

Laborator 3 - 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

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

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**.

Exerciții propuse

- I.1 Scrieți o funcție care să reprezinte grafic densitatea legii normale $N(\mu, \sigma^2)$.
- I.2 Scrieți o funcție care să determine pentru legea normală intervalele centrate în medie și de lungime egală cu două deviații standard, patru deviații standard respectiv șase deviații standard.
- I.3 Aplicați funcțiile de mai sus următoarele legi normale: $N(0, 4)$, $N(2, 5)$, $N(1, 9)$;

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")
```

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

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

Exerciții propuse

II.1 Scrieți într-un script funcția descrisă mai sus.

II.2 Creați un fișier care să conțină următorul eșantion (care provine dintr-o populație cu dispersia $\sigma^2 = 2$) și apoi determinați-i media de selecție folosind funcția anterioară.

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

II.3 Modificați funcția de mai sus pentru a determina și deviația standard a mediei de selecție (dispersia populației va fi un parametru al funcției). Aplicați funcția astfel modificată fișierului de mai sus.

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 $1 -$ nivelul de încredere. Media de selecție, dacă nu este dată, se poate calcula astfel:

$$\bar{x}_n = mean(date_eșantion)$$

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 = 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].

Exerciții propuse

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 greutateii 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 cu o deviație standard de 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%.