BAZELE STATISTICII

Conf.univ.dr. Ciprian Chirilă

2. Analiza unei serii statistice univariate

2.1. Variabile cantitative

Variabilă discretă

1. Prezentarea seriei (distribuției) statistice

- seria simplă $X:(x_i)$, cu i=1,n, când $n_1=n_2=\ldots=n_i=1$ seria cu frecvențe 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 unități 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 unităților 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) și unirea acestora prin segmente de dreaptă.
- aproximează forma unei distribuţii.
- b. Curba frecvenţelor:
- presupune ajustarea printr-o linie curbă, continuă a poligonului frecvențelor.
- aproximează mai bine forma de distribuţie a colectivităţii după variabila considerată.

3. Analiza seriei folosind metode numerice

Presupune calculul indicatorilor statisticii descriptive, cunoscuți și sub denumirea de caracteristici numerice ale unei distribuții.

3.1. Indicatori ai tendinței centrale (mărimi medii)

a. Definire:

- mediile sunt acele valori în jurul cărora se repartizează unitățile unei populații.
- cele mai importante mărimi medii sunt media artitmetică, modul și mediana.

3. Analiza seriei folosind metode numerice

b. Media aritmetică ($\overline{\boldsymbol{\mathcal{X}}}$)

- Media aritmetică este valoarea pe care am observa-o dacă unitățile statistice ar înregistra aceleași valori ale variabilei (dacă nu ar exista variații ale valorilor înregistrate de unitățile statistice).
- Mod de calcul în cazul seriilor simple şi seriilor cu frecvențe diferite (variabilă discretă).

□ Media simplă:

$$\bar{x} = \frac{\sum_{i} x_i}{n}$$

Media ponderată.

$$\bar{x} = \frac{\sum_{i} x_{i} \cdot n_{i}}{\sum_{i} n_{i}} \qquad sau \quad \bar{x} = \sum_{i} x_{i} \cdot f_{i}$$

Observație:

Media aritmetică este sensibilă la prezența valorilor extreme (outliers).

Cele mai importante proprietăți ale mediei aritmetice:

1. Media unei distribuții 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}) .

Observație:

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

Interpretare: Cele mai multe unități 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 unitățile colectivității: jumătate din unități au valori mai mici decât mediana, iar jumătate au valori mai mari decât mediana.
- corespunde locului unității mediane calculate astfel:

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

Aflarea medianei se face diferit în funcție de tipul seriei:

1. Serii simple:

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

2. Serii cu frecvențe 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.

Observație:

mediana nu este influențată de valorile extreme.

e. Relații între cele trei mărimi medii

Arată forma unei distribuții:

- 1. Când $\bar{x} = Mo = Me$ distribuția 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 părți egale.
- reprezentare grafică și mod de calcul (Q_1 , Q_2 , Q_3).

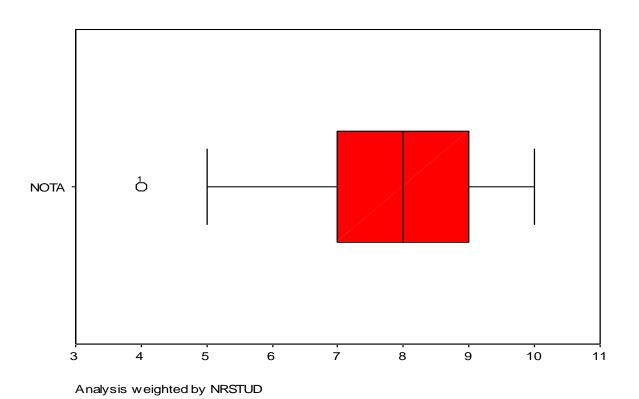
g. Decile

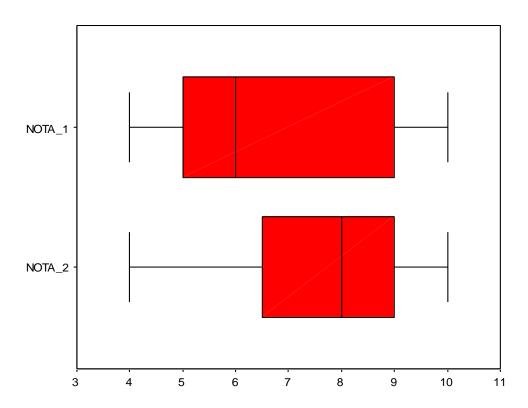
- decila unu (D_1) și decila $9 (D_9)$.
- h. Reprezentarea diagramei "box-plot" sau "box-and-whiskers"
- Forma diagramei $(x_{min}, D_1, Q_1, Q_2, Q_3, D_9, x_{max_n});$

□ Avantaje:

- permite aprecierea nivelului mediu (*Me*), dispersiei și asimetriei unei distribuției;
- facilitează compararea mai multor distribuții (prin reprezentarea simultană a diagramelor).

Diagrama box-plot





3.2. Indicatori ai dispersiei (variației)

Definire:

- dispersia exprimă gradul de variație a valorilor individuale ale unei variabile față 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

$$\overline{d} = \frac{\sum_{i} |x_i - \overline{x}|}{n} \quad \text{, respectiv} \quad \overline{d} = \frac{\sum_{i} |x_i - \overline{x}| \cdot n_i}{\sum_{i} n_i}$$

- arată cu cât variază, în medie, valorile x_i ale variabilei față de nivelul mediu al distribuției, în sens pozitiv și negativ.
- se exprimă în aceeași unitate de măsură cu cea a variabilei.

2. Varianța

$$s^{2} = \frac{\sum_{i} (x_{i} - \overline{x})^{2}}{n} \qquad \text{, respectiv} \qquad s^{2} = \frac{\sum_{i} (x_{i} - \overline{x})^{2} \cdot n_{i}}{\sum_{i} n_{i}}$$

Varianța este întotdeauna pozitivă, nu are unitate de măsură și nu se interpretează.

Prin ridicarea la pătrat a abaterilor valorilor x_i față de medie

crește "influența" valorilor extreme asupra nivelului varianței.

- 3. Abaterea standard (s)
- arată cu cât variază, în medie, valorile x_i ale variabilei față de nivelul mediu al distribuției, în sens pozitiv și negativ.
- se exprimă în aceeași unitate de măsură cu cea a variabilei.
- 4. Coeficientul de variație (v)

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

- 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 față de valoarea mediei: cu cât media este mai apropiată de zero, cu atât coeficientul de variație 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 variație interquartilic

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

- În mod sintetic, cele mai importante caracteristici numerice ale unei distribuții pot fi "cuplate" astfel:
- media abaterea standard (valoare absolută) coeficientul de variație (valoare relativă)
- □ mediana intervalul interquartilic (valoare absolută) coeficientul de variație 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 frecvențelor, 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 frecvențelor.
- 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

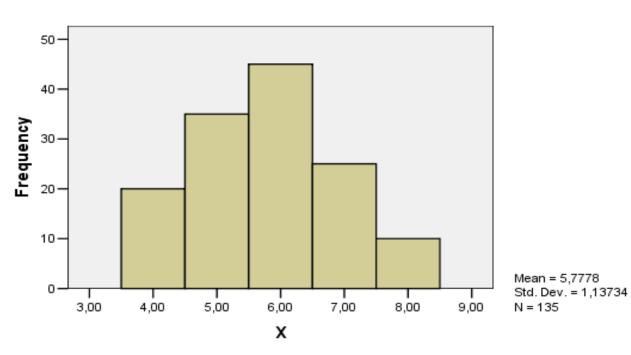
B. Variabilă continuă

- 1. Prezentarea seriei statistice
- gruparea unităților statistice este realizată pe intervale de variație.

Observație:

- Gruparea pe intervale de variație duce la pierderea unei părți a informației inițiale.
- 2. Prelucrarea seriei statistice folosind metode grafice
- a. Histograma

Histogram



Cases weighted by ni

- b. Poligonul frecvențelor
- c. Curba frecvențelor
- d. Curba frecvențelor cumulate:
- este reprezentarea grafică a funcției de repartiție a frecvențelor unei variabile: F(X<x_i).
- 3. Indicatori ai statisticii descriptive
 - se calculează în mod identic, prin "discretizarea" variabilei (calculul mijlocului intervalelor de variație).

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. <u>Variabile ordinale</u>

II. Reprezentare grafică

a) Variabile nominale:

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

b) Variabile ordinale:

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

Exemplu:

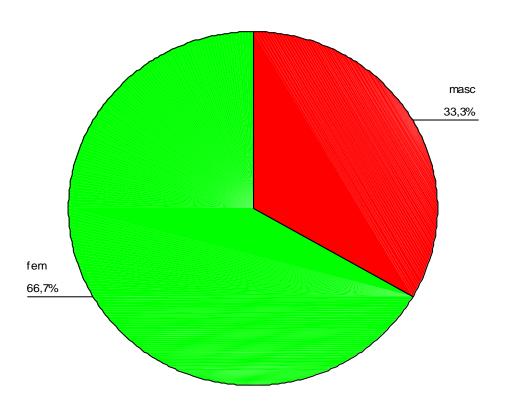


Figura 1. Distribuţia studenţilor unei grupe pe sexe, la 1 ianuarie 2011

III. Indicatori statistici specifici

a) Variabile nominale:

1. Mărimi relative

frecvențe relative (f_i)

2. Indicatori ai tendinței centrale

<u>modul</u> arată categoria cea mai frecvent observată.

b) Variabile ordinale:

1. Mărimi relative

- $frecvenţe relative (f_i)$
- $frecvențe relative cumulate <math>(F_i)$

2. Indicatori ai tendinței centrale

- mediana și modul.