Laborator 3 - 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**.

Exerciţii propuse

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

I.2 Scrieţi o funcţie care să determine pentru legea normală intervalele centrate în medie şi de lungime egală cu două deviaţii standard, patru deviaţii standard respectiv şase deviaţii standard.

I.3 Aplicaţi funcţiile de mai sus următoarele legi normale: $N(0,4)$, $N(2,5)$, $N(1,9)$;

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")
```

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

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

Exerciţii propuse

II.1 Scrieţi într-un script funcţia descrisă mai sus.

II.2 Creaţi un fişier care să conţină următorul eşantion (care provine dintr-o populaţie cu dispersia $\sigma^2 = 2$) şi apoi determinaţi-i media de selecţie folosind funcţia anterioară.

33 47 56 34 24 42 51 34 36 49 28 55 32 37

II.3 Modificaţi funcţia de mai sus pentru a determina şi deviaţia standard a mediei de selecţie (dipsersia populaţiei va fi un parametru al funcţiei). Aplicaţi funcţia astfel modificată fisierului de mai sus.

III. Intervale de în credere 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 = 100
> n = 20
> 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 Scrieţi într-un script o funcţie (numită **zconfidence_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.

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 cu o deviaţie standard de 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%.