Laborarator 12 – Probabilități și Statistică Matematică

STATISTICA DESCRIPTIVA

Statistica descriptiva are rolul de a descrie trasaturile principale ale unor esantioane si consta in determinarea unor masuri simple si analize grafice ale datelor din esantion. Analiza univariata reprezinta studiul unui singur atribut (trasatura) a esantionului. Acest atribut al membrilor unui esantion este o proprietate sau o cantitate masurata (observata). Acest atribut (care se presupune ca este variabil) poate fi clasificat in cel putin doua moduri:

A1) Atribut discret: intr-un interval in care pot fi observate masuratorile acestea pot lua intotdeauna un numar finit de valori (cunoscute). Exemplu: numar de accidente pe autostrada, grade de dificultate (foarte usor, usor, obisnuit, dificil etc), clasificari (asiatic/european/amerindian) etc.

A2) Atribut continuu: in intervalul in care pot fi observate masuratorile acestea pot lua practic orice valoare reala. Exemplu: greutate, inaltime, viteza etc.

B1) Atribut cantitativ care poate fi:
- discret (numar de erori, numar de copii pe familie etc);
- continuu (viteza, volum, greutate etc).

B2) Atribut calitativ (sau categoric):
- ordinal (mai bun/la fel/mai rau; pro/neutru/contra; grade de dificultate);
- nominal (angajat/¸somer; european/neeuropean; casatorit/necasatorit).

I. Reprezentarea grafica distributiei esantionului

Datele sunt grupate in categorii (de exemplu intervale) si fiecarui interval i se asociaza numarul de indivizi (din esantion) a caror valoare cade in intervalul respectiv. (Frecventele se pot inlocui cu procente).

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

Tipuri de reprezentari grafice:

1. *Stem and leaf plot*: pentru atribute cantitative (de obicei discrete) in numar 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 stanga punctului zecimal reprezinta ”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					

EXERCITIU REZOLVAT: Sa se creeze in R un stem-and-leaf plot pentru urmatorul esantion

11 4 21 32 17 24 21 35 52 44 21 28 36 49 41 19 20 34 37 29

```
> 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 imparte domeniul intr-un numar de interval si se reprezinta grafic sub forma unor coloane alaturate frecventele de pe fiecare interval. Functia utilizata este *hist()*.

EXERCITIU REZOLVAT: In fisierul date.txt avem un esantion pentru care vom reprezenta histograma astfel:

```
> sample = scan(”sample.txt”)

> min = min(sample)

> max = max(sample)

> min [1] 41

> max [1] 96
```

Putem alege sa impartim valorile pe intervalele [40,50), [50,60) etc. ultimul interval fiind [90,100) - sunt sase intervale. Histograma va fi reprezentata 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 contine vectorul capetelor de interval (de la 40 la 100) sau un numar care indica numarul de interval;

◦ *right* ne spune ca intervalele sunt inchise la dreapta (TRUE) sau deschise la dreapta;

◦ un parametru similar *include.lowest* priveste capatul din stanga;

◦ *freq* indica daca reprezentarea este a frecventelor (TRUE) sau a procentelor corespunzatoare (FALSE) - inaltimea relativa a coloanelor va fi aceeasi.

3. *Bar chart (Pareto):* este o reprezentare asemanatoare histogramei, se foloseste mai ales pentru atribute discrete (calitative sau cantitative). Aceasta reprezentare presupune determinarea anterioara a frecventelor (se construieste o tabela a frecventelor numarand observatiile care cad in aceeasi categorie sau interval). Functia utilizata este *barplot()*.

EXERCITIU REZOLVAT: Sa presupunem ca urmatoarele valori reprezinta frecventele unui esantion 9 8 12 3 17 41 29 35 40 19 8

Reprezentarea lor se face astfel:

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

> barplot(frecv, space = 0)
```

II. Analiza tendintei centrale

Analiza tendintei centrale este o aproximare a ”centrului” distributiei esantionului. (In cele ce urmeaza presupunem ca datele din esantion sunt ordonate $x_1 \leq x_2 \leq \cdots \leq x_n$, desi nu toate statisticile de mai jos necesita ordonarea lor). Cele mai importante masuri ale tendintei centrale sunt:

- *Media* - uzual media aritmetica a datelor din esantion; de exemplu pentru esantion 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

$$M = \frac{1}{n}\sum_{k=1}^{n} x_k$$ in R: *mean(esantion)*

- *Mediana*: se ordoneaza crescator datele din esantion si, daca dimensiunea esantionului este impara mediana este chiar valoarea din mijloc, iar daca dimensiunea este para, mediana este media celor doua valori din mijloc.

Pentru esantionul 3,6,4,3,6,7,8,5, dupa ordonare: 3,3,4,5,6,6,7,8, gasim ca mediana este

Me = (5 + 6)/ 2 = 11/ 2 = 5.5.

Pentru esantionul 3,6,4,5,2,6,9,7,8,5,4, dupa ordonare: 2,3,4,4,5,5,6,6,7,8,9 mediana este Me = 5.

$$\text{Me} = \begin{cases} x_{k+1}, daca\ n = 2k+1 \\ \frac{x_k + x_{k+1}}{2}, daca\ n = 2k \end{cases}$$ in R: *median(esantion)*

- *Modul* este valoarea care are cea mai mare frecventa in esantion. In cazul in care exista mai multe valori cu frecventa maxima, distributia se va numi multi-modala.

Pentru e¸=santionul 3,6,4,3,6,7,8,5,3,6, valorile 3 si 6 apar de cele mai multe ori - avem o distributie bi-modala.

Pentru esantionul 2,6,4,3,6,7,8,5,6,4, modul este 6 (care apare de un numar maxim de ori) - distributia esantionului este uni-modala.

O functie care sa determine modul in R standard nu exista (doar anumite pachete o contin).

III. Imprastierea si valorile aberante

Imprastierea (sau dispersia datelor) reuneste un grup de valori care masoara imprastierea datelor in jurul tendintei centrale.

- *domeniul datelor (range)* este diferenta dintre valoarea maxima si valoarea minima a datelor. Pentru esantionul 2,6,4,3,6,7,8,5,6,4 domeniul este 8−2 = 6.

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

- *deviatia standard* a esantionului (s)¸

$$s = \sqrt{\frac{\sum_{k=1}^{n}(x_k - M)^2}{n-1}}$$ in R: *sd(esantion)*

- *dispersia* esantionului (s^2):

$$s^2 = \frac{\sum_{k=1}^{n}(x_k - M)^2}{n-1}$$ in R: *var(esantion)*

- *quartilele si intervalul interquartilic (IQR):* prima quartila Q_1 este mediana segmentului de esantion cuprins intre cea mai mica valoare din esantion (x_1) si mediana, a treia quartila Q_3 este mediana segmentului de esantion cuprins intre mediana si cea mai mare valoare din esantion (x_n).

Functia *quantile(esantion)* returneaza sub forma unui obiect (*date frame*), urmatoarele valori: minimul, prima quartila, mediana, a doua quartila ¸si maximul. O quartila poate fi obtinuta astfel

Q_i in R: *as.vector(quantile(esantion))[i+1]*

$$IQR = Q_3 - Q_1$$

EXERCITIU REZOLVAT: Functia *summary(esantion)* determina ceea ce se numeste sumarul celor sase valori: min, Q_1, Me, M, Q_3 si 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 esantion care au frecventa redusa si sunt fie mult prea mici, fie mult prea mari fata de valoarea medie calculata. Valorile aberante se datoreaza fie unor greseli de masura, fie unor cauze naturale si pot afecta semnificativ valoarea mediei. La acest nivel indepartarea lor se poate face prin doua metode:

- Cu ajutorul deviatiei standard a esantionului: sunt considerate valori aberante acele valori care sunt in afara intervalului $(M - 2s, M + 2s)$.

- (regula $1.5 \cdot IQR$) cu ajutorul quartilelor: sunt considerate aberante acele valori din esantion care se gasesc in afara intervalului $(Q1 - 1.5 \cdot IQR, Q3 + 1.5 \cdot IQR)$.

EXERCITIU REZOLVAT: Pentru esantionul de mai jos determinati 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
```

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.

APLICATII:

1. Consideram urmatorul esantion aleator simplu care contine masele a 45 de indivizi;

84 72 88 78 76 84 84 82 87 80 81 69 73 79 79 75 68 80 74 68 77 80 78 81 76 75 70 76 78 82

72 73 86 79 91 70 84 73 69 70 83 76 47 67 76

Determinati mediana, media, deviatia standard, cvartilele si valorile aberante (daca exista).

2. Se considera urmatorul esantion format din notele de admitere ale unui grup de studenti:

6.50 8.60 9.60 7.25 8.50 9.95 6.66 6.40 7.75 7.66 8.60 9.33 7.80 9.85 9.50 5.50 7.60 7.25 8.50 9.70 9.50 8.25 7.50 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

Sa se determine media, mediana, deviatia standard, quartilele si sa se afle (daca exista) valorile aberante ale esantionului.