BAZELE STATISTICII

Conf.univ.dr. Ciprian Chirilă

2. Analiza unei serii statistice univariate

2.1. Variabile cantitative

A. Variabilă discretă

1. Prezentarea seriei (distribuţiei) statistice

- seria simplă $X:(x_i)$, cu i=1,n , când $n_1=n_2=\ldots=n_i=1$
- seria cu frecvenţe diferite X: $\begin{pmatrix} x_i \\ n_i \end{pmatrix}$, când $n_i \neq n_j$.

- *Frecvenţe absolute cumulate crescător ($N_i \downarrow$) sau descrescător ($N_i \uparrow$)*

- exprimă *numărul de unităţi statistice* cumulate "până la" sau "peste" nivelul considerat al caracteristicii, adică valori $\leq x_i$, respectiv $\geq x_i$.

$$N_i \downarrow = N_{i-1} \downarrow + n_i = \sum_{h=1}^{i} n_h$$

$$N_i \uparrow = N_{i+1} \uparrow + n_i = \sum_{h=i}^{m} n_h$$

- *Frecvenţe relative cumulate crescător ($F_i \downarrow$) sau descrescător ($F_i \uparrow$)*

  - exprimă *ponderea unităţilor statistice* cumulate "până la" sau "peste" nivelul considerat al caracteristicii, adică valori $\leq x_i$, respectiv $\geq x_i$.

2. Caracterizarea seriei folosind metode grafice

a. *Poligonul frecvenţelor:*

- construirea acestuia presupune găsirea locului geometric al punctelor A_i de coordonate (x_i,n_i) sau (x_i,f_i) şi unirea acestora prin segmente de dreaptă.
- aproximează forma unei distribuţii.

b. *Curba frecvenţelor:*

- presupune ajustarea printr-o linie curbă, continuă a poligonului frecvenţelor.
- aproximează mai bine forma de distribuţie a colectivităţii după variabila considerată.

3. Analiza seriei folosind metode numerice

Presupune calculul indicatorilor statisticii descriptive, cunoscuţi şi sub denumirea de caracteristici numerice ale unei distribuţii.

3.1. Indicatori ai tendinţei centrale (mărimi medii)

a. Definire:

- mediile sunt acele valori în jurul cărora se repartizează unităţile unei populaţii.
- cele mai importante mărimi medii sunt media artitmetică, modul şi mediana .

3. Analiza seriei folosind metode numerice

b. Media aritmetică ($\overline{x}$)

- Media aritmetică este valoarea pe care am observa-o dacă unităţile statistice ar înregistra aceleaşi valori ale variabilei (dacă nu ar exista variaţii ale valorilor înregistrate de unităţile statistice).

- Mod de calcul în cazul seriilor simple şi seriilor cu frecvenţe diferite (variabilă discretă).

- Media simplă: $$\bar{x} = \frac{\sum_i x_i}{n}$$
- Media ponderată.

$$\bar{x} = \frac{\sum_i x_i \cdot n_i}{\sum_i n_i} \qquad sau \qquad \bar{x} = \sum_i x_i \cdot f_i$$

Observaţie:

Media aritmetică este sensibilă la prezenţa valorilor extreme (*outliers*).

Cele mai importante proprietăţi ale mediei aritmetice:

1. Media unei distribuţii este o valoare internă:

$$x_{min} \leq \bar{x} \leq x_{max.}$$

2. Media este o mărime normală: suma abaterilor valorilor individuale ale unei variabile *X* de la media lor este egală cu zero.

c. Modul (Mo)

- este valoarea variabilei cea mai frecvent observată într-o distribuţie, adică valoarea x_i care corespunde frecvenţei maxime (n_{imax}).

Observaţie:

- modul poate fi aflat doar în cazul seriilor cu frecvenţe diferite.
- o distribuţie poate avea una, două sau mai multe valori modale (serii unimodale, bimodale sau plurimodale).

Interpretare: Cele mai multe unităţi înregistrează valoarea modală.

d. Mediana (Me)

- este acea valoare a variabilei unei serii ordonate, crescător sau descrescător, până la care şi peste care sunt distribuite în număr egal unităţile colectivităţii: jumătate din unităţi au valori mai mici decât mediana, iar jumătate au valori mai mari decât mediana.
- corespunde locului unităţii mediane calculate astfel:

$$U^{Me} = \frac{n+1}{2}$$

Aflarea medianei se face diferit în funcţie de tipul seriei:

1. Serii simple:

- număr impar de termeni.
- număr par de termeni.

2. Serii cu frecvenţe diferite

- se calculează unitatea mediană (U^{Me}).
- se calculează $N_i \downarrow$

- se află prima valoare $N_i \downarrow \geq U^{Me}$
- valoarea x_i corespunzătoare acesteia este *Me*.

Observaţie:

- mediana nu este influenţată de valorile extreme.

e. Relaţii între cele trei mărimi medii

Arată forma unei distribuţii:

1. Când $\bar{x} = Mo = Me$ distribuţia este simetrică.
2. Când $\bar{x} > Me > Mo$ distribuţia este asimetrică la dreapta (asimetrie pozitivă).
3. Când $\bar{x} < Me < Mo$ distribuţia este asimetrică la stânga (asimetrie negativă).

f. Quartilele

- sunt valori ale variabilei care împart volumul eşantionului în 4 părţi egale.
- reprezentare grafică şi mod de calcul (Q_1, Q_2, Q_3).

g. Decile

- decila unu (D_1) şi decila 9 (D_9).

h. Reprezentarea diagramei "box-plot" sau "box-and-whiskers"

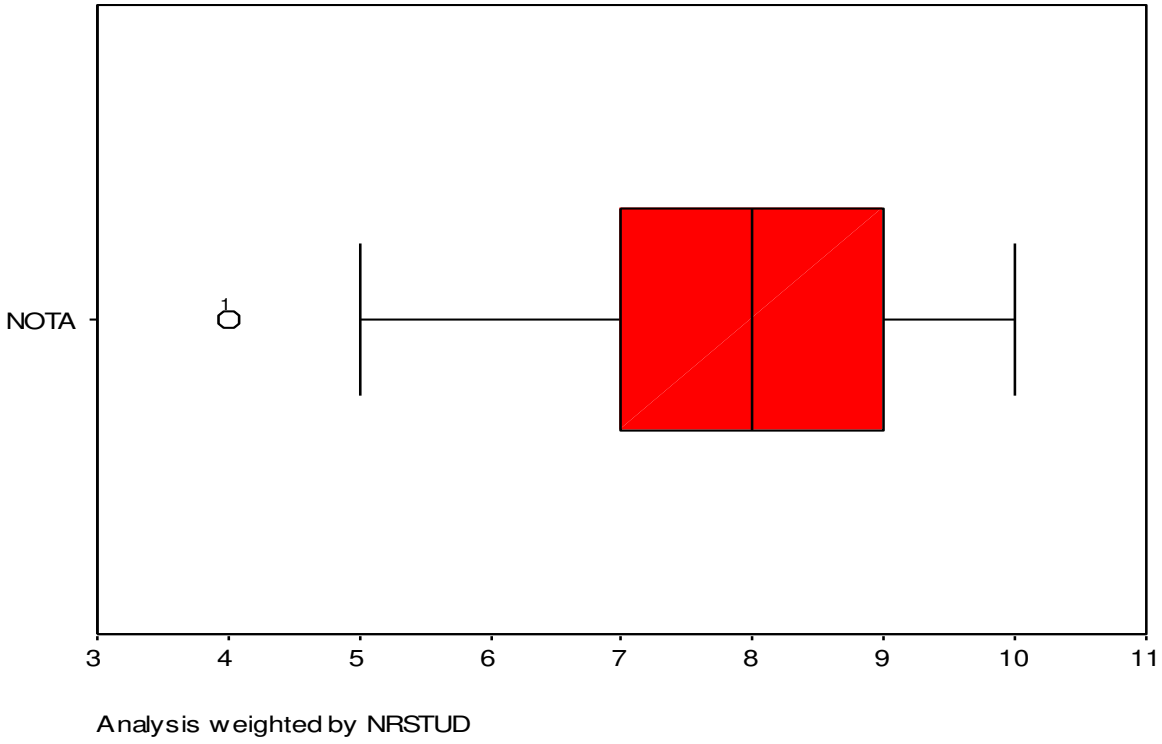
- Forma diagramei (x_{min}, D_1, Q_1, Q_2, Q_3, D_9, x_{max});

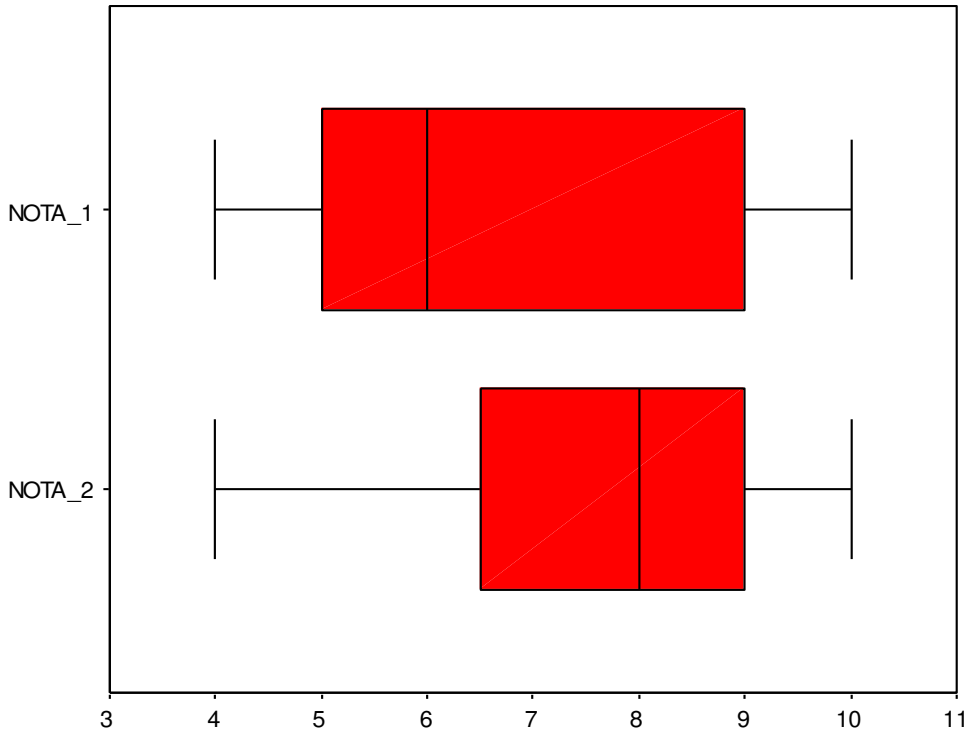
- Avantaje:

  - permite aprecierea nivelului mediu (*Me*), dispersiei şi asimetriei unei distribuţiei;

  - facilitează compararea mai multor distribuţii (prin reprezentarea simultană a diagramelor).

Diagrama box-plot




NOTA_1
NOTA_2
3
4
5
6
7
8
9
10
11

3.2. Indicatori ai dispersiei (variaţiei)

Definire:

- dispersia exprimă gradul de variaţie a valorilor individuale ale unei variabile faţă de nivelul mediu.
- aprecierea fenomenului de dispersie al unei distribuţii permite identificarea gradului de reprezentativitate a mediei unei distribuţii.

Indicatori sintetici ai dispersiei:

1. Abaterea medie liniară

$$\bar{d} = \frac{\sum_i |x_i - \bar{x}|}{n}$$, respectiv $$\bar{d} = \frac{\sum_i |x_i - \bar{x}| \cdot n_i}{\sum_i n_i}$$

- arată cu cât variază, în medie, valorile x_i ale variabilei faţă de nivelul mediu al distribuţiei, în sens pozitiv şi negativ.
- se exprimă în aceeaşi unitate de măsură cu cea a variabilei.

2. Varianţa

$$s^2 = \frac{\sum_i (x_i - \bar{x})^2}{n}$$, respectiv $$s^2 = \frac{\sum_i (x_i - \bar{x})^2 \cdot n_i}{\sum_i n_i}$$

Varianţa este întotdeauna pozitivă, nu are unitate de măsură şi nu se interpretează.

Prin ridicarea la pătrat a abaterilor valorilor x_i faţă de medie

creşte “influenţa” valorilor extreme asupra nivelului varianţei.

3. Abaterea standard (*s*)

- arată cu cât variază, în medie, valorile x_i ale variabilei faţă de nivelul mediu al distribuţiei, în sens pozitiv şi negativ.
- se exprimă în aceeaşi unitate de măsură cu cea a variabilei.

4. Coeficientul de variaţie (*v*)

$$v = \frac{s}{\bar{x}} \cdot 100$$

- se exprimă în procente.
- valori ridicate ale acestui coeficient (v>50%) arată o distribuţie eterogenă, care se caracterizează printr-o variaţie mare a valorilor x_i faţă de nivelul mediu şi o medie nereprezentativă.
- este sensibil faţă de valoarea mediei: cu cât media este mai apropiată de zero, cu atât coeficientul de variaţie este mai dificil de folosit (tinde spre infinit).

5. Intervalul interquartilic

$I_Q = Q_3 - Q_1$.

- cuprinde 50% din volumul eşantionului.

6. Coeficientul de variaţie interquartilic

$$V_Q = \frac{\frac{Q_3 - Q_1}{2}}{Me} \cdot 100$$

- În mod sintetic, cele mai importante caracteristici numerice ale unei distribuţii pot fi “cuplate” astfel:
- media - abaterea standard (valoare absolută) - coeficientul de variaţie (valoare relativă)
- mediana - intervalul interquartilic (valoare absolută) - coeficientul de variaţie interquartilic (valoare relativă)

3.3. Indicatori ai formei

1. Asimetria:

- reprezintă o deviere de la forma simetrică a unei distribuţii.

Asimetria poate fi apreciată:

- *pe cale grafică:* curba frecvenţelor, diagrama box-plot.
- *pe cale numerică:* - prin calculul indicatorilor de asimetrie (*Skewness*).

a.

b. Coeficientul de asimetrie Fisher

$$\gamma_1 = \frac{\mu_3}{s^3}$$

2. Boltirea

- este definită prin compararea distribuţiei empirice cu distribuţia normală din punctul de vedere al variaţiei variabilei X şi a frecvenţei n_i.

  Boltirea poate fi apreciată:

– *pe cale grafică*: curba frecvenţelor.

- *numeric:* prin calculul indicatorilor boltirii (*kurtosis*).

a.

b. *Coeficientul de boltire Fisher*:

$$\gamma_2 = \frac{\mu_4}{\mu_2^2} - 3 = \frac{\mu_4}{s^4} - 3$$

3.4. Indicatorii statisticii descriptive în Excel

Column1	
Mean	8.6
Standard Error	0.347735
Median	8
Mode	10
Standard Deviation	1.904622
Sample Variance	3.627586
Kurtosis	-0.14315
Skewness	-0.40554
Range	8
Minimum	4
Maximum	12
Sum	258
Count	30

B. Variabilă continuă

1. *Prezentarea seriei statistice*

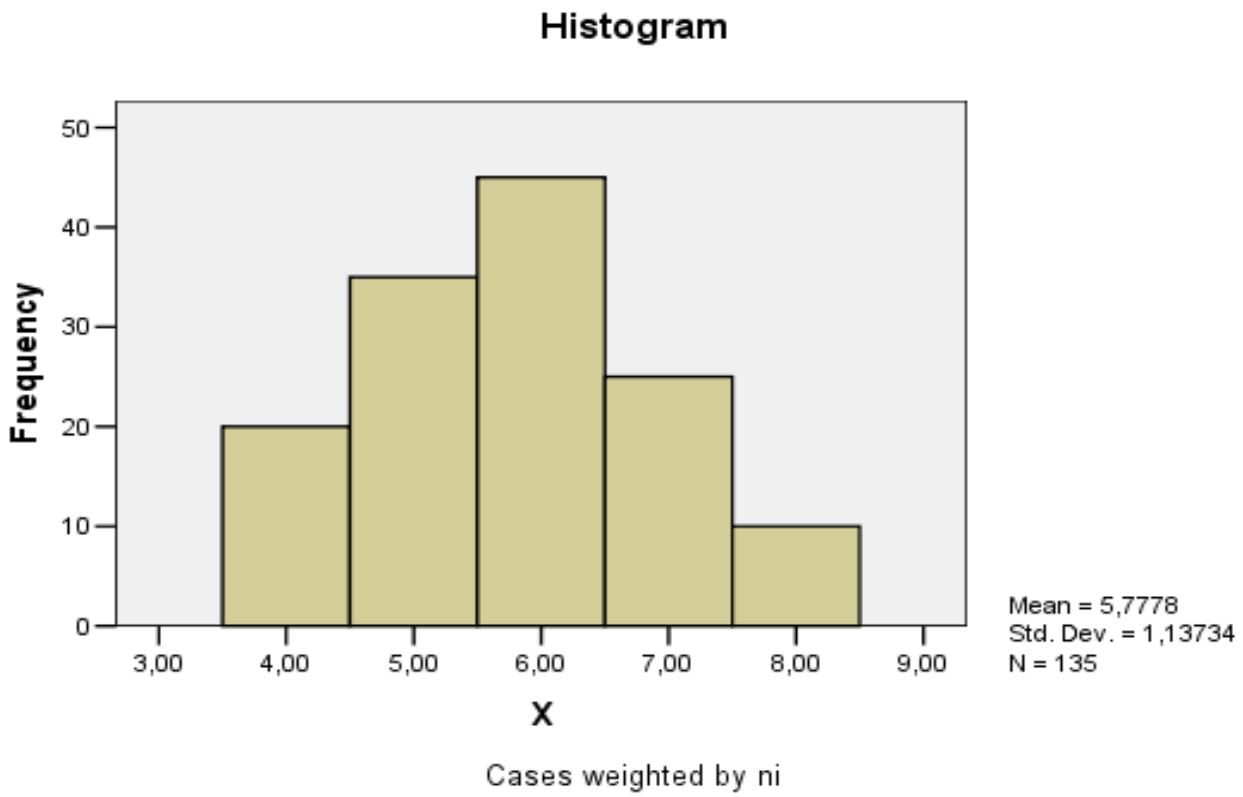
- gruparea unităţilor statistice este realizată pe intervale de variaţie.

Observaţie:

- Gruparea pe intervale de variaţie duce la pierderea unei părţi a informaţiei iniţiale.

2. Prelucrarea seriei statistice folosind metode grafice

a. Histograma


Histogram
Frequency
50
40
30
20
10
0
3,00
4,00
5,00
6,00
7,00
8,00
9,00
X
Mean = 5,7778
Std. Dev. = 1,13734
N = 135
Cases weighted by ni

b. Poligonul frecvenţelor

c. Curba frecvenţelor

d. Curba frecvenţelor cumulate:

- este reprezentarea grafică a funcţiei de repartiţie a frecvenţelor unei variabile: $F(X<x_i)$.

3. *Indicatori ai statisticii descriptive*

- se calculează în mod identic, prin "discretizarea" variabilei (calculul mijlocului intervalelor de variaţie).

2. Analiza unei serii univariate

2.1. Variabilă cantitativă

A. Variabilă discretă

B. Variabilă continuă

2.2. Variabilă calitativă

I. Tipuri de variabile

A. Variabile nominale

B. Variabile ordinale

II. Reprezentare grafică

a) Variabile nominale:

- Pentru a reprezenta structura pe categorii la nivelul unui eşantion se calculează frecvenţe relative;
- Reprezentarea structurii unui eşantion se realizează folosind diagrame de structură: dreptunghiul, pătratul şi cercul de structură (*Bar Chart*, *Pie Chart*).

b) Variabile ordinale:

Reprezentarea structurii unui eşantion se realizează folosind diagrame de structură (*Bar Chart*, *Pie Chart*)

Exemplu:

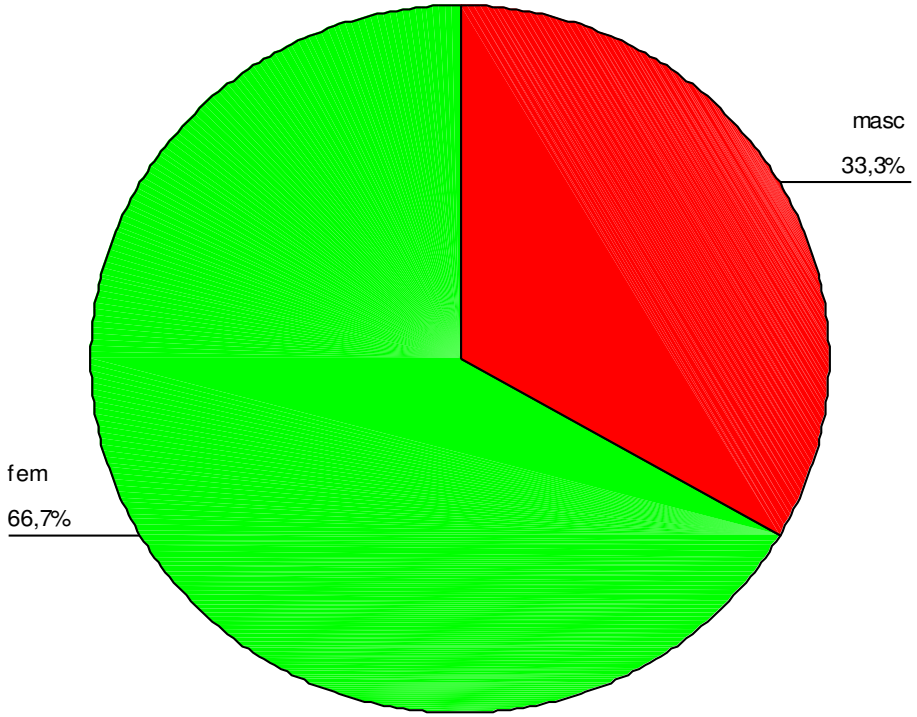


Figura 1. Distribuţia studenţilor unei grupe pe sexe, la 1 ianuarie 2011

III. Indicatori statistici specifici

a) Variabile nominale:

1. Mărimi relative

- *frecvenţe relative (f_i)*

2. Indicatori ai tendinţei centrale

- modul arată categoria cea mai frecvent observată.

b) Variabile ordinale:

1. Mărimi relative

- *frecvenţe relative (f_i)*
- *frecvenţe relative cumulate (F_i)*

2. Indicatori ai tendinţei centrale

- mediana şi modul.