

Laborator 2 - Statistică descriptivă

Statistica descriptivă are rolul de a descrie trăsăturile principale ale unor eșantioane și constă în determinarea unor măsuri simple și analize grafice ale datelor din eșantion.

Analiza univariată reprezintă studiul unui singur atribut (trăsătură) a eșantionului. Acest **atributul** al membrilor unui eșantion este o proprietate sau o cantitate măsurată (observată). Acest atribut (care se presupune că este variabil) poate fi clasificat în cel puțin două moduri:

- a1) Atribut discret: într-un interval în care pot fi observate măsurătorile acestea pot lua întotdeauna un număr finit de valori (cunoscute). Exemplu: număr de accidente pe autostradă, grade de dificultate (foarte ușor, ușor, obișnuit, dificil etc), clasificări (asiatic/european/amerindian) etc.
- a2) Atribut continuu: în intervalul în care pot fi observate măsurătorile acestea pot lua practic orice valoare reală. Exemplu: greutate, înălțime, viteză etc.
- b1) Atribut cantitativ care poate fi:
 - discret (număr de erori, număr de copii pe familie etc);
 - continuu (viteză, volum, greutate etc).
- b2) Atribut calitativ (sau categoric)¹:
 - ordinal (mai bun/la fel/mai rău; pro/neutru/contra; grade de dificultate);
 - nominal (angajat/șomer; european/neeuropean; căsătorit/necăsătorit).

I. Reprezentarea grafică distribuției eșantionului

Datele sunt grupate în categorii (de exemplu intervale) și fiecărui interval i se asociază numărul de indivizi (din eșantion) a căror valoare cade în intervalul respectiv. (Frecvențele se pot înlocui cu procente).

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

Tipuri de reprezentări grafice:

1. **Stem and leaf** plot: pentru atribute cantitative (de obicei discrete) în număr relativ mic (cel mult 30 - 40?). Exemplu. Pentru datele de mai jos care pot fi un atribut continuu sau discret (cantitativ oricum)

```
0.6 0.2 1.6 2.0 1.1 0.5 1.5 2.3 3.4 1.9
0.4 0.5 1.2 0.9 2.1 1.6 1.8 2.6 3.1 2.5,
```

cifra de la stânga punctului zecimal reprezintă "stem"-ul, iar cea de la dreapta punctului zecimal este frunza ("leaf"):

```
0 || 6 2 5 4 5 9
1 || 6 1 5 9 2 6 8
2 || 0 3 1 6 5
3 || 4 1
```

Exercițiu rezolvat. Să se creeze în R un stem-and-leaf plot pentru următorul eșantion

11 14 21 32 17 24 21 35 52 44 21 28 36 49 41 19 20 34 37 29

¹Atributele calitative sunt discrete deoarece categoriilor li se poate asocia o valoare numerică: 1, 2, 3 etc.

```
> x = c(11, 14, 21, 32, 17, 24, 21, 35, 52, 44, 21, 28, 36, 49, 41, 19, 20, 34, 37, 29)
> stem(x)
The decimal point is 1 digit(s) to the right of the |
1 | 1479
2 | 0111489
3 | 24567
4 | 149
5 | 2
```

2. **Histograme:** se împarte domeniul într-un număr de intervale² și se reprezintă grafic sub forma unor coloane alăturate frecvențele de pe fiecare interval. Funcția utilizată este *hist()*.

Exercițiu rezolvat. În fișierul *date.txt* avem un eșantion pentru care vom reprezenta histograma astfel:

```
> sample = scan("sample.txt")
> min = min(sample)
> max = max(sample)
> min
[1] 41
> max
[1] 96
```

Putem alege să împărțim valorile pe intervalele [40, 50), [50, 60) etc ultimul interval fiind [90, 100) - sunt șase intervale. Histograma va fi reprezentată cu

```
> interval = seq(40, 100, 10)
> hist(sample, breaks = interval, right = F, freq = T)
```

Sau cu

```
> int = 6
> hist(sample, breaks = int, right = F, col = "blue")
```

- o *breaks* este un parametru care conține vectorul capetelor de interval (de la 40 la 100) sau un număr care indică numărul de intervale,
- o *right* ne spune că intervalele sunt în chise la dreapta (TRUE) sau deschise la dreapta,
- o un parametru similar *include.lowest* (sic) privește capătul din stânga,
- o *freq* indică dacă reprezentarea este a frecvențelor (TRUE) sau a procentelor core-spunzătoare (FALSE) - înălțimea relativă a coloanelor va fi aceeași.

3. **Bar chart** (Pareto): este o reprezentare asemănătoare histogramei, se folosește mai ales pentru atribute discrete (calitative sau cantitative). Această reprezentare presupune determinarea anterioară a frecvențelor (se construiește o tabelă a frecvențelor numărând observațiile care cad în aceeași categorie sau interval). Funcția utilizată este *barplot()*.

Exercițiu rezolvat. Să presupunem că următoarele valori reprezintă frecvențele unui eșantion

11 14 21 32 17 24 21 35 52 44 21 28 36 49 41 19 20 34 37 29

Reprezentarea lor se face astfel

²În cazul în care lungimea lor comună nu este evidentă, se poate recurge la următoarea formulă: $L = 1 + \frac{\ln n}{\ln 2}$

```
> freqv = c(9, 8, 12, 3, 17, 41, 29, 35, 32, 40, 19, 8)
> barplot(freqv, space = 0)
```

Exerciții propuse.

I.1 reprezentați stem-and-leaf plot pentru eșantionul din fișierul "sample1.txt".

I.2 Fișierul "unemploy2012.csv" conține rate ale șomajului în 2012 din majoritatea țărilor europene (cu două coloane numite 'country' și 'rate'). Reprezentați histograma ratelor șomajului folosind intervalele (0, 4], (4, 6], (6, 8], (8, 10], (10, 12], (12, 14] și (14, 300].

Indicație: citiți eșantionul astfel

```
> tablou = read.csv("unemploy2012.csv", header = T)
> rate = tablou[['rate']]
```

I.3 Fișierul "life_expect.csv" conține speranța de viață (la naștere, în 2012) din majoritatea țărilor europene (cu trei coloane numite 'country', 'female' și 'male'). Reprezentați histogramele speranței de viață pentru cele două grupe, împărțind eșantioanele în câte șapte intervale.

II. Analiza tendinței centrale

Analiza *tendinței centrale* este o aproximare a "centrului" distribuției eșantionului. (În cele ce urmează presupunem că datele din eșantion sunt ordonate $x_1 \leq x_2 \leq \dots \leq x_n$, deși nu toate statisticile de mai jos necesită ordonarea lor). Cele mai importante măsuri ale tendinței centrale sunt:

- **Media** - uzual media aritmetică a datelor din eșantion; de exemplu pentru eșantion de mai jos

3, 6, 4, 3, 6, 7, 8, 5

media este $M = (3 + 6 + 4 + 3 + 6 + 7 + 8 + 5)/8 = 42/8 = 5.25$

$$M = \frac{1}{n} \left(\sum_{k=1}^n x_k \right) \quad \text{în R: } \text{mean(eșantion)}$$

- **Mediana:** se ordonează crescător datele din eșantion și, dacă dimensiunea eșantionului este impară mediana este chiar valoarea din mijloc, iar dacă dimensiunea este pară, mediana este media celor două valori din mijloc.

Pentru eșantionul 3, 6, 4, 3, 6, 7, 8, 5, după ordonare: 3, 3, 4, 5, 6, 6, 7, 8, găsim că mediana este $Me = \frac{5+6}{2} = \frac{11}{2} = 5.5$.

Pentru eșantionul 3, 6, 4, 5, 2, 6, 9, 7, 8, 5, 4, după ordonare: 2, 3, 4, 4, 5, 5, 6, 6, 7, 8, 9, mediana este $Me = 5$.

$$Me = \begin{cases} x_{k+1}, & \text{dacă } n = 2k + 1 \\ \frac{x_k + x_{k+1}}{2}, & \text{dacă } n = 2k \end{cases} \quad \text{în R: } \text{median(eșantion)}$$

- **Mòdul** este valoarea care are cea mai mare frecvență în eșantion. În cazul în care există mai multe valori cu frecvență maximă, distribuția se va numi multi-modală.

Pentru eșantionul 3, 6, 4, 3, 6, 7, 8, 5, 3, 6, valorile 3 și 6 apar de cele mai multe ori - avem o distribuție bi-modală.

Pentru eșantionul 2, 6, 4, 3, 6, 7, 8, 5, 6, 4, mòdul este 6 (care apare de un număr maxim de ori) - distribuția eșantionului este uni-modală.

O funcție care să determine mòdul în R standard nu există (doar anumite pachete o conțin).

Exerciții propuse.

II.1 Calculați media și mediana eșantionului din fișierul "sample1.txt".

II.2 Calculați media și mediana eșantionului din fișierul "life_expect.csv".

II.3* Scrieți o funcție care să calculeze mòdul pentru un eșantion dat.

III. Împrăștierea și valorile aberante

Împrăștierea (sau *dispersia* datelor) reunește un grup de valori care măsoară împrăștierea datelor în jurul tendinței centrale.

- **domeniul datelor** (*range*) este diferența dintre valoarea maximă și valoarea minimă a datelor.

Pentru eșantionul 2, 6, 4, 3, 6, 7, 8, 5, 6, 4, lungimea intervalului este $8 - 2 = 6$.

$$Range = \max_{1 \leq k \leq n} x_k - \min_{1 \leq k \leq n} x_k$$

- **deviația standard** a eșantionului (*s*),

$$s = \sqrt{\frac{\sum_{k=1}^n (x_k - M)^2}{n - 1}} \quad \text{în R: } sd(eșantion)$$

eroarea standard a mediei eșantionului (*se*):

$$se = \frac{s}{\sqrt{n}}$$

- **dispersia** eșantionului (*s*²):

$$s^2 = \frac{\sum_{k=1}^n (x_k - M)^2}{n - 1} \quad \text{în R: } var(eșantion)$$

- **quartilele și intervalul interquartilic (IQR)**: prima quartilă Q_1 este mediana segmentului de eșantion cuprins între cea mai mică valoare din eșantion (x_1) și mediană, a treia quartilă Q_3 este mediana segmentului de eșantion cuprins între mediană și cea mai mare valoare din eșantion (x_n).

Funcția `quantile(eșantion)` returnează sub forma unui obiect (*date frame*), următoarele valori: minimul, prima quartilă, mediana, a doua quartilă și maximul. O quartilă poate fi obținută astfel

Q_i în R: `as.vector(quantile(eșantion))[i+1]`

$$IQR = Q_3 - Q_1$$

Exercițiu rezolvat. Funcția `summary(eșantion)` determină ceea ce se numește sumarul celor șase (sic) valori: \min , Q_1 , Me , M , Q_3 și \max .

```
> sample = c(9, 8, 12, 3, 17, 41, 29, 35, 32, 40, 19, 8)
> summary(sample)
  Min.   1st Qu.  Median Mean   3rd Qu.  Max.
  3.00   8.75   18.00  21.08  32.75   41.00
```

- **Valorile aberante** (*outliers*) sunt acele date dintr-un eșantion care au frecvență redusă și sunt fie mult prea mici, fie mult prea mari față de valoarea medie calculată. Valorile aberante se datorează fie unor greșeli de măsură, fie unor cauze naturale și pot afecta semnificativ valoarea mediei. La acest nivel îndepărtarea lor se poate face prin două metode:

- Cu ajutorul deviației standard a eșantionului: sunt considerate valori aberante acele valori care sunt în afara intervalului³ $(M - 2s, M + 2s)$.
- (Regula $1.5 \cdot IQR$) Cu ajutorul quartilelor: sunt considerate aberante acele valori din eșantion care se găsesc în afara intervalului⁴ $(Q_1 - 1.5 \cdot IQR, Q_3 + 1.5 \cdot IQR)$.

Exercițiu rezolvat. Pentru eșantionul de mai jos determinați valorile aberante folosind prima dintre metodele de mai sus.

1 91 38 72 13 27 11 19 5 22 20 19 8 17 11 15 13 23 14 17

```
> sample = c(1, 91, 38, 72, 13, 27, 11, 85, 5, 22, 20, 19, 8, 17, 11, 15, 13, 23, 14, 17)
> m = mean(sample)
> s = sd(sample)
> new_sample = vector()
> j = 0
> for(i in 1:length(sample))
>   if(sample[i] >= m - 2*s & sample[i] <= m + 2*s) {
>     j = j + 1
>     new_sample[j] = sample[i]
>   }
> new_sample
[1] [1] 1 38 72 13 27 11 5 22 20 19 8 17 11 15 13 23 14 17
```

Exerciții propuse.

III.1 Scrieți într-un script o funcție `outliers_mean(eșantion)` care să determine valorile aberante folosind prima metodă expusă mai sus. Verificați-o pe eșantionul din exemplul de mai sus.

³În general intervalul este de forma $(M - k \cdot s, M + k \cdot s)$, k putând fi chiar și mai mic decât 2.

⁴Intervalul este în general de forma $(Q_1 - k \cdot IQR, Q_3 + k \cdot IQR)$, $k \in \mathbb{R}$.

III.2 Scrieți în același script o funcție *outliers_iqr(eșantion)* care să determine valorile aberante folosind cea de-a doua metodă expusă mai sus (3/2 IQR).

III.3 Aplicați funcția *summary()* dar și funcțiile de mai sus eșantionului din fișierul "sample2.txt". Rezultatele sunt similare?

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**.