Laborarator 13 – Probabilități și Statistică Matematică

STATISTICA INFERENTIALA (1)

O populatie statistica este o multime de indivizi al caror atribut (greutate, inaltime etc) este supus unor variatii aleatoare. Statistica inferentiala are drept scop determinarea cu un anumit grad de acuratete (aproximarea, in cele mai multe cazuri) a parametrilor unei populatii statistice (cum ar fi medie sau deviatie standard). Inferenta asupra parametrilor populatiei se realizeaza astfel:

- se alege un esantion aleator simplu (alegerea indivizilor se face in mod independent si fiecare individ are aceeasi probabilitate de a fi ales);

- se calculeaza una sau mai multe statistici utilizand esantionul;

- utilizand statistica matematica si teoria probabilitatilor, cu ajutorul statisticilor calculate, se formuleaza o afirmatie (se infereaza) asupra unui parametru al populatiei.

I. Legea normala, reprezentare grafica

RStudio. Nu uitati sa va setati directorul de lucru: Session → Set Working Directory → Choose Directory.

<u>Exercitiu rezolvat</u>: *Reprezentarea grafica a functiei de densitate normale standard* ($\mu = 0, \sigma = 1$) se face din linia de comanda astfel:

```
> t = seq(-6, 6, length = 400)

> f = 1/sqrt(2*pi)*exp(-tˆ2/2)

> plot(t, f, type = ”l”, lwd = 1)
```

Acest rezultat se poate transforma intr-o functie care se va scrie intr-un script R astfel:

File → New File → R Script si in ferestra de editare se scrie urmatorul cod

```
normal_density < − function(limit) {

 t = seq(-limit, limit, length = 400)

f = 1/sqrt(2*pi)*exp(-tb2/2)

 plot(t, f, type = ”l”, lwd = 1)

}

 normal_density(6)
```

RStudio. Dupa editare, scriptul este salvat (Ctrl+S) cu un nume de tipul ”my script.R” si este incarcat cu Code → Source File (Ctrl+Shift+O) sau din linia de comanda cu source(script file)
RStudio. O data incarcat scriptul, o functie care face parte din acest script se poate executa din linia de comanda: normal density(8) sau din fereastra de editare astfel: se selecteaza liniile dorite a fi executate si Ctrl+Enter, iar scriptul in intregime se executa cu Ctrl+Alt+R.

II. Estimarea mediei unei populatii: Media de selectie

Consideram o populatie cu media μ si dispersia σ^2, careia i se masoara atributul X. Din aceasta populatie se extrage un esantion aleator simplu de dimensiune n: $X_1, X_2, ..., X_n$. Aceste valori pot fi privite si ca variabile aleatoare independente si identic repartizate cu variabila X.

O functie pentru determinarea mediei de selectie a unui esantion dat intr-un fisier:

```
selection_mean < − function(filename) {

x = scan(filename);

m = mean(x) }

selection_mean(”sample.txt”)
```

RStudio. Fisierul cu numele filename trebuie sa fie in directorul de lucru.

III. Intervale de incredere pentru media unei populatii cu dispersia cunoscuta

Se considera o populatie cu dispersia cunoscuta σ^2. Se cauta un interval in care media μ, necunoscuta a populatiei sa se gaseasca cu probabilitate mare (0.90, 0.95 sau 0.99).

Un astfel de interval este urmatorul: $(\bar{x}_n - z*\cdot\ \sigma/\sqrt{n},\ \bar{x}_n + z*\cdot\ \sigma/\sqrt{n})$

unde z∗, numit valoarea critica, se determina 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 incredere.

Media de selectie, daca nu este data, se poate calcula astfel: $\bar{x}_n = mean(date\_esantion)$

<u>Exercitiu rezolvat</u>. *Durata vietii unui tip de baterie urmeaza cu aproximatie o lege normala cu dispersia de 9 ore. Pentru un esantion de 100 de baterii se masoara o medie de viata de 20 de ore. Sa se determine un interval de incredere de 90% pentru media de viata a intregii populatii.*

```
> 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].

APLICATII:

1. Se cauta un interval de incredere de 90% pentru media unei populatii normale cu dispersia cunoscuta $\sigma^2 = 100$. Pentru aceasta se utilizeaza un esantion aleator simplu de 25 de indivizi a carui medie de selectie (calculata) este 67.53.
2. Intr-o institutie publica exista un automat de cafea reglat in asa fel incat cantitatea de cafea dintr-un pahar urmeaza o lege normala cu deviatia standard $\sigma = 0.5$ oz. Pentru un esantion de $n = 50$ de pahare ales la intamplare, se masoara o medie a greutatii pentru un pahar de 5 oz. Sa se determine un interval de incredere de 95% pentru media de greutate a unui pahar de cafea.
3. Intr-o incercare disperata de a concura General Electric, compania ACME introduce un nou tip de becuri. ACME fabrica initial 100 de becuri a caror medie de viata masurata este 1280 de ore (deviatia standard a populatiei este 140 de ore). Sa se gaseasca un interval de ˆıncredere de 99% pentru media de viata a becurilor.

IV. Intervale de ıncredere pentru media unei populatii cu dispersia necunoscuta

Se considera o populatie careia nu i se cunoaste dispersia. In acest caz se foloseste drept estimator al deviatiei standard σ, deviatia standard a esantionului s.

Se cauta un interval in care media populatiei μ, necunoscuta si ea, sa se gaseasca cu probabilitate prescrisa (0.9, 0.95 sau 0.99).

Un astfel de interval este urmatorul: $(\bar{x}_n - t * \cdot \ s/\sqrt{n}, \bar{x}_n \ + \ t * \cdot \ s/\sqrt{n})$

unde t∗, numit valoarea critica, se determina astfel

t∗ = −qt(α/2,n−1) = qt(1−α/2,n−1)

α este egal cu 1− nivelul de incredere, iar s este deviatia standard a esantionului.

In cazul in care sunt cunoscute valorile din esantion, $\bar{x}_n$ si s se calculeaza astfel:

$\bar{x}_n$= mean(date-e¸santion), s = sd(date-e¸santion)

In calculele de mai jos vom folosi un estimator pentru eroarea standard a mediei, anume

$se = s/\sqrt{n}$.

<u>Exercitiu rezolvat</u>. *O companie ce produce jucarii doreste sa afle cat de interesante sunt produsele sale. 60 de copiii dintr-un esantion sunt rugati sa raspunda cu o valoare intre 0 si 5 si se determina o medie egala cu 3.3, cu o deviatie standard s = 0.4.*
Cat de interesante, in medie, sunt jucariile companiei (95% nivel de incredere)?

```
> 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]

APLICATII:

1. 196 de studenti alesi aleator au fost intrebati cat de multi bani au investit in cumparaturi online saptamana trecuta. Media a fost calculata la 44.65$, cu o dispersie (a esantionului) egala cu $s^2 = 2.25$. Calculati un interval de incredere de 99% pentru media populatiei (despre care se presupune ca urmeaza o lege normala).

2. O companie de dulciuri considera ca nivelul de zahar in produsele sale poate avea valori intre 1 si 20, urmand o lege normala. Se considera un esantion de 49 de produse. Media nivelului de zahar este 12 iar deviatia standard a esantionului este de 1.75.

(a) Determinati intervalele de incredere de 99% si 95% pentru media nivelului de zahar.

(b) Dupa modificarea retetei, s-au testat 49 produse si s-a gasit ca media nivelului de zahar este de 13.5 cu o deviatie standard de 1.25. Determinati un interval de incredere de 95% pentru media nivelului de zahar.

V. Testarea ipotezelor statistice - Testul z asupra proportiilor

<u>Exercitiu rezolvat</u>. *Un politician sustine ca va primi mai putin de 60% dintre voturi in colegiul sau. Un esantion dintr-o 100 de alegatori arata ca 63 dintre ei au votat pentru acest politician. Putem respinge afirmatia politicianului? (1% nivel de semnificatie)*

```
> 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 nula nu se poate respinge.