Laborator 2 - Statistică descriptivă

Statistica descriptivă are rolul de a descrie trăsăturile principale ale unor eşantioane şi constă în determinarea unor măsuri simple şi analize grafice ale datelor din eşantion.

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

a1) Atribut discret: într-un interval în care pot fi observate măsuratorile acestea pot lua întotdeauna un număr finit de valori (cunoscute). Exemplu: număr de accidente pe autostradă, grade de dificultate (foarte uşor, uşor, obişnuit, 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, înălţime, 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/şomer; european/neeuropean; căsătorit/necăsătorit).

I. Reprezentarea grafică a *distribuţiei* eşantionului

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

RStudio. Nu uitaţi să va setaţi 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					

   Exerciţiu rezolvat. Să se creeze în R un stem-and-leaf plot pentru următorul eşantion

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

[1] Atributele 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 intervale[2] şi se reprezintă grafic sub forma unor coloane alăturate frecvenţele de pe fiecare interval. Funcţia utilizată este *hist*().

   Exerciţiu rezolvat. În fisierul sample1.txt avem un eşantion 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 împărţim valorile pe intervalele $[40, 50)$, $[50, 60)$ etc ultimul interval fiind $[90, 100)$ - sunt şase intervale. Histograma va fi reprezentată cu

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

   Sau cu

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

   - *breaks* este un parametru care conţine vectorul capetelor de interval (de la 40 la 100) sau un număr care indică numărul de intervale,
   - *right* ne spune că intervalele sunt închise la dreapta (TRUE) sau deschise la dreapta,
   - un parametru similar *include.lowest* (sic) priveşte capătul din stânga,
   - *freq* indică daca reprezentarea este a frecvenţelor (TRUE) sau a procentelor corespunzătoare (FALSE) - înălţimea relativă a coloanelor va fi aceeaşi.

3. **Bar chart** (Pareto): este o reprezentare asemănătoare histogramei, se foloseşte mai ales pentru atribute discrete (calitative sau cantitative). Această reprezentare presupune determinarea anterioară a frecvenţelor (se construieşte o tabelă a frecvenţelor numărând observaţiile care cad în aceeaşi categorie sau interval). Funcţia utilizată este *barplot*().

   Exerciţiu rezolvat. Să presupunem că următoarele valori reprezintă frecvenţele unui eşantion

   9 8 12 3 17 41 29 35 40 19 8

   Reprezentarea lor se face astfel

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

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

Exerciţii propuse.

I.1 Reprezentaţi stem-and-leaf plot pentru eşantionul din fişierul "sample1.txt".

I.2 Fişierul "unemploy2012.csv" conţine rate ale şomajului în 2012 din majoritatea ţărilor europene (cu două coloane numite 'country' şi 'rate'). Reprezentaţi histograma ratelor şomajului folosind intervalele $(0, 4]$, $(4, 6]$, $(6, 8]$, $(8, 10]$, $(10, 12]$, $(12, 14]$ şi $(14, 300]$.

Indicaţie: citiţi eşantionul astfel

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

I.3 Fişierul "life_expect.csv" conţine speranţa de viaţă (la naştere, în 2012) din majoritatea ţărilor europene (cu trei coloane numite 'country', 'female' şi 'male'). Reprezentaţi histogramele speranţei de viaţă pentru cele două grupe, împărţind eşantioanle în câte şapte intervale.

II. Analiza *tendinţei centrale*

Analiza ***tendinţei centrale*** este o aproximare a "centrului" distribuţiei eşantionului. (În cele ce urmează presupunem că datele din eşantion sunt ordonate $x_1 \leqslant x_2 \leqslant \ldots \leqslant x_n$, deşi nu toate statisticile de mai jos necesită ordonarea lor). Cele mai importante măsuri ale tendinţei centrale sunt:

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

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

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

$$\boxed{M = \frac{1}{n}\left(\sum_{k=1}^{n} x_k\right)}$$ în R: *mean(eşantion)*

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

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

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

$$\boxed{Me = \begin{cases} x_{k+1}, & \text{dacă } n = 2k+1 \\ \dfrac{x_k + x_{k+1}}{2}, & \text{dacă } n = 2k \end{cases}}$$ în R: *median(eşantion)*

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

  Pentru eşantionul $3, 6, 4, 3, 6, 7, 8, 5, 3, 6$, valorile 3 şi 6 apar de cele mai multe ori - avem o distribuţie bi-modală.

  Pentru eşantionul $2, 6, 4, 3, 6, 7, 8, 5, 6, 4$, mòdul este 6 (care apare de un număr maxim de ori) - distribuţia eşantionului este uni-modală.

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

Exerciţii propuse.

II.1 Calculaţi media şi mediana eşanionului din fişierul ”sample1.txt”.

II.2 Calculaţi media şi mediana eşanioanelor din fişierul ”life_expect.csv”.

II.3* Scrieţi o funcţie care să calculeze mòdul pentru un eşantion dat.

III. Împrăştierea şi valorile aberante

Împrăştierea (sau *dispersia* datelor) reuneşte un grup de valori care măsoară împrăştierea datelor în jurul tendinţei centrale.

- ***domeniul datelor*** (***range***) este diferenţa dintre valoarea maximă şi valoarea minimă a datelor.

  Pentru eşantionul $2, 6, 4, 3, 6, 7, 8, 5, 6, 4$ domeniul este $8 - 2 = 6$.

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

- ***deviaţia standard*** a eşantionului ($\boldsymbol{s}$)

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

- ***dispersia*** eşantionului ($\boldsymbol{s}^2$):

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

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

  Funcţia *quantile*(*eşantion*) returnează sub forma unui obiect (*date frame*), următoarele valori: minimul, prima quartilă, mediana, a doua quartilă şi maximul. O quartilă poate fi obţinută astfel

$$Q_i \text{ în R: } \boxed{as.vector(quantile(eşantion))[\mathrm{i+1}]}$$

$$\boxed{IQR = Q_3 - Q_1}$$

Exerciţiu rezolvat. Funcţia *summary*(*eşantion*) determină ceea ce se numeşte sumarul celor şase (sic) valori: min, Q_1, Me, M, Q_3 şi 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 eşantion care au frecvenţă redusă şi sunt fie mult prea mici, fie mult prea mari faţă de valoarea medie calculată. Valorile aberante se datorează fie unor greşeli de măsură, fie unor cauze naturale şi pot afecta semnificativ valoarea mediei. La acest nivel îndepărtarea lor se poate face prin două metode:
  - Cu ajutorul deviaţiei standard a eşantionului: sunt considerate valori aberante acele valori care sunt în afara intervalului[3] $(M - 2s, M + 2s)$.
  - (regula $1.5 \cdot IQR$) cu ajutorul quartilelor: sunt considerate aberante acele valori din eşantion care se găsesc în afara intervalului[4] $(Q_1 - 1.5 \cdot IQR, Q_3 + 1.5 \cdot IQR)$.

Exerciţiu rezolvat. Pentru eşantionul de mai jos determinaţi 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)
> outliers = vector()
> j = 0
> for(i in 1:length(sample))
>    if(sample[i] <m - 2*s — sample[i] > m + 2*s) {
>       j = j + 1
>       outliers[j] = sample[i]
>    }
> outliers
[1] 91 85
```

Exerciţii propuse.

III.1 Scrieţi într-un script o funcţie *outliers_mean(eşantion)* care să determine valorile aberante folosind prima metodă expusă mai sus. Verificaţi-o pe eşantionul din exemplul de mai sus.

III.2 Scrieţi în acelaşi script o funcţie *outliers_iqr(eşantion)* care să determine valorile aberante folosind cea de-a doua metodă expusă mai sus (3/2 IQR).

III.3 Aplicaţi funcţia *summary*() dar şi funcţiile de mai sus eşantionului din fişierul ”sample2.txt”. Rezultatele sunt similare?

[3] În general intervalul este de forma $(M - k \cdot s, M + k \cdot s)$, k putând fi chiar şi mai mic decât 2.
[4] Intervalul este în general de forma $(Q_1 - k \cdot IQR, Q_3 + k \cdot IQR)$, $k \in \mathbb{R}$.

RStudio. După editare, scriptul este salvat (**Ctrl+S**) cu un nume de tipul ”my_script.R” şi este încărcat cu **Code** $\rightarrow$ **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**.

Temă pentru acasă.

3 puncte [1p: A1] + [2p: A2 sau A3]

A1. (1 punct) Reprezentaţi grafic funcţia de masă de probabilitate a distribuţiei binomiale $B(n, p)$, şi apoi funcţiile de masă de probabilitate $Poisson(\lambda)$ şi $Geometric(p)$ - doar primele k valori, k fiind un parametru dat.

A2. (2 puncte) Considerăm următorul eşantion aleator simplu care conţine masele a 45 de indivizi

79	71	89	57	76	64	82	82	67	80	81	65	73	79	79
60	58	83	74	68	78	80	78	81	76	65	70	76	58	82
59	73	72	79	87	63	74	90	69	70	83	76	61	66	71
60	57	81	57	65	81	78	77	81	81	63	71	66	56	62
75	64	74	74	70	71	56	69	63	72	81	54	72	91	92

Determinaţi mediana, media, deviaţia standard, cvartilele şi valorile aberante (dacă există). Reprezentaţi grafic distribuţia frecvenţelor cu intervalele $(40, 50], [51, 60], [61, 70], \ldots$.

A3. (2 puncte) Se consideră următorul eşantion format din notele de admitere ale unui grup de studenţi:

9.50	5.50	6.60	7.25	8.50	9.70	7.50	8.25	8.50	8.66	7.50	9.00	8.50	9.33
8.33	9.90	8.75	5.60	6.50	6.75	8.20	8.33	9.50	8.66	6.50	7.25	9.50	9.33
7.50	8.60	5.60	7.25	8.50	9.95	6.66	6.40	7.75	7.66	6.60	9.33	7.80	9.85
6.66	8.66	5.75	8.25	8.33	9.75	8.25	6.33	7.50	8.25	8.66	8.33	5.75	9.33
8.75	7.25	6.60	9.50	7.50	6.85	6.75	5.75	5.66	6.75	7.60	7.33	6.85	5.66

Să se determine media, mediana, deviaţia standard, quartilele şi să se afle (dacă există) valorile aberante ale eşantionului. Reprezentaţi grafic distribuţia frecvenţelor cu intervalele $(4, 5], (5, 6], (6, 7], \ldots$.

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