

Laborator 5 - Statistică inferențială

O populație statistică este o mulțime de indivizi¹ al căror atribut (greutate, înălțime etc) este supus unor variații aleatoare. Statistica inferențială are drept scop determinarea cu un anumit grad de acuratețe (aproximarea, în cele mai multe cazuri) a parametrilor unei populații statistice (cum ar fi medie sau deviație standard). Inferența asupra parametrilor populației se realizează astfel:

- se alege un eșantion aleator simplu (alegerea indivizilor se face în mod independent și fiecare individ are aceeași probabilitate de a fi ales);
- se calculează una sau mai multe statistici utilizând eșantionul;
- utilizând statistica matematică și teoria probabilităților, cu ajutorul statisticilor calculate, se formulează o afirmație (se inferează) asupra unui parametru al populației.

I. Legea normală, reprezentare grafică

RStudio. Nu uitați să va setați directorul de lucru: **Session** → **Set Working Directory** → **Choose Directory**.

O variabilă aleatoare normală cu parametrii μ și σ^2 are următoarea funcție de densitate

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right].$$

Dacă $X : N(\mu, \sigma^2)$, atunci

$$M(X) = \mu \text{ și } D^2(X) = \sigma^2$$

Distribuția $N(0, 1)$ se numește *normală standard*. Valorile unei variabile distribuite normal au următoarea împărțire:

%68 se găsesc la cel mult o deviație standard față de medie;

%95 se găsesc la cel mult două deviații standard față de medie;

%99.7 se găsesc la cel mult trei deviații standard față de medie;

Exercițiu rezolvat. Reprezentarea grafică a funcției de densitate normale standard ($\mu = 0$, $\sigma = 1$) se face din linia de comandă astfel

```
> t = seq(-6, 6, length = 400)
> f = 1/sqrt(2*pi)*exp(-t^2/2)
> plot(t, f, type = "l", lwd = 1)
```

Acest rezultat se poate transforma într-o funcție care se va scrie într-un *script R* astfel: **File** → **New File** → **R Script** și în fereastra de editare se scrie următorul cod

```
normal_density <- function(limit) {
  t = seq(-limit, limit, length = 400)
  f = 1/sqrt(2*pi)*exp(-t^2/2)
  plot(t, f, type = "l", lwd = 1)
}
normal_density(6)
```

¹În sens larg.

RStudio. După editare, scriptul este salvat (**Ctrl+S**) cu un nume de tipul "my_script.R" și este încărcat cu **Code** → **Source File** (**Ctrl+Shift+O**) sau din linia de comandă cu **source(script_file)**

RStudio. O dată încărcat scriptul, o funcție care face parte din acest script se poate executa din linia de comandă: **normal_density(8)** sau din fereastra de editare astfel: se selectează liniile dorite a fi executate și **Ctrl+Enter**, iar scriptul în întregime se execută cu **Ctrl+Alt+R**.

Exercițiu propus

I.1 Scrieți o funcție care să reprezinte grafic densitatea legii normale $N(\mu, \sigma^2)$.

II. Estimarea mediei unei populații: Media de selecție

Considerăm o populație cu media μ și dispersia σ^2 , căreia i se măsoară atributul² X . Din această populație se extrage un eșantion aleator simplu de dimensiune n : X_1, X_2, \dots, X_n . Aceste valori pot fi privite și ca variabile aleatoare independente și identic repartizate cu variabila X . Media de selecție se definește astfel:

$$\bar{x}_n = \frac{X_1 + X_2 + \dots + X_n}{n}$$

și este o statistică, dar în același timp, pentru un eșantion generic, poate fi văzută ca o variabilă aleatoare. Proprietăți ale mediei de selecție:

- \bar{x}_n este un estimator (nedeplasat) al mediei populației, μ , din care provine eșantionul.
- privită ca variabilă aleatoare:

$$M(\bar{x}_n) = \mu, D^2(\bar{x}_n) = \frac{\sigma^2}{n}$$

- dacă populația din care provine eșantionul este distribuită normal $N(\mu, \sigma^2)$, atunci media de selecție urmează o distribuție normală $N\left(\mu, \frac{\sigma^2}{n}\right)$;
- dacă dimensiunea eșantionului este suficient de mare ($n \geq 30$), atunci media de selecție urmează cu aproximație o distribuție normală $N\left(\mu, \frac{\sigma^2}{n}\right)$.

O funcție pentru determinarea mediei de selecție a unui eșantion dat într-un fișier.

```
selection_mean <- function(filename) {  
  x = scan(filename);  
  m = mean(x)  
}  
selection_mean("sample.txt")
```

RStudio. Fișierul cu numele *filename* trebuie să fie în directorul de lucru.

Exercițiu propus

II.1 Scrieți într-un script funcția descrisă mai sus și aplicați-o fișierului *history.txt*.

² X este o variabilă aleatoare cu media μ și dispersia σ^2 .

III. Intervale de încredere pentru media unei populații cu dispersia cunoscută

Se consideră o populație cu dispersia cunoscută σ^2 . Se caută un interval în care media μ , necunoscută a populație să se găsească cu probabilitate mare (0.90, 0.95 sau 0.99). Un astfel de interval este următorul:

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

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

$$z^* = -qnorm(\alpha/2, mean = 0, sd = 1) = qnorm(1 - \alpha/2, mean = 0, sd = 1)$$

iar α este egal cu $1 - \text{nivelul de încredere}$. Media de selecție, dacă nu este dată, se poate calcula astfel:

$$\bar{x}_n = \text{mean}(\text{date_eșantion})$$

Exercițiu rezolvat. Durata vieții unui tip de baterie urmează cu aproximație o lege normală cu dispersia de 9 ore. Pentru un eșantion de 100 de baterii se măsoară o medie de viață de 20 de ore. Să se determine un interval de încredere de 90% pentru media de viață a întregii populații.

```
> alfa = 0.1
> sample_mean = 20
> n = 100
> sigma = sqrt(9)
> critical_z = qnorm(1 - alfa/2, 0, 1)
> a = sample_mean - critical_z*sigma/sqrt(n)
> b = sample_mean + critical_z*sigma/sqrt(n)
> interval = c(a, b)
> interval
```

Rezultatul este intervalul [19.50654, 20.49346].

Exerciții propuse (III.1 și încă două dintre III.2-III.6)

- III.1 Scrieți într-un script o funcție (numită **zconfidence.interval**) care să calculeze intervalul de încredere ca mai sus (parametrii funcției vor fi: n , \bar{x}_n , α etc). Funcția aceasta va fi utilizată la rezolvarea exercițiilor de mai jos.
- III.2 Se caută un interval de încredere de 90% pentru media unei populații normale cu dispersia cunoscută $\sigma^2 = 100$. Pentru aceasta se utilizează un eșantion aleator simplu de 25 de indivizi a cărui medie de selecție (calculată) este 67.53.
- III.3 Într-o instituție publică există un automat de cafea reglat în așa fel încât cantitatea de cafea dintr-un pahar urmează o lege normală cu deviația standard $\sigma = 0.5$ oz. Pentru un eșantion de $n = 50$ de pahare ales la întâmplare, se măsoară o medie a greutății pentru un pahar de 5 oz. Să se determine un interval de încredere de 95% pentru media de greutate a unui pahar de cafea.
- III.4 Într-o încercare disperată de a concura General Electric, compania ACME introduce un nou tip de becuri. ACME fabrică inițial 100 de becuri a căror medie de viață măsurată este 1280 de ore (deviația standard a populației este 140 de ore). Să se găsească un interval de încredere de 99% pentru media de viață a becurilor.

- III.5 Se măsoară greutatea pentru un eșantion de 35 de atleți și se găsește o medie de 60 kg. Se presupune că deviația standard a populației este 5 kg. Să se determine intervalele de încredere de 90%, 95% respectiv 99% pentru media populației. Intervalul de 95% încredere este mai mare sau mai mic decât cel de 99%? De ce?
- III.6 Modificați funcția de mai sus pentru cazul când eșantionul este dat într-un fișier (trebuie calculată media de selecție și dimensiunea eșantionului). Aplicați funcția astfel modificată fișierului construit la exercițiul II.2 pentru a determina un interval de încredere de 95%.

IV. 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 σ , 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 μ , 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

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

α 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 (IV.1 și încă două dintre IV.2-IV.5)

- IV.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 , α etc). Funcția aceasta va fi utilizată la rezolvarea exercițiilor de mai jos.
- IV.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).
- IV.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.
- Determinați intervalele de încredere de 99% și 95% pentru media nivelului de zahăr.
 - Dupa 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.
- IV.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.
- IV.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.

Testarea ipotezelor statistice

Avem o populație statistică căreia nu i se cunoaște complet distribuția. Un test statistic asupra unor aspecte ale distribuției³ 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.

³De exemplu asupra mediei sau dispersiei.

- **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

V. Testul z asupra proporțiilor

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: α (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

- V.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 α , n , numărul de succese, p_0). Funcția aceasta va fi utilizată, pentru rezolvarea exercițiilor de mai jos.
- V.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%?
- V.3 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?

Temă pentru acasă.

4 puncte [2p: D1 sau D2] + [2p: D3 sau D4]

D1. (2 puncte) Nivelul de sodiu al unei persoane este măsurat independent de opt ori (aceste măsurători urmează o distribuție normală cu deviația standard $\sigma = 10$). Media acestor măsurători este 140 (mEq/L); determinați un interval de încredere de 95% pentru nivelul mediu de sodiu.

D2. (2 puncte) Determinați un interval de încredere de 99% pentru media unei populații distribuite normal, folosind un eșantion aleator simplu format din 196 de indivizi cu o medie de selecție egală cu 35 și cu dispersia eșantionului $s^2 = 16$.

D3. (2 puncte) O companie își schimbă managementul de PR deoarece 30% dintre clienți declară ca sunt nemulțumiți de relațiile cu publicul oferite. După schimbarea managementului, dintr-un eșantion de 100 clienți, 32 sunt nemulțumiți. Se poate trage concluzia că schimbarea a fost inutilă? (1% și 5% nivel de semnificație.)

D4. (2 puncte) Biroul de recensământ al SUA (US Census Bureau) raportează că proporția studentelor din colegii este 50.9%. La un anumit colegiu 48 dintre cei 100 de studenți dintr-un eșantion aleator simplu sunt fete. Se poate trage concluzia că la acest colegiu proporția este mai mică decât la nivel național? (Cu 1% și 5% nivel de semnificație.)

Rezolvările acestor exerciții (funcțiile R și apelurile lor) vor fi redactate într-un script R.