



SEMINAR BAZELE STATISTICII

SEMINAR 4

Analiza unei serii statistice univariate. Cazul unei variabile cantitative discrete

1. Mărimi medii:

- media aritmetică, modul, mediana.
- Quartile
- Decile

2. Indicatori ai dispersiei

- varianta (s^2),
- abaterea standard (s)
- coeficientul de variație (v)

3. Indicatori ai formei

- Asimetrie (Skewness)
- Boltire (Kurtosis)

RECAPITULARE SEMINAR 3

Media aritmetica (\bar{x})

Media simplă: $\bar{x} = \frac{\sum x_i}{n}$

Media ponderată: $\bar{x} = \frac{\sum x_i \cdot n_i}{\sum n_i}$ sau $\bar{x} = \sum x_i \cdot f_i$

Mediana (M_e)

este acea valoare a variabilei unei serii ordonate, crescător sau descrescător, până la care și peste care sunt distribuite în număr egal unitățile colectivității:

Modul (M_o)

este valoarea variabilei cea mai frecvent observată într-o distribuție, adică valoarea x_i care corespunde frecvenței maxime ($n_{i\max}$).

Unitatea mediană (U^{Me}).

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

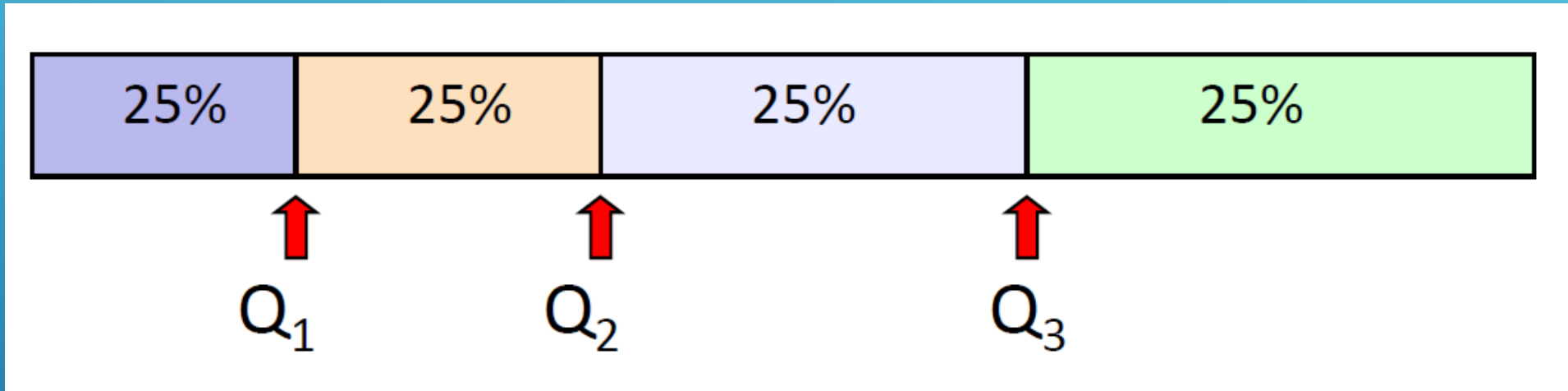
QUARTILE & DECILE

indică **poziția** unei valori în raport cu întregul set de date

Several thin, parallel white lines of varying lengths and slopes are positioned in the bottom right corner of the slide, creating a modern, abstract graphic element.

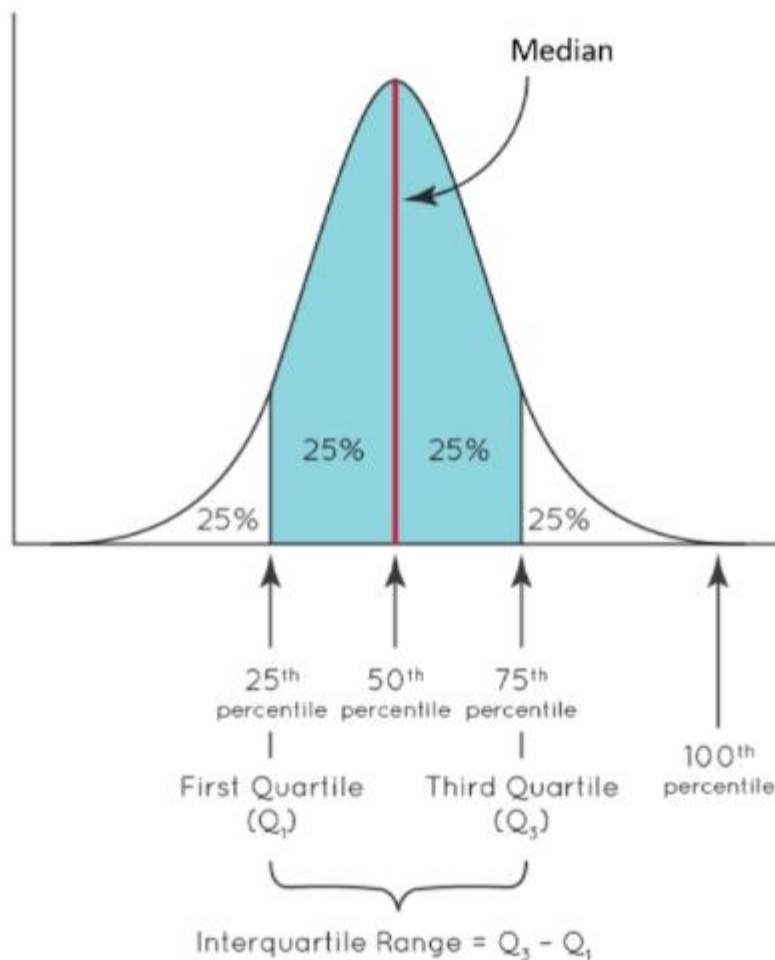
Quartilele

- sunt valori ale variabilei care împart volumul eșantionului în 4 părți egale.



Interpretare:

- Prima cuartilă, Q_1 , este valoarea pentru care 25% dintre observații sunt mai mici, iar 75% sunt mai mari
- Q_2 este același cu mediana (50% sunt mai mici, 50% sunt mai mari)
- 25% dintre observații sunt mai mari decât a treia quartila (Q_3)



- Prima quartilă (Q₁) este definită ca numărul mijlociu dintre cel mai mic număr (minimum) și mediana setului de date. Este, de asemenea, cunoscut sub numele de quartila inferioară sau a 25-a, deoarece 25% din date sunt sub acest punct.
- A doua quartilă (Q₂) este mediana unui set de date; astfel 50% din date se află sub acest punct.
- A treia quartilă (Q₃) este valoarea medie dintre mediană și cea mai mare valoare (maximum) a setului de date. Este cunoscut sub numele de quartila superioară sau a 75-a, deoarece 75% din date se află sub acest punct.

Exemplu : 11 12 13 16 16 17 18 21 22



$n = 9$

Q_1

Q_2

Q_3

$$Q_1 = 0.25 (9 + 1) = 2.5 \text{ pozitia} - Q_1 = 12.5$$

$$Q_2 = 0.50 (9 + 1) = 5 \text{ pozitia} - Q_2 = 16$$

$$Q_3 = 0.75 (n + 1) = 7.5 \text{ pozitia} - Q_3 = 19.5$$

Decile

Reprezinta valori ale caracteristici(variabilei) care impart volumul colectivitatii in 10 parti egale

Mod de calcul

$$D_1 : U^{D_1} = \frac{n+1}{10}$$

$$D_2 : U^{D_2} = \frac{2(n+1)}{10}$$

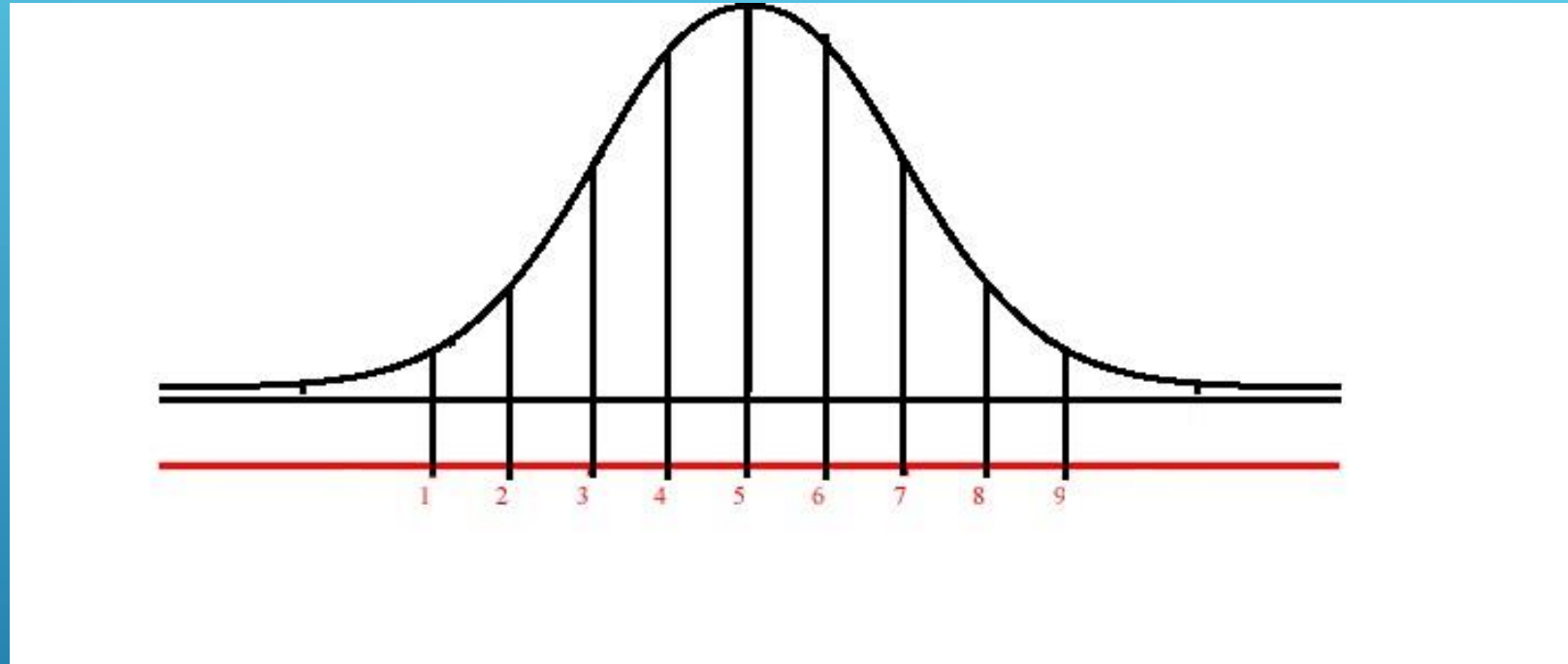
·
·

$$D_5 = Me = Q_2$$

·
·

$$D_9 : U^{D_9} = \frac{9(n+1)}{10}$$

Unde $U^{D_{1...9}}$ - reprezinta unitati decilice



Indicatori sintetici ai dispersiei:

1. Abaterea medie liniară $\bar{d} = \frac{\sum_i |x_i - \bar{x}|}{n}$ sau $\bar{d} = \frac{\sum_i |x_i - \bar{x}| \cdot n_i}{\sum_i n_i}$

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

2. Varianța $s^2 = \frac{\sum_i (x_i - \bar{x})^2}{n}$ (serie simpla) sau

$$s^2 = \frac{\sum_i (x_i - \bar{x})^2 \cdot n_i}{\sum_i n_i} \quad (\text{serie cu frecvente diferite})$$

- **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) (deviația standard)

$$s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n}}$$
 serie simpla

serie cu frecvente diferite $s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2 * n_i}{\sum n_i}}$

- 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.

Exemplu:

Salarii (mii lei) : 3; 7; 8 ; 12; 15 $\bar{x} = \frac{45}{5} = 9$ mii lei

$$\begin{aligned} s &= \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n}} = \sqrt{\frac{(3-9)^2 + (7-9)^2 + (8-9)^2 + (12-9)^2 + (15-9)^2}{5}} \\ &= \sqrt{\frac{36+4+1+9+36}{5}} = \sqrt{\frac{86}{5}} = \sqrt{17.2} = 4.147 \text{ mii lei} \end{aligned}$$

Interpretare:

Salariatii obtin un salariu care variaza (se abate), in medie, de la salariul mediu (9 mii lei) cu ± 4.15 mii lei (in sens pozitiv si in sens negativ)

4. Coeficientul de variație (v)

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

- se exprimă în procente.
- **$v > 50\%$** - arată o **distribuție eterogenă**, care se caracterizează printr-o variație mare a valorilor x_i față de nivelul mediu și o **medie nerepresentativă**.
- **$v < 50\%$** - arată o **distribuție omogenă**, care se caracterizează printr-o variație mică a valorilor x_i față de nivelul mediu și o **medie reprezentativă**.
- 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).

Coeficient de variație (CV)	Interpretare: populația poate fi considerată
$CV < 0,10$	<u>omogenă</u>
$0,10 \leq CV < 0,20$	<u>relativ omogenă</u>
$0,20 \leq CV < 0,30$	<u>relativ eterogenă/relativ heterogenă</u>
$> 0,30$	<u>eterogenă/heterogenă</u>

Exemplu:

Eșantionul 1:

$$\overline{x_1} = 25 \text{ ani}$$

$$s_1 = 2 \text{ ani}$$

Eșantionul 2:

$$\overline{x_2} = 50 \text{ ani}$$

$$s_2 = 2 \text{ ani}$$

$$v_1 = \frac{s_1}{\overline{x_1}} * 100 = \frac{2}{25} * 100 = 8\%$$

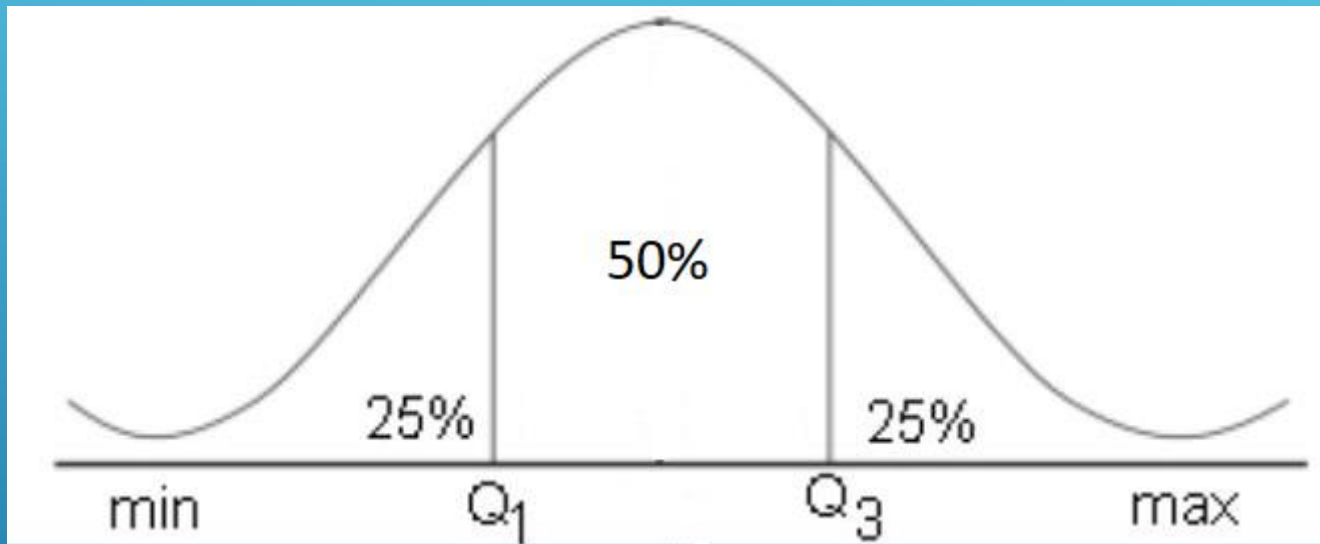
$$v_2 = \frac{s_2}{\overline{x_2}} * 100 = \frac{2}{50} * 100 = 4\%$$

v_1 și v_2 au valori $< 50\%$ → :
ambele distribuții sunt
omogene, se caracterizează
prin dispersie mică, iar mediile
sunt reprezentative.

5. Amplitudinea intervalului interquartilic

$$I_Q = Q_3 - Q_1$$

- cuprinde 50% din volumul eșantionului.



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).

Indicatori ai formei

1. Asimetria (Skewness):

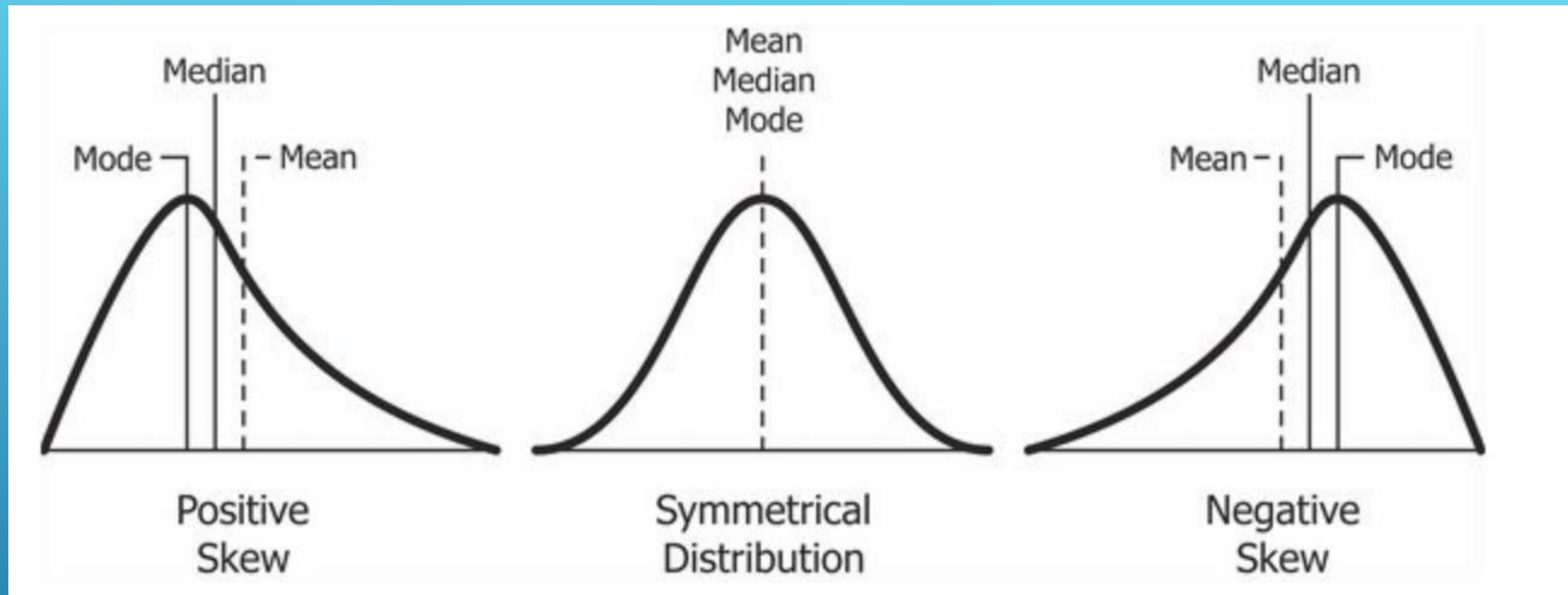
- reprezintă o deviere de la forma simetrică a unei distribuții.

Asimetria poate fi apreciată:

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

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

μ_3 este momentul centrat de ordin 3



**Distributie asimetrica
la dreapta (pozitiva)**

$\gamma > 0$ – asimetrie la dreapta

**Distributie simetrica
– distributie normala**

$\gamma = 0$ – distributie simetrica

**Distributie asimetrica
la stanga (negativa)**

$\gamma < 0$ – asimetrie la stanga

2. Boltirea (Kurtosis)

- este definită prin compararea distribuției empirice cu distribuția normală din punctul de vedere al variației variabilei X și a frecvenței n_i .

Boltirea poate fi apreciată:

- *pe cale grafică*: curba frecvențelor
- *numeric*: prin calculul coeficientului de boltire Fisher (*kurtosis*):

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

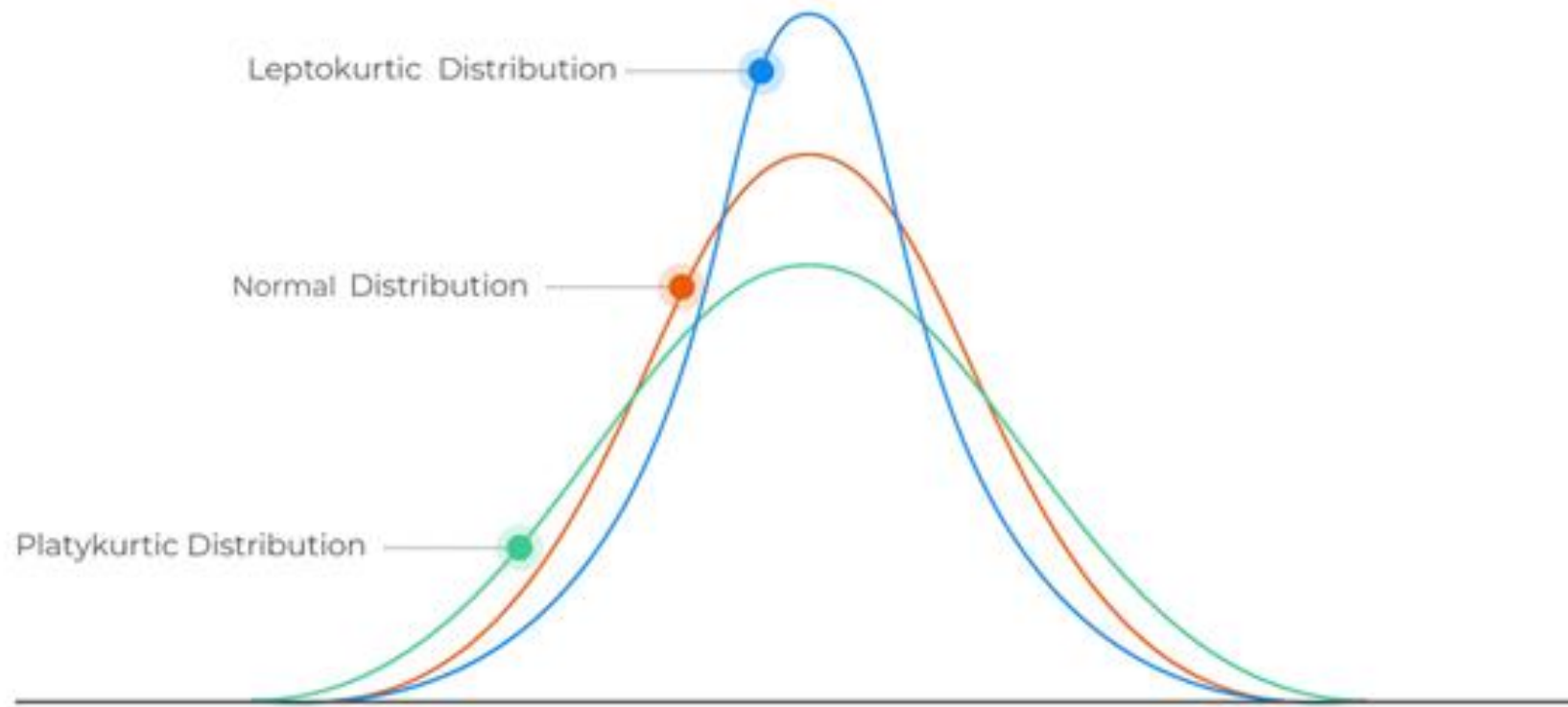
μ_2 este momentul centrat de ordin 2

μ_4 este momentul centrat de ordin 4

$\gamma_2 > 0$ – distributie leptocurtica

$\gamma_2 = 0$ – distributie mezocurtica (normala)

$\gamma_2 < 0$ – distributie platicurtica



APLICATII

1. Se înregistrează vânzările unor firme în anul 2019 (mil. lei): 13, 11, 10, 12, 18, 17, 17.

Se cere să afle valoarea quartilei 3.

Several thin, parallel white lines are drawn diagonally across the bottom right corner of the slide, extending from the middle towards the bottom right edge.

3. Se înregistrează salariile lunare pentru un eșantion de persoane și se obțin următoarele rezultate: 3, 6, 9, 7, 13 (mii lei).

Să se aprecieze :

- a) Tipul distribuției statistice.
- b) Omogenitatea distribuției.

4. Distribuția numărului de copii pentru familiile dintr-un bloc se prezintă astfel:

Nr. copii	0	1	2	3	4
Nr. familii	10	15	30	3	2

Se cere să se aprecieze reprezentativitatea mediei.

7. În urma prelucrării datelor privind salariul lunar obținut (mii lei) pe un eșantion de salariați, s-au obținut următoarele rezultate:

Se cere să aprecieze:

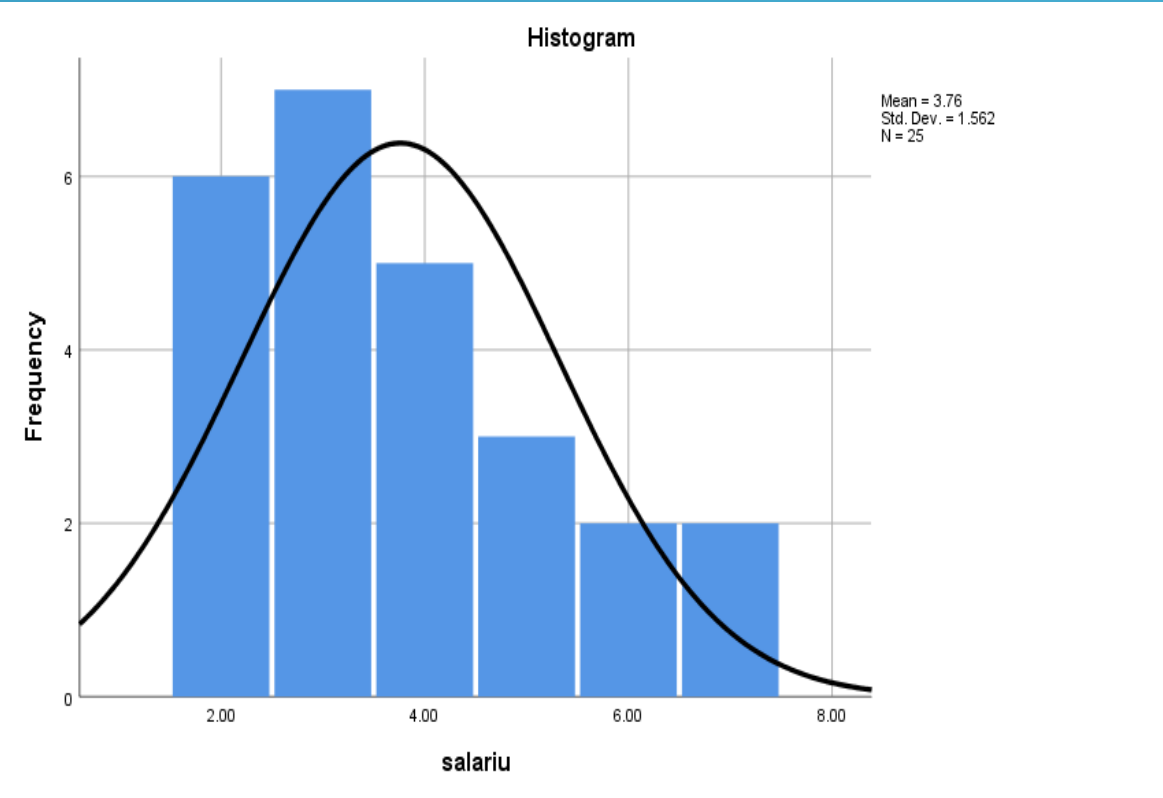
- a) reprezentativitatea mediei;
- b) asimetria și boltirea.

Column1	
Mean	4.53
Median	4.00
Mode	4.00
Standard Deviation	1.46
Sample Variance	2.12
Kurtosis	-0.45
Skewness	0.32
Range	5.00
Minimum	2.00
Maximum	7.00
Sum	68.00
Count	15.00

8.. În urma prelucrării datelor privind salariul lunar obținut (mii lei) pe un eșantion de salariați, s-au obținut următoarele rezultate:

Se cere:

- a) să se interpreteze valorile quartilelor
- b) să se aprecieze forma distribuției.



Statistics		
salariu		
N	Valid	25
	Missing	0
Mean		3.7600
Median		3.0000
Mode		3.00
Std. Deviation		1.56205
Variance		2.440
Skewness		.719
Kurtosis		-.378
Range		5.00
Minimum		2.00
Maximum		7.00
Percentiles	25	2.5000
	50	3.0000
	75	5.0000

1. Media aritmetica, ca indicator fundamental al tendintei centrale:

- a) reprezinta valoarea care modifica nivelul totalizator al seriei;
- b) se foloseste cel mai frecvent si reprezinta suma valorilor raportata la numarul lor;
- c) se calculeaza adunand valorile existente;
- d) nu poate determina numarul de unitati din colectivitate.

2. Media aritmetica reprezinta:

- a) valoarea pe care ar purta-o fiecare unitate statistica daca distributia ar fi eterogena;
- b) valoarea pe care ar purta-o fiecare unitate statistica daca distributia ar fi omogena;

3. 25% din salariatii unei intreprinderii au un salariu de cel mult 4 milioane lei. Aceasta valoare reprezinta:

- a) quartila unu;
- b) mediana;
- c) quartila trei.

4. 75% din salariatii unei intreprinderii au un salariu de cel mult 7 milioane lei. Aceasta valoare reprezinta:

- a) quartila unu;
- b) mediana;
- c) quartila trei.

5. 50% din salariatii unei intreprinderii au un salariu de cel mult 5 milioane lei. Aceasta valoare reprezinta:

- a) quartila unu;
- b) mediana;
- c) quartila trei.

6. Intr-o repartitie normala valoarea fata de care 75% din valorile individuale sunt mai mici iar 25% din valorile individuale sunt mai mari este:

- a. cuartila a doua;
- b. cuartila a treia;
- c. cuartila intai;
- d. valoarea modala.

7. Daca dorim sa obtinem un indicator al tendintei centrale pentru date masurate pe orice scala, vom determina:

- a) media aritmetica;
- b) modul;

8. Daca dorim sa obtinem un indicator al tendintei centrale care sa nu fie afectat de valorile extreme, sa poata fi calculat pentru date masurate pe orice scala (cu exceptia celei nominale) si sa aiba o singura valoarea, vom calcula:

- a) media aritmetica
- b) modul;
- c) mediana.

9. Quartilele sunt marimi de pozitie care impart colectivitatea in:

- a. $n-1$ parti egale;
- b. 4 parti egale;
- c. 5 parti egale;

10. O distributie este platicurtica atunci cand:

- a) $\gamma_2 = 0$
- b) $\gamma_2 > 0$
- c) $\gamma_2 < 0$