Laborator 4 - Statistică inferenţială

I. Intervale de încredere pentru media unei populaţii cu dispersia necunoscută

Se consideră o populaţie căreia nu i se cunoaşte dispersia. În acest caz se foloseşte drept estimator al deviaţiei standard σ, deviaţia standard a eşantionului s. În acest caz, scorul $t = \frac{\overline{x}_n - \mu}{s/\sqrt{n}}$ este distribuit Student cu $n-1$ grade de libertate: $t(n-1)$.

Se caută un interval în care media populaţiei μ, necunoscută şi ea, să se găsească cu probabilitate prescrisă (0.9, 0.95 sau 0.99). Un astfel de interval este următorul:

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

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

$$\boxed{t^* = -qt\left(\alpha/2, n-1\right) = qt\left(1-\alpha/2, n-1\right)}$$

α este egal cu 1− nivelul de încredere, iar s este deviaţia standard a eşantionului. În cazul în care sunt cunoscute valorile din eşantion, $\overline{x}_n$ şi s se calculează astfel:

$$\boxed{\overline{x}_n = mean(\text{date-eşantion}),\ s = sd(\text{date-eşantion})}$$

În calculele de mai jos vom folosi *eroarea standard a mediei* $se = \frac{s}{\sqrt{n}}$.

Exerciţiu rezolvat. O companie ce produce jucării doreşte să afle cât de interesante sunt produsele sale. 60 de copiii dintr-un eşantion sunt rugaţi să răspundă cu o valoare între 0 şi 5 şi se determină o medie egală cu 3.3, cu o deviatie 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 $[3.19667, 3.40333]$.

Exerciţii propuse

I.1 Scrieţi într-un script o funcţie (numită **t_conf_interval**) care să calculeze intervalul de în credere ca mai sus (parametrii funcţiei vor fi: n, $\overline{x}_n$, α etc). Funcţia aceasta va fi utilizată la rezolvarea exerciţiilor de mai jos.

I.2 196 de studenţi aleşi aleator au fost întrebaţi cât de mulţi bani au investit în cumpărături online săptămâna trecută. Media a fost calculată la 44.65$, cu o dispersie (a eşantionului) egală cu $s^2 = 2.25$. Calculaţi un interval de încredere de 99%pentru media populaţiei.

I.3 O companie de dulciuri consideră că nivelul de zahăr în produsele sale poate avea valori între 1 şi 20. Se consideră un eşantion de 49 de produse. Media nivelului de zahăr este 12 iar deviaţia standard a eşantionului este de 1.75.

(a) Determinaţi intervalele de încredere de 99% si 95% pentru media nivelului de zahăr.

(b) Dupa modificarea reţetei, s-au testat 49 produse şi s-a găsit că media nivelului de zahar este de 13.5 cu o deviaţie standard de 1.25. Determinaţi un interval deîincredere de 95% pentru media nivelului de zahăr.

I.4 Modificaţi funcţia de mai sus pentru cazul când eşantionul este dat într-un fişier (trebuie calculată media de selecţie şi dimensiunea eşantionului). Aplicaţi funcţia astfel modificată pentru a rezolva şi următorul exerciţiu.

I.5 Pentru un eşantion aleator simplu dintr-o populaţie 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% şi 99% pentru media populaţiei.

II. Testarea ipotezelor statistice

Avem o populaţie statistică căreia nu i se cunoaşte distribuţia. Un test statistic asupra distribuţiei[1] populaţiei urmează următoarea procedură generală:

- se formulează o ipoteză, numită ipoteza nulă H_0, care precizează complet distribuţia populaţiei.

- ipoteza nulă este "atacată" de o ipoteză alternativă H_a, care susţine o presupunere diferită asupra distribuţiei populaţiei.

- în cazul în care există dovezi suficiente (statistic semnificative) **ipoteza nulă**, H_0, **este respinsă şi 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, deşi, 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, deşi, 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_a este adevărată	eroare de tip II	corect

II. Testul z asupra proporţiilor

[1] 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 proporţiilor inferează asupra probabilităţii p. Se notează cu $p' = \dfrac{X}{n}$ frecvenţa dată de eşantion. Deoarece $M(X) = np$ şi $D^2(X) = np(1-p)$, vom avea

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

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

Testul asupra proporţiilor decurge astfel:

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

$$\boxed{H_0 : \ p = p_0}$$

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

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

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

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

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

4. se calculează scorul testului:

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

5. se determină valoarea critică z^*:

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

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

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

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

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

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

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

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

Exerciţiu rezolvat. Un politician susţine ca va primi mai puţin de 60% dintre voturi în colegiul său. Un eşantion dintr-o 100 de alegători arată că 63 dintre ei au votat pentru acest politician. Putem respinge afirmaţia politicianului? (1% nivel de semnificaţie)

```
> 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.

Exerciţii propuse

II.1 Scrieţi într-un script o funcţie (numită **test_proportion**) care să calculeze şi să returneze valoarea critică şi scorul testului proporţiilor (parametrii vor fi α, n, numărul de succese, p_0). Funcţia aceasta va fi utilizată, pentru rezolvarea exerciţiilor 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 şi se determină că 20 dintre ele sunt defecte. Se poate afirma cu nivel de semnificaţie de 5% că procentul componentelor defecte este mai mare decât 10%?

II.3 O agenţie imobiliară îşi schimbă managementul deoarece 10% dintre clienţi declară ca sunt nemulţumiţi de serviciile oferite. După schimbarea managementului, dintr-un eşantion de 110 clienţi, 15 sunt nemulţumiţi. Se poate trage concluzia că schimbarea a fost inutilă?

II.4 Datele istorice arată că 2.5% dintre pacienţii care supravieţuiesc unei proceduri pe cord (Fontan) au un IQ sub 70 de puncte. Dintr-un eşantion de 128 de pacienţi 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. Proporţia este numărul de purici înlăturaţi (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?