Laborator 5 - Statistică inferenţială

O populaţie statistică este o mulţime de indivizi[1] al căror atribut (greutate, înalţime etc) este supus unor variaţii aleatoare. Statistica inferenţială are drept scop determinarea cu un anumit grad de acurateţe (aproximarea, în cele mai multe cazuri) a parametrilor unei populaţii statistice (cum ar fi medie sau deviaţie standard). Inferenţa asupra parametrilor populaţiei se realizează astfel:

- se alege un eşantion aleator simplu (alegerea indivizilor se face în mod independent şi fiecare individ are aceeaşi probabilitate de a fi ales);
- se calculează una sau mai multe statistici utilizând eşantionul;
- utilizând statistica matematică şi teoria probabilităţilor, cu ajutorul statisticilor calculate, se formuleaza o afirmaţie (se inferează) asupra unui parametru al populaţiei.

I. Legea normală, reprezentare grafică

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

O variabilă aleatoare normală cu parametrii μ şi σ^2 are următoarea funcţie 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

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

Distribuţia $N(0, 1)$ se numeşte *normală standard*. Valorile unei variabile distribuite normal au următoarea împrăştiere:

%68 se găsesc la cel mult o deviaţie standard faţă de medie;

%95 se găsesc la cel mult două deviaţii standard faţă de medie;

%99.7 se găsesc la cel mult trei deviaţii standard faţă de medie;

Exerciţiu rezolvat. Reprezentarea grafică a funcţiei 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 funcţie care se va scrie într-un *script R* astfel: **File → New File → R Script** şi în ferestra 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)
```

[1] În sens larg.

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

RStudio. O dată încărcat scriptul, o funcţie 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 şi **Ctrl+Enter**, iar scriptul în întregime se execută cu **Ctrl+Alt+R**.

Exerciţiu propus

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

II. Estimarea mediei unei populaţii: Media de selecţie

Considerăm o populaţie cu media μ şi dispersia σ^2, căreia i se măsoară atributul[2] X. Din această populaţie se extrage un eşantion aleator simplu de dimensiune n: $X_1, X_2, \ldots, X_n$. Aceste valori pot fi privite şi ca variabile aleatoare independente şi identic repartizate cu variabila X. Media de selecţie se defineşte astfel:

$$\overline{x}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

şi este o statistică, dar în acelaşi timp, pentru un eşantion generic, poate fi văzută ca o variabilă aleatoare. Proprietăţi ale mediei de selecţie:

- $\overline{x}_n$ este un estimator (nedeplasat) al mediei populaţiei, μ, din care provine eşantionul.
- privită ca variabilă aleatoare:

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

- dacă populaţia din care provine eşationul este distribuită normal $N(\mu, \sigma^2)$, atunci media de selecţie urmează o distribuţie normală $N\left(\mu, \frac{\sigma^2}{n}\right)$;
- dacă dimensiunea eşantionului este suficient de mare ($n \geqslant 30$), atunci media de selecţie urmează cu aproximaţie o distribuţie normală $N\left(\mu, \frac{\sigma^2}{n}\right)$.

O funcţie pentru determinarea mediei de selecţie a unui eşantion dat într-un fişier.

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

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

Exerciţiu propus

II.1 Scrieţi într-un script funcţia descrisă mai sus şi aplicaţi-o fişierului *history.txt*.

[2] X este o variabilă aleatoare cu media μ şi dispersia σ^2.

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

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

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

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

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

iar α este egal cu 1− nivelul de încredere. Media de selecţie, dacă nu este dată, se poate calcula astfel:

$$\boxed{\overline{x}_n = mean(\text{date\_eşantion})}$$

Exerciţiu rezolvat. Durata vieţii unui tip de baterie urmează cu aproximaţie o lege normală cu dispersia de 9 ore. Pentru un eşantion de 100 de baterii se măsoară o medie de viaţa de 20 de ore. Să se determine un interval de încredere de 90% pentru media de viaţă a întregii populaţii.

```
> 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 $[19.50654, 20.49346]$.

Exerciţii propuse (III.1 şi încă două dintre III.2-III.6)

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

III.2 Se caută un interval de încredere de 90% pentru media unei populaţii normale cu dispersia cunoscută $\sigma^2 = 100$. Pentru aceasta se utilizează un eşantion aleator simplu de 25 de indivizi a cărui medie de selecţie (calculată) este 67.53.

III.3 Într-o instituţie publică există un automat de cafea reglat în aşa fel încât cantitatea de cafea dintr-un pahar urmează o lege normală cu deviaţia standard $\sigma = 0.5$ oz. Pentru un eşantion de $n = 50$ de pahare ales la întâmplare, se măsoară o medie a greutăţii 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 iniţial 100 de becuri a căror medie de viaţă măsurată este 1280 de ore (deviaţia standard a populaţiei este 140 de ore). Să se găsească un interval de încredere de 99% pentru media de viaţă a becurilor.

III.5 Se măsoară greutatea pentru un eşantion de 35 de atleţi şi se găseşte o medie de 60 kg. Se presupune că deviaţia standard a populaţiei este 5 kg. Să se determine intervalele de încredere de 90%, 95% respectiv 99% pentru media populaţiei. Intervalul de 95% încredere este mai mare sau mai mic decât cel de 99%? De ce?

III.6 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 fişierului construit la exerciţiul II.2 pentru a determina un interval de încredere de de 95%.

IV. 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 $t = \dfrac{\overline{x}_n - \mu}{s/\sqrt{n}}$ 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 un estimator pentru *eroarea standard a mediei*, anume $se = \dfrac{s}{\sqrt{n}}$.

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 (IV.1 şi încă două dintre IV.2-IV.5)

IV.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.

IV.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 (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 şi 20, urmÂnd o lege normală. 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.

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

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

Testarea ipotezelor statistice

Avem o populaţie statistică căreia nu i se cunoaşte complet distribuţia. Un test statistic asupra unor aspecte ale distribuţiei[3] 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.

[3] De exemplu asupra mediei sau dispersiei.

- **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_0 este falsă	eroare de tip II	corect

V. Testul z asupra proporţiilor

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

V.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.

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 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%?

V.3 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?

Temă pentru acasă.

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

D1. (2 puncte) Nivelul de potasiu al unei persoane este măsurat independent de opt ori (aceste măsurători urmează o distribuţire normală cu deviaţia standard $\sigma = 0.5$). Media acestor măsurători este 2.75; determinaţi un interval de în credere de 99% pentru nivelul mediu de potasiu.

D2. (2 puncte) Determinaţi un interval de încredere de 90% pentru media unei populaţii distribuite normal, folosind un eşantion aleator simplu format din 144 de indivizi cu o medie de selecţie egală cu 20 şi cu dispersia $s^2 = 18$.

D3. (2 puncte) O agenţie imobiliară îşi schimbă managementul deoarece 10% dintre clienţi declară ca sunt nemulţumiţi de serviciile oferite. După schimbarea managementului, dintr-un eşantion de 112 clienţi, 14 sunt nemulţumiţi. Se poate trage concluzia că schimbarea a fost inutilă? (1% şi 5% nivel de semnificaţie.)

D2. (2 puncte) Datele istorice arată că 2.5% dintre pacienţii care supravieţuiesc unei proceduri pe cord 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?

Rezolvările acestor exerciţii (funcţiile R şi apelurile lor) vor fi redactate într-un script R.