adecvată pentru datele de mai jos. Proporția este numărul de purici înlăturați (killed fleas) supra numărul total de purici (fleas).

My dog has so many fleas,  
They do not come off with ease.  
As for shampoo, I have tried many types  
Even one called Bubble Hype,  
Which only killed 25% of the fleas,  
Unfortunately I was not pleased.

I've used all kinds of soap,  
Until I had give up hope  
Until one day I saw  
An ad that put me in awe.

A shampoo used for dogs  
Called GOOD ENOUGH to Clean a Hog  
Guaranteed to kill more fleas.

I gave Fido a bath  
And after doing the math  
His number of fleas  
Started dropping by 3's!

Before his shampoo  
I counted 42.  
At the end of his bath,  
I redid the math  
And the new shampoo had killed 17 fleas.  
So now I was pleased.

Now it is time for you to have some fun  
With the level of significance being 0.01,  
You must help me figure out  
Use the new shampoo or go without?