The background of the slide is composed of several overlapping geometric shapes in various colors: dark purple, orange, yellow, green, blue, and light purple. These shapes create a modern, abstract design.

Studying medicine is  
like true love..  
You suffer, fail, learn,  
fail again, fight and grow,  
but you never want to let go.

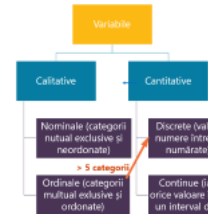
# **STATISTICĂ DESCRIPTIVĂ VARIABLE CANTITATIVE**

Sorana D. Bolboacă

# Despre ...

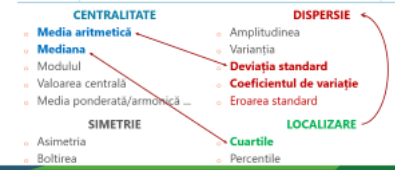
## Obiective educaționale

- Indicatori statistici de sumarizare a datelor cantitative: centralitate, dispersie, localizare, simetrie și boltire
- Sumarizarea grafică: graficul de distribuție (histograma și cutia cu mustați), graficul de asociere (scatter)
- Ce raportăm în funcție de tipul variabilei ...



## INDICATORI STATISTICI DESCRIPTIVI

STATISTICA EȘANTIONULUI = ESTIMATOR PUNCTUAL AL PARAMETRULUI POPULAȚIEI



## Indicatori de centralitate

o Statistica eșantionului = estimator punctual al parametrului populației

- Media aritmetică
- Mediana
- Modulul
- Media ponderată
- Valoarea centrală

## Indicatori de dispersie

- Amplitudinea
- Variația
- Deviația standard
- Eroarea standard
- Coeficientul de variație

Varabilitate vs. Diversitate

## Indicatori de localizare (dispersie): cvartile

$IC = Q_3 - Q_1$   
unde  $Q_3$  = cvartila 3 (percentila 75%),  $Q_1$  = cvartila 1 (percentila 25%)  
 $IC$  = intervalul cvartilic

Măsură a dispersiei pentru 50% din datele de mijloc.



## Indicatori de simetrie și boltire

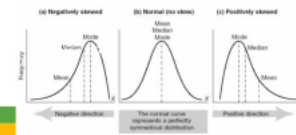
Interpretarea – aplicată populației

Asimetria  $< -1$  sau  $> +1$  → distribuție înalt asimetrică

Asimetria  $< -1$  → asimetrie la stânga / pozitivă

Asimetria  $> +1$  → asimetrie la dreapta / negativă

**SKEW**  
**SKEW.P**



## Tipuri de grafice

**Histograma**  
1 variabilă cantitativă  
(distribuție)

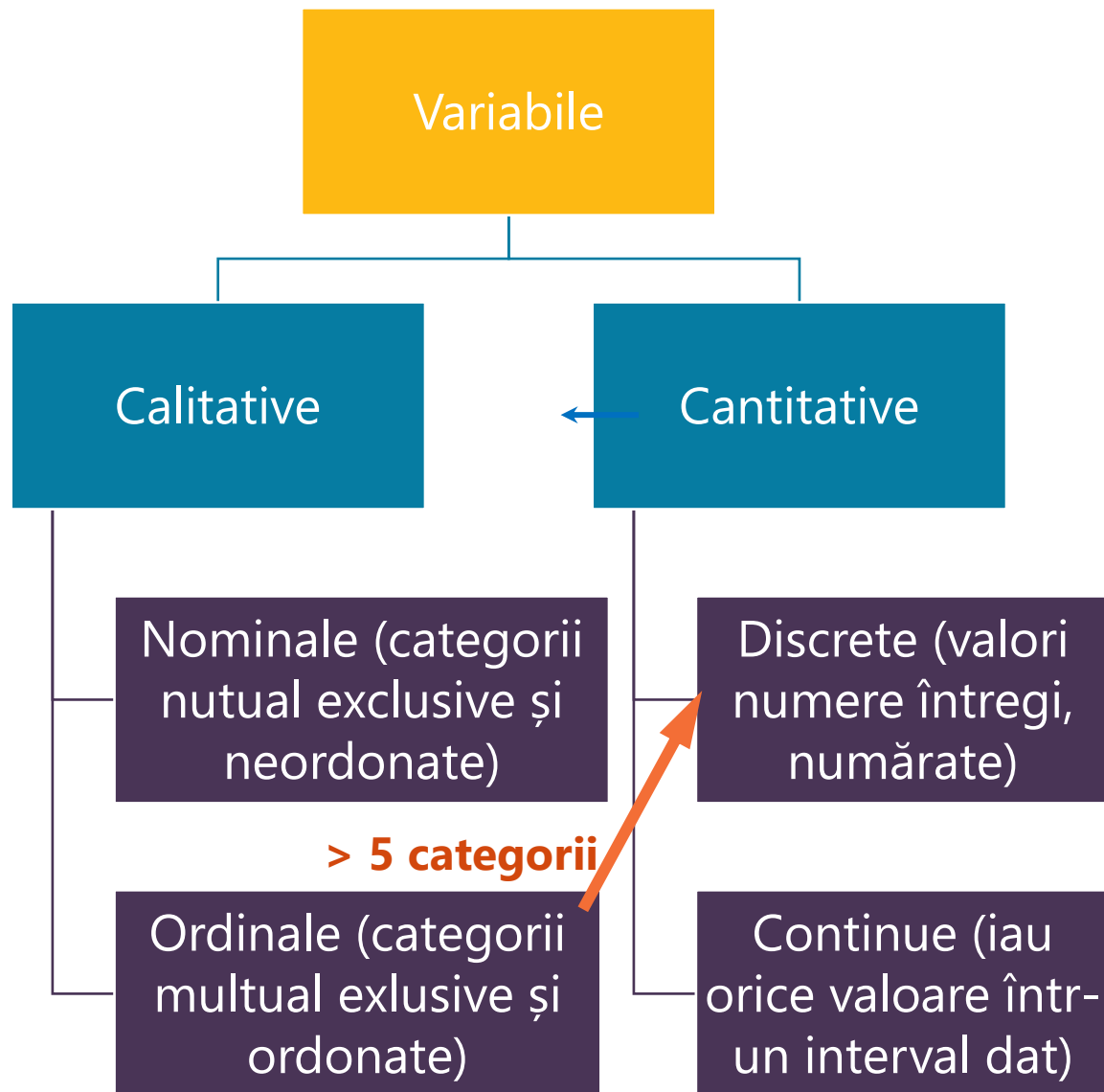
**Cutia cu mustați**  
1 variabilă cantitativă +  
1 variabilă de grupare  
(nominală sau ordinală)  
(distribuție)

**Nor de puncte**  
2 variabile cantitative  
(asociere / relație)

# Obiective educaționale

- Indicatori statistici de sumarizare a datelor cantitative: centralitate, dispersie, localizare, asimetrie și boltire
- Sumarizarea grafică: graficul de distribuție (histograma și cutia cu mustăți), graficul de asociere (scatter)
- Ce raportăm în funcție de tipul variabilei ...





# INDICATORI STATISTICI DESCRIPTIVI

STATISTICA EȘANTIONULUI = ESTIMATOR PUNCTUAL AL PARAMETRULUI POPULAȚIEI

## CENTRALITATE

- **Media aritmetică**
- **Mediana**
- Modulul
- Valoarea centrală
- Media ponderată/armonică ...

## SIMETRIE

- Asimetria
- Boltirea

## DISPERSIE

- Amplitudinea
- Varianța
- **Deviația standard**
- **Coeficientul de variație**
- Eroarea standard

## LOCALIZARE

- **Cuartile**
- Percentile



# Indicatori de centralitate

- o Statistica eșantionului = estimator punctual al parametrului populației

Media aritmetică

Mediana

Modulul

Media ponderată

Valoarea centrală

# Indicatori de centralitate

## AVERAGE

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$$

## MEDIAN

Eșantion → statistica

Impar (3, 5, 7 ...)

$$Me = X_{\frac{n+1}{2}}$$

$$\mu = \frac{\sum_{i=1}^N x_i}{N}$$

Populație → parametru

Par (2, 4, ...)

$$Me = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2}$$

Modulul: cea mai frecventă observație a seriei

**MODE=MODE.SNGL      MODE.MULT**

# Indicatori de centralitate

Scorurile obținute la examenul practic de o grupă de 11 studenți:

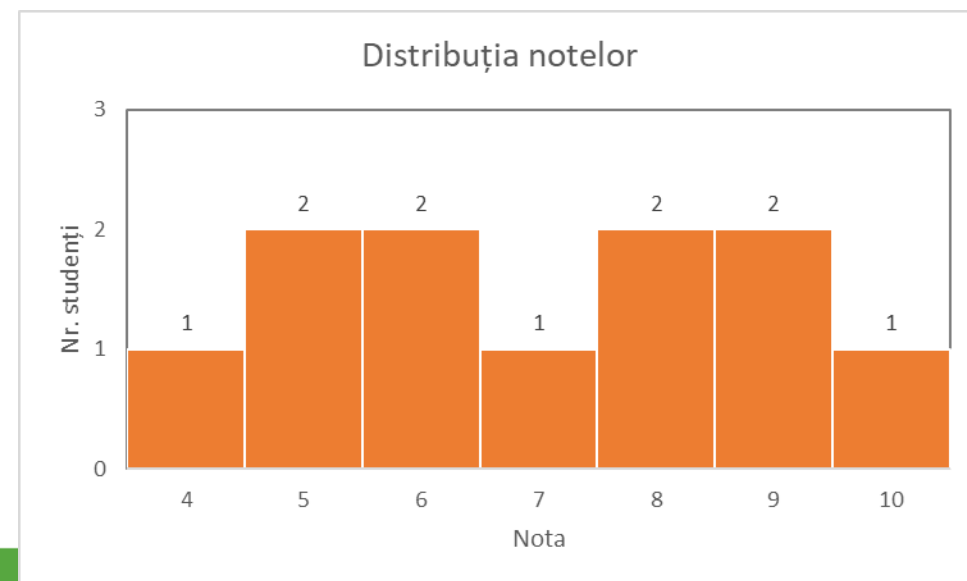
4, 9, 5, 8, 6, 7, 9, 10, 8, 6, 5

Media aritmetică =  $(4+9+5+8+6+7+9+10+8+6+5)/11 = 7$

Modulul: 5, 6, 8, 9 → serie multimodală

Mediana: 4, 5, 5, 6, 6, 7, 8, 8, 9, 9, 10

- $n$  (volumul eșantionului) = 11
- $Me = X_{(n+1)/2} = X_6 = 7$





# Indicatori de centralitate

Scorurile obținute la examenul practic de o grupă de 12 studenți:

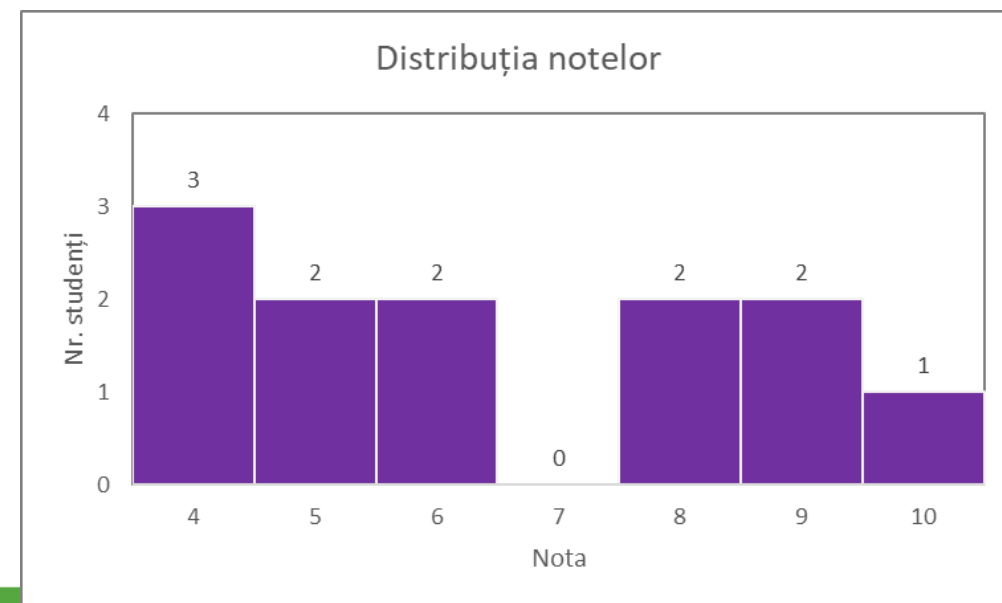
4, 9, 5, 8, 6, 4, 9, 10, 8, 6, 5, 4

Media =  $(4+9+5+8+6+4+9+10+8+6+5+4)/12 = 6.5$

Modulul: 4 → serie unimodală

Mediana: 4, 4, 4, 5, 5, 6, 6, 8, 8, 9, 9, 10

- $n$  (volumul eșantionului) = 12
- $Me = (X_{n/2} + X_{n/2+1})/2 = (X_6 + X_7)/2 = (6+6)/2 = 6$



# Indicatori de centralitate

## Media ponderată

$$m_x = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

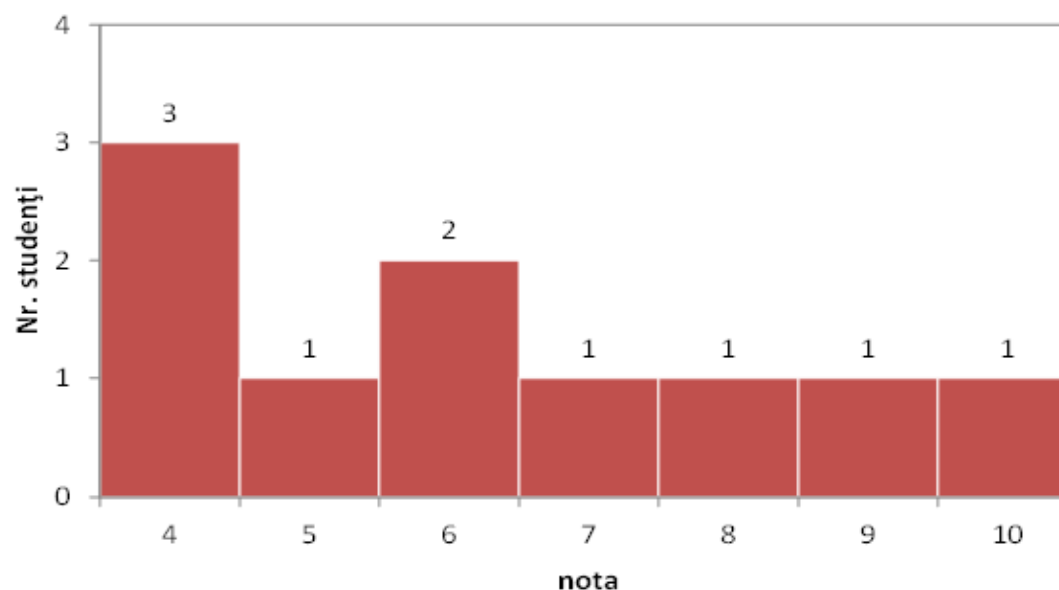
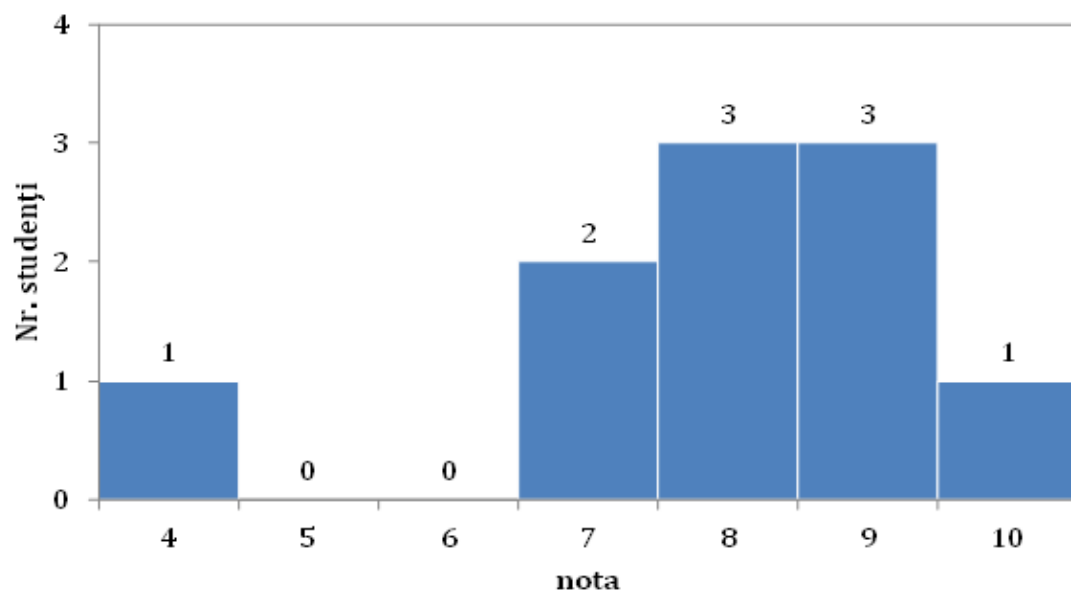
$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$$

Credits	6	7	6	4	2	2	2	2		
ID	Anatomie	Biofizică	Biologie celulară	Biostatistică și informatică medicală	Psihologie	Etică medicală	Comunicare medicală	Opțional	Mean	WMean
1	9	7	6	8	9	10	10	10	8.6	8.0
2	5	6	7	5	10	6	8	8	6.9	6.4
3	10	8	6	5	5	7	7	9	7.1	7.4
4	6	8	6	6	9	10	8	10	7.9	7.3
5	8	6	5	7	6	10	6	9	7.1	6.8
6	8	5	8	10	6	8	8	5	7.3	7.3
7	6	6	6	8	5	10	10	8	7.4	6.8
8	8	10	6	5	8	10	10	6	7.9	7.8
9	9	6	10	9	9	6	9	6	8.0	8.1
10	9	8	9	9	9	8	7	7	8.3	8.5
11	8	10	7	7	5	9	6	5	7.1	7.7
12	7	10	6	8	7	7	10	5	7.5	7.7
13	10	6	6	6	8	6	10	8	7.5	7.3
14	6	9	9	5	9	9	9	8	8.0	7.8
15	9	7	8	5	8	9	5	5	7.0	7.3

Media aritmetică este un caz special al mediei ponderate ( $W_i$ , ponderi egale).

# Indicatori de centralitate

Valoarea centrală =  $(x_{\min} + x_{\max})/2$



# Indicatori de centralitate

Denumire	Avantaj	Dezavantaj
<b>Media</b>	Utilizează toate datele Ușor de aplicat	Influențată de valori extreme Nerepresentativă dacă datele nu au o distribuție simetrică
<b>Mediana</b>	Nu e influențată de valori extreme Neinfluențată de asimetria datelor	Ignoră majoritatea datelor din serie
<b>Modulul</b>	Aplicabil și variabilelor calitative	Ignoră majoritatea datelor din serie
<b>Media geometrică</b>	Aplicabilă datelor asimetrice spre dreapta	Adecvată dacă logaritmarea datelor produce o distribuție normală
<b>Media ponderată</b>	Cuantifică importanța relativă a fiecărei observații	Ponderile trebuie să fie cunoscute sau estimate

# Indicatori de dispersie

Amplitudinea

Variația

Deviația standard

Eroarea standard

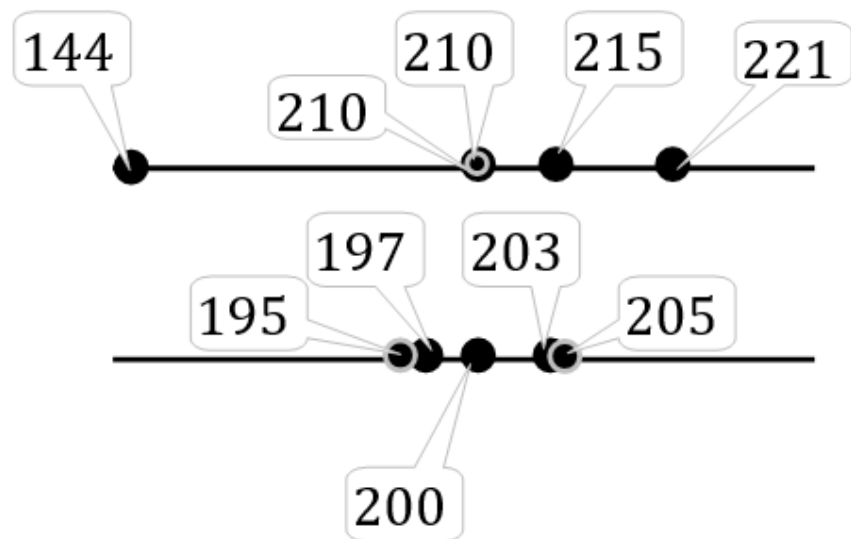
Coeficientul de variație

Variabilitate vs. Diversitate

# Indicatori de dispersie

Dispersie relativ la indicatorul de centralitate

Datele sunt cu atât mai dispersate cu cât valorile seriei sunt mai diferite unele față de altele!



Colesterol (mg/dL) – Metoda I

Colesterol (mg/dL) – Metoda II

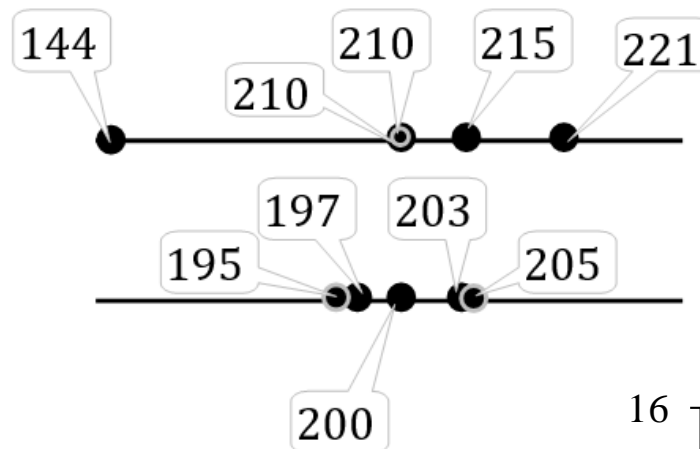
Care din cele două metode e mai precisă?

# Indicatori de dispersie: Amplitudinea

$$A = X_{\max} - X_{\min}$$

Metoda 1:  $A = 221 - 144 = 77$

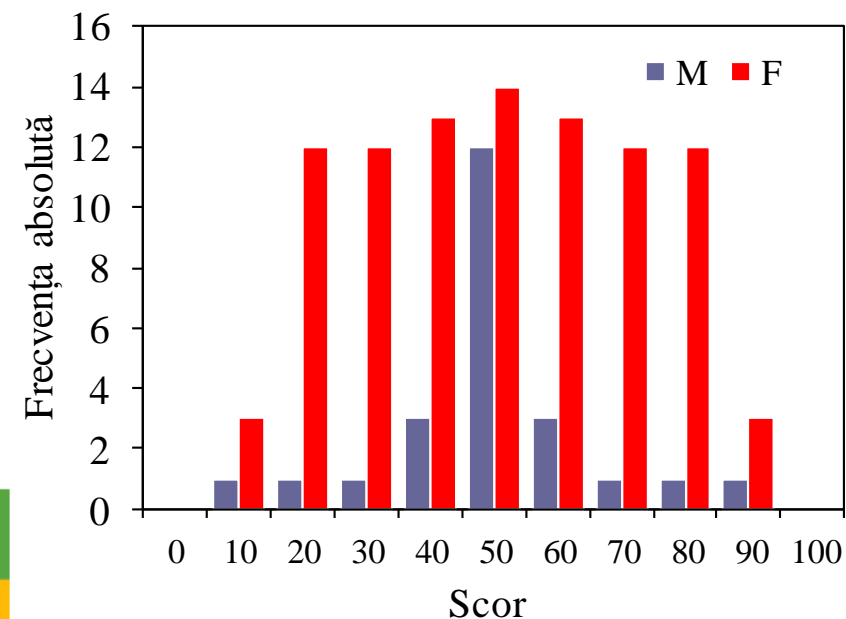
Metoda 2:  $A = 205 - 195 = 10$



Colesterol (mg/dL) – Metoda I

Colesterol (mg/dL) – Metoda II

- Consideră exclusiv valorile extreme
- Nu oferă informații cu privire la distribuția datelor între valorile extreme



# Indicatori de dispersie: Media deviației

## Metoda 1

Media abaterii de la medie = 22,4

Media abaterii de la mediană = 16,4

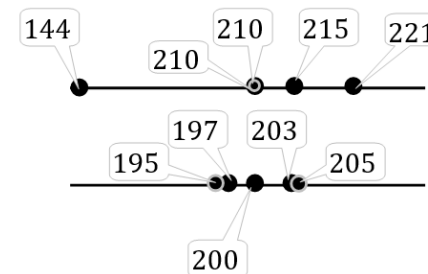
$$\frac{\sum_{i=1}^n |x_i - \bar{X}|}{n}$$

## Metoda 2

Media abaterii de la medie = 3,2

Media abaterii de la mediană = 3,2

$$\frac{\sum_{i=1}^n |x_i - Me|}{n}$$



Colesterol (mg/dL) – Metoda I

Colesterol (mg/dL) – Metoda II

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n}$$

$$S^2 = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n-1}$$

**Variația eșantionului**

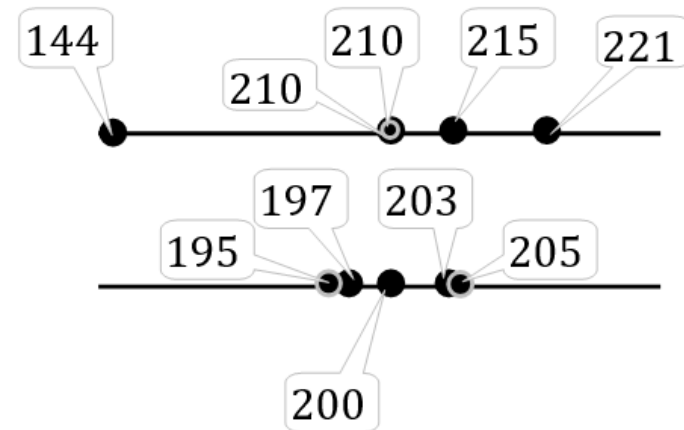
**Variația de eșantionare / VARIANȚA**



# Indicatori de dispersie: Media deviației

Rădăcina pătratică a variației:  $s$

Abaterea standard sau deviația standard (de eșantionare):  $S$



Colesterol (mg/dL) – Metoda I

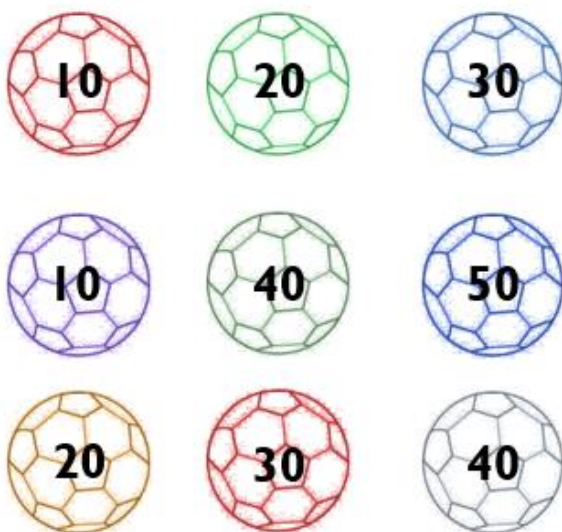
Colesterol (mg/dL) – Metoda II

	metoda 1	metoda2
	144	195
	210	197
	210	200
	215	203
	221	205
media	200	200
mediana	210	200
var	1000.5	17
stdev	31.63	4.12

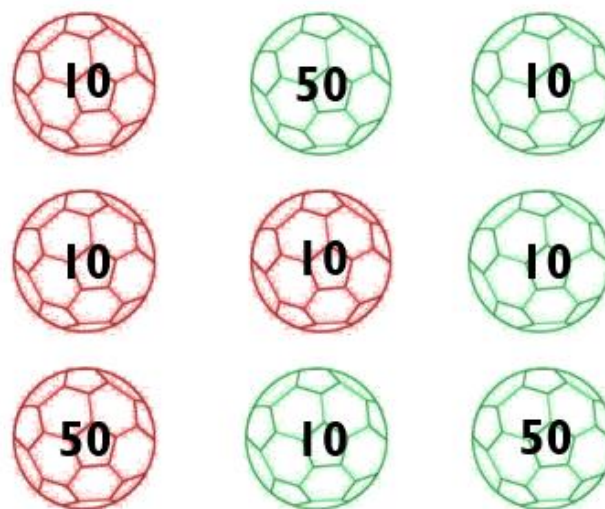
# Indicatori de dispersie: Variabilitate vs. Diversitate

1. Care din grupurile de mai jos prezintă o compoziție a culorilor mai diversă?
2. Care din grupurile de mai jos prezintă o variabilitate mai mare a orelor de utilizare?

**Grup 1**



**Grup 2**



**Grup 1 →  $s = 15,81$**

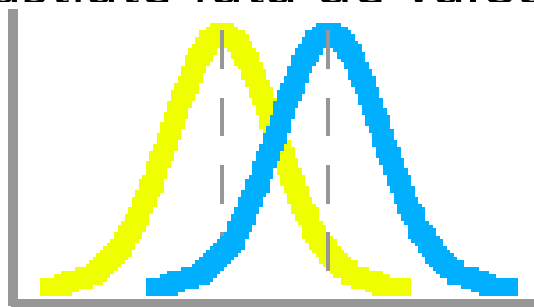
**Grup 2 →  $s = 21,91$**

# Indicatori de dispersie: deviația standard

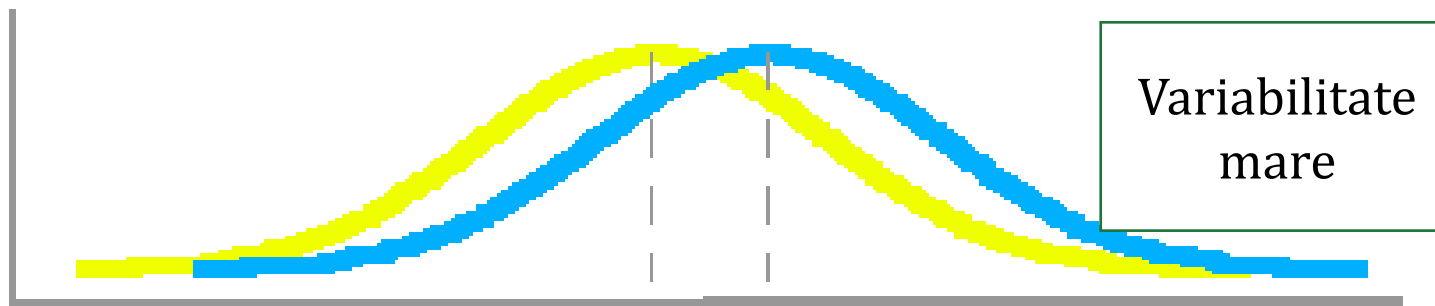
↓  $s$  → datele seriei au valori apropiate de ale mediei (variabilitate mică)

↑  $s$  → datele seriei sunt împrăștiate față de valoarea medie (variabilitate mare)

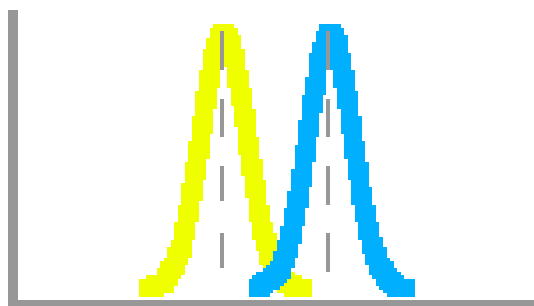
Variabilitate  
medie



Variabilitate  
mare



Variabilitate  
mică



# Indicatori de dispersie: eroarea standard

*SEM = standard error of the mean*

Indicator al acurateții mediei aritmetice:  $SEM = s/\sqrt{n}$

Relație invers proporțională dintre volumul eșantionului și SEM

Utilizat în statistica inferențială (intervalul de încredere)



# Indicatori de dispersie: variația și deviația standard

Populație → Parametrul

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$$

**VAR.P**

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}$$

**STDEV.P**

Eșantion → Statistica

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n - 1}$$

**VAR.S = VAR**

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n - 1}}$$

**STDEV.S=STDEV**

# Indicatori de dispersie: CV (coeficientul de variație)

Dispersie relativă

Nu are unitate de măsură

Se calculează doar pentru variabilele cantitative de tip rație/raport (valori pozitive)

Compararea variabilității a două populații când unitățile de măsură sunt diferite (mg/dL vs mmol/L – colesterol)

$$CV(\%) = \frac{s}{\bar{X}} \times 100$$

$$CQV(\%) = \frac{(Q_3 - Q_1)}{(Q_3 + Q_1)} \times 100$$

Omogen

CV < 10%

Relative omogen

10% ≤ CV < 20%

Relativ eterogen

20% ≤ CV < 30%

Eterogen

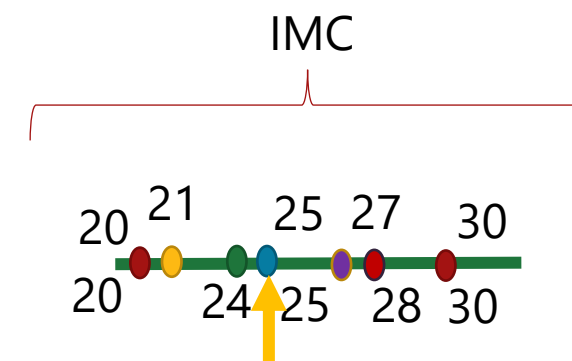
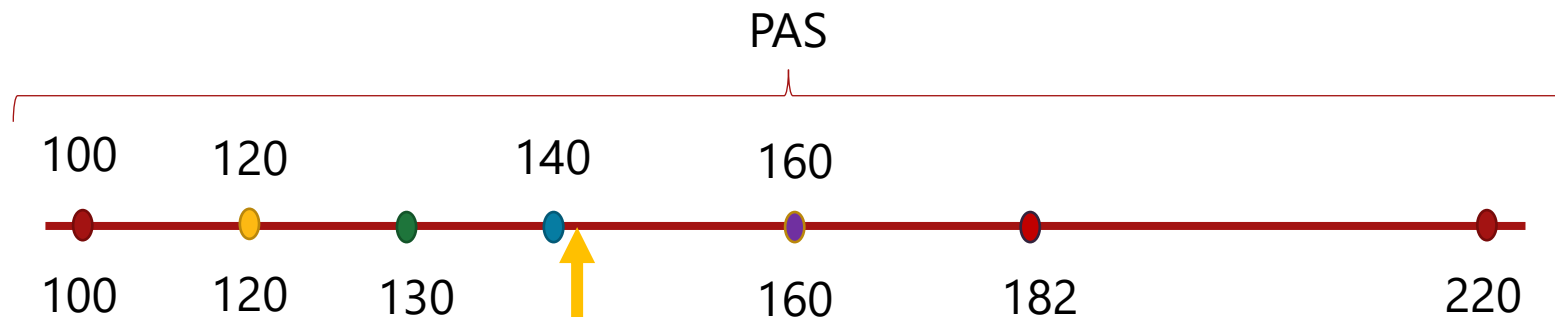
CV ≥ 30%

# Indicatori de dispersie: CV

Se dorește evaluarea variabilității presiunii arteriale sistolice (PAS) comparativ cu cea a indicelui de masă corporală (IMC). Datele aparținând la 10 subiecți sunt redată în tabelul de mai jos:

	1	2	3	4	5	6	7	8	9	10
PAS (mmHg)	220	182	100	130	120	100	140	160	120	160
IMC (kg/m <sup>2</sup> )	28	30	21	27	25	20	30	25	24	20

- PAS:  $CV(\%) = 36,58/144 \times 100 = 25,40$
- IMC:  $CV(\%) = 3,8/25 \times 100 = 15,20$



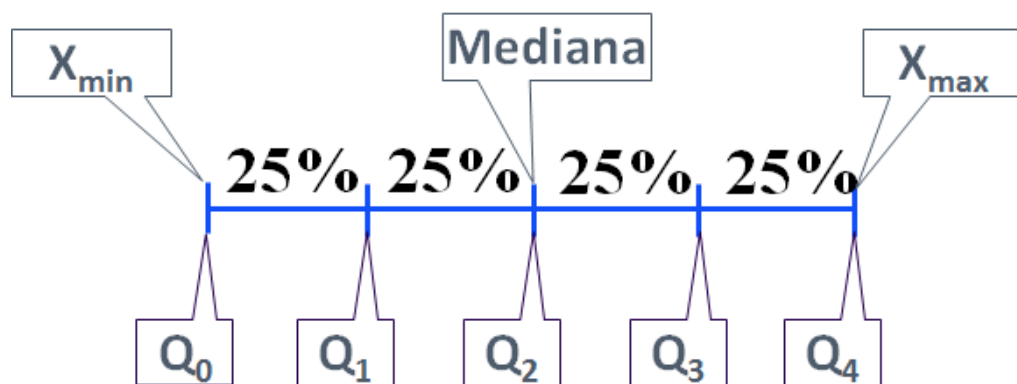
# Indicatori de localizare (dispersie): cvartile

$$IC = Q_3 - Q_1$$

unde  $Q_3$  = cuartila 3 (percentila 75%),  $Q_1$  = cuartila 1 (percentila 25%)

IC = intervalul cuartilic

Măsură a dispersiei pentru 50% din datele de mijloc.



QUARTILE.EXC

QUARTILE.INC

QUARTILE = QUARTILE.INC



# Indicatori de localizare (dispersie): cuartile

Variabila:  
note la  
examenul  
practic

$x_1$	9
$x_2$	6
$x_3$	4
$x_4$	9
$x_5$	4
$x_6$	8
$x_7$	8
$x_8$	9
$x_9$	7
$x_{10}$	4
$x_{11}$	10
$x_{12}$	10

Ordonare

	A	B
1	$x_{10}$	4
2	$x_3$	4
3	$x_5$	4
4	$x_2$	6
5	$x_9$	7
6	$x_6$	8
7	$x_7$	8
8	$x_1$	9
9	$x_4$	9
10	$x_8$	9
11	$x_{11}$	10
12	$x_{12}$	10

Me=8

$$Q_3 - Q_1 = 9 - 5.5 = 3.5$$

$$A = 10 - 4 = 6$$

$$Me = [X_{12/2} + X_{(12/2+1)}] / 2 = (X_6 + X_7) / 2 = (8 + 8) / 2 = 8$$

**Formule Excel:**

**(Mediana) Me:**

$$= \text{MEDIAN}(B1:B12)$$

**(Intervalul dintre cvartila 3 și 1)**

**IC:**

$$= \text{QUARTILE}(B1:B12, 3) - \text{QUARTILE}(B1:B12, 1)$$

**(Amplitudinea) A:**

$$= \text{MAX}(B1:B12) - \text{MIN}(B1:B12)$$

# Indicatori de localizare (dispersie): cuartile

	A	B
1	$X_{10}$	4
2	$X_3$	4
3	$X_5$	4
4	$X_2$	6
5	$X_9$	7
6	$X_6$	8
7	$X_7$	8
8	$X_1$	9
9	$X_4$	9
10	$X_8$	9
11	$X_{11}$	10
12	$X_{12}$	10

**Me=8**

$$Q_3 - Q_1 = 9 - 5.5 = 3.5$$

$$A = 10 - 4 = 6$$

**Q1:** 25% din studenți au note  $< 5.5$

**Q3:** 75% din studenți au note  $\leq 9$

**IC:** 50% din studenți au note care nu diferă una față de alta cu mai mult de 3,5 puncte

**A:** Diferența dintre nota maximă și nota minimă a fost de 6 puncte

# Indicatori de dispersie

Estimator	Avantaj	Dezavantaj
<b>Amplitudinea</b>	Ușor de calculat	Utilizează doar datele a două observații Influențată de valori extreme Tinde să aibă valoare mai mare odată cu creșterea volumului eșantionului
<b>IC</b>	Neafectată de valorile extreme Independentă de volumul eșantionului Adecvată pentru date cu distribuție asimetrică	Dificil de calculat Neadekvată pentru eșantioane mici Utilizează doar două observații
<b>Variația</b>	Utilizează toate datele seriei statistice	Influențată de valorile extreme Neadekvată pentru datele asimetrice
<b>Deviația standard</b>	Utilizează toate datele seriei statistice Unitatea de măsură este identică cu cea a datelor brute Se interpretează ușor	Influențată de valorile extreme Neadekvată pentru datele asimetrice
<b>CV</b>	Independent de unitățile de măsură Permite compararea serilor statistice cu diferite unități de măsură	Media aproape de zero → CV se apropie de infinit fiind sensibil la modificări mici ale mediei

# Indicatori de simetrie și boltire

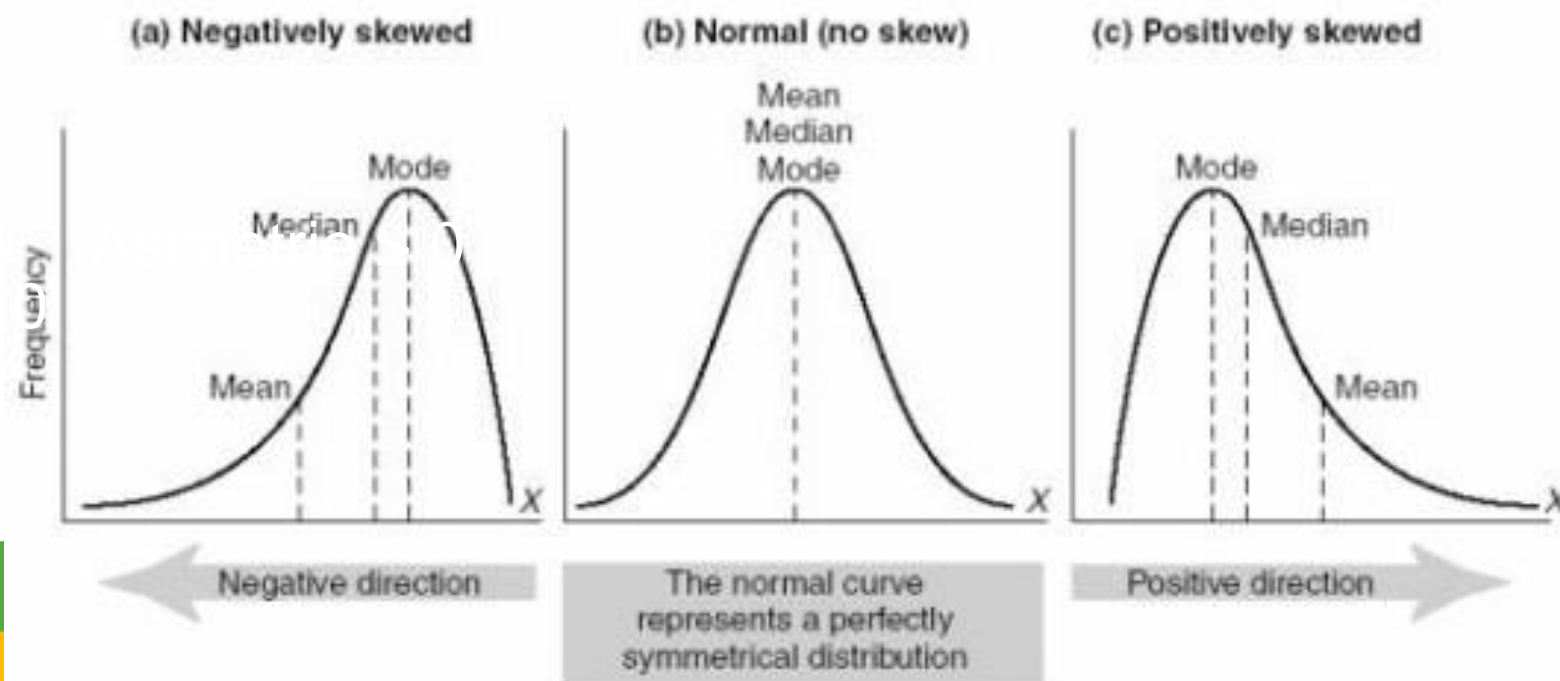
**Intepretarea** – aplicată populației

Asimetria  $< -1$  sau  $> +1 \rightarrow$  distribuție înalt asimetrică

Asimetria  $< -1 \rightarrow$  asimetrie la stânga / pozitivă

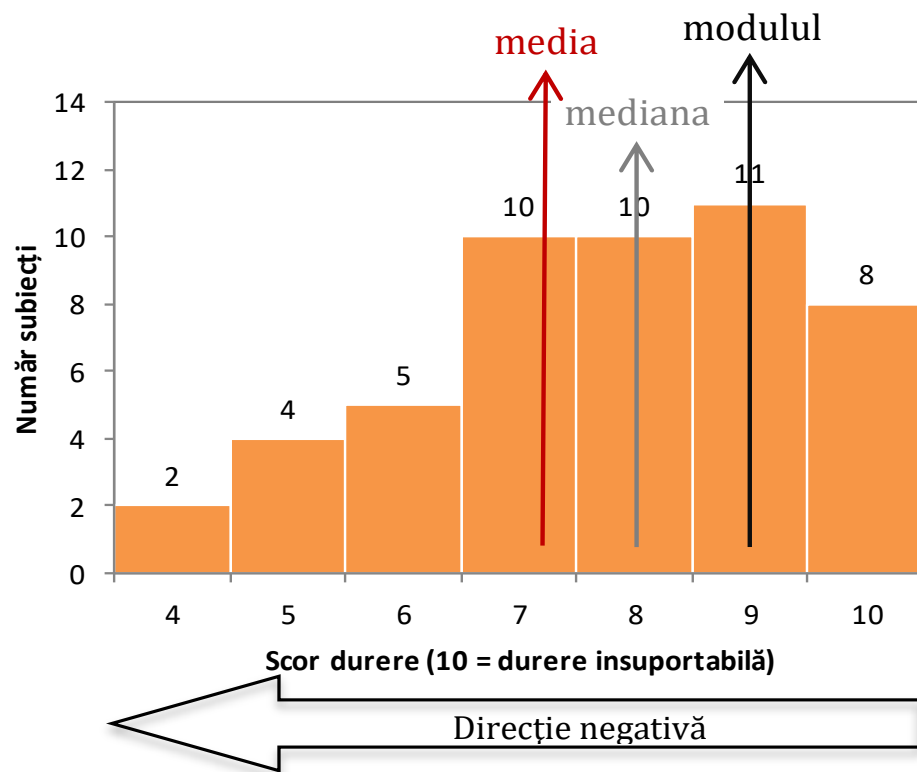
Asimetria  $> +1 \rightarrow$  asimetrie la dreapta / negativ

**SKEW**  
**SKEW.P**



# Indicatori de simetrie și boltire

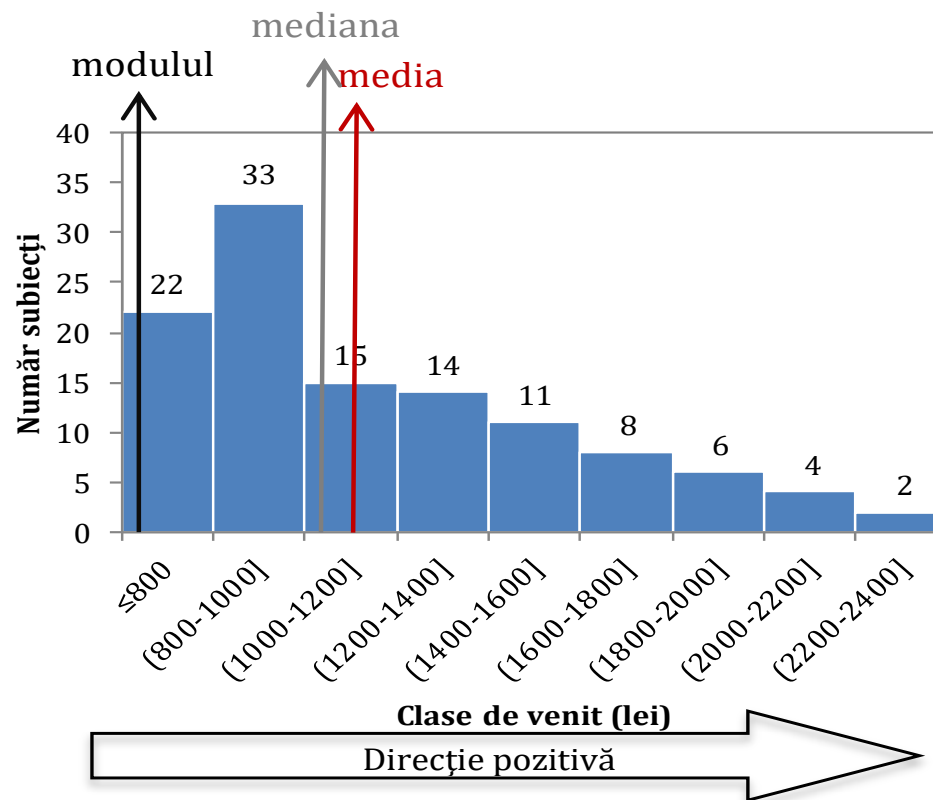
$$\text{SKEW}(\text{array}) < -1$$



Asimetrie la stânga

**Modulul > Mediana > Media**

$$\text{SKEW}(\text{array}) > 1$$



Asimetrie la dreapta

**Modulul < Mediana < Media**

# Indicatori de simetrie și boltire

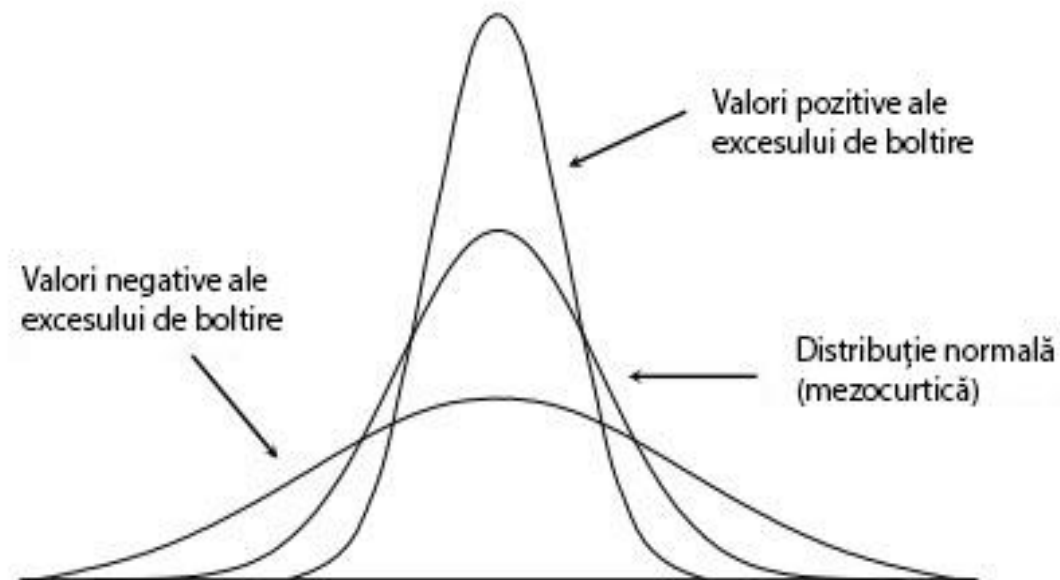
Distribuția normală are boltirea în jurul valorii 3.

Excesul de boltire (funcția KURT) =  
(Boltire – 3)

Distribuția normală: boltirea  $\cong 3$  (excesul de boltire  $\cong 0$ )  $\rightarrow$  mezocurtic

Distribuția cu boltirea  $< 3$  (excesul de boltire  $< -1$ )  $\rightarrow$  platicurtică (aplatizată)

Distribuția cu boltirea  $> 3$  (excesul de boltire  $> 1$ )  $\rightarrow$  leptocurtică (ascuțită)



## KURT

# Indicatori de simetrie

Examinarea simetriei unei distribuții cu ajutorul cuartilelor

$Q_2 - Q_1 \approx Q_3 - Q_2$  ( $\approx$  aproape egal)  $\rightarrow$  distribuția este aproximativ simetrică

$Q_2 - Q_1 \neq Q_3 - Q_2 \rightarrow$  distribuția este asimetrică

2,80	2,97	3,05	3,25	3,40	3,45	3,80	4,10	4,30	4,40
$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$	$x_8$	$x_9$	$x_{10}$

- $Q_1 = 3,03$

- $Q_2 = 3,43$

- $Q_3 = 4,15$

$$Q_2 - Q_1 = 3,43 - 3,03 = 0,40$$

$$Q_3 - Q_2 = 4,15 - 3,43 = 0,72$$

Cum interpretăm acest rezultat?

# Tipuri de grafice

## Histograma

*1 variabilă cantitativă*  
(distribuție)

## Cutia cu mustăți

1 variabilă cantitativă +  
1 variabilă de grupare  
(nominală sau ordinală)  
(distribuție)

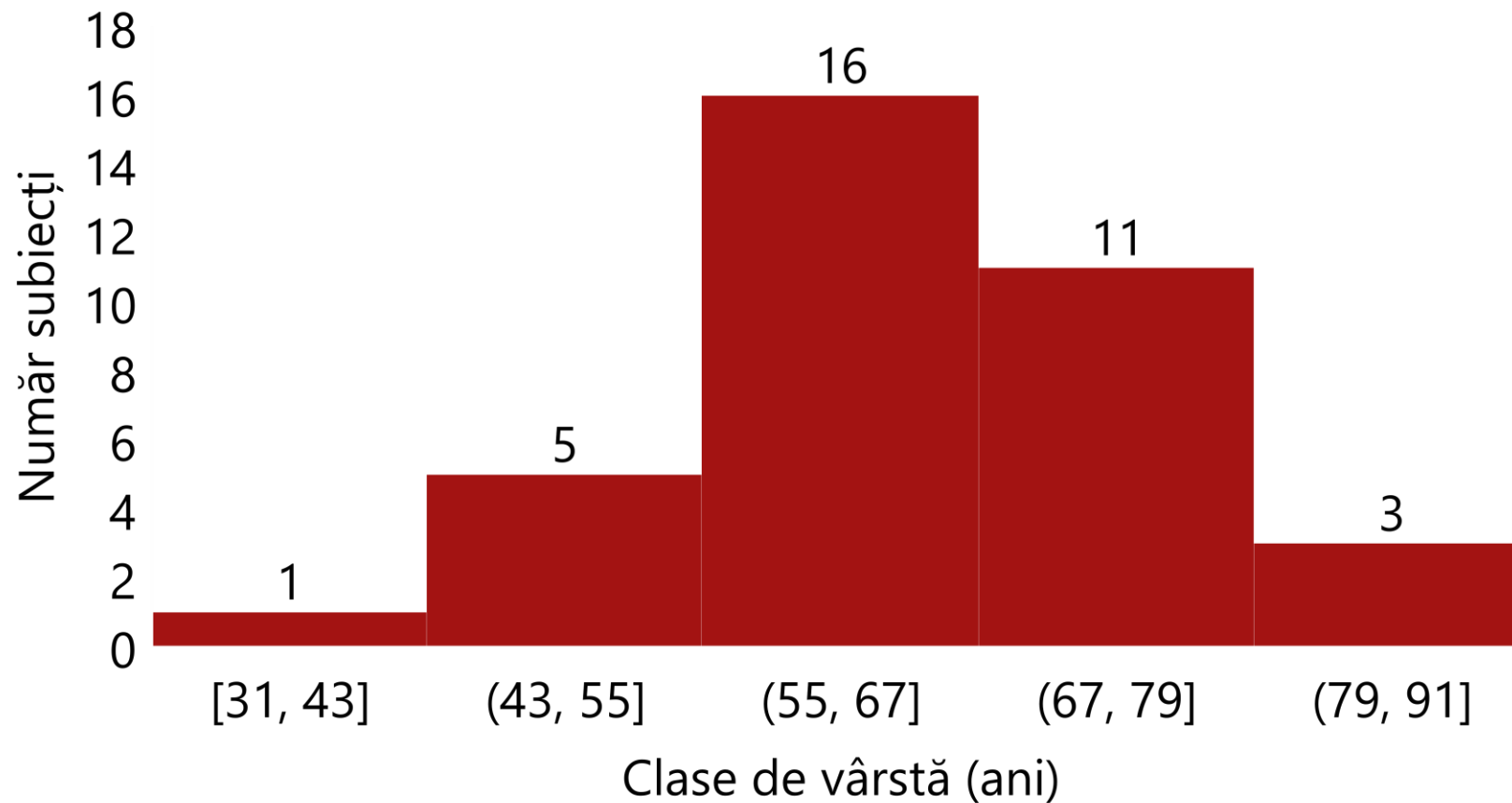
## Nor de puncte

2 variabile cantitative  
(asociere / relație)



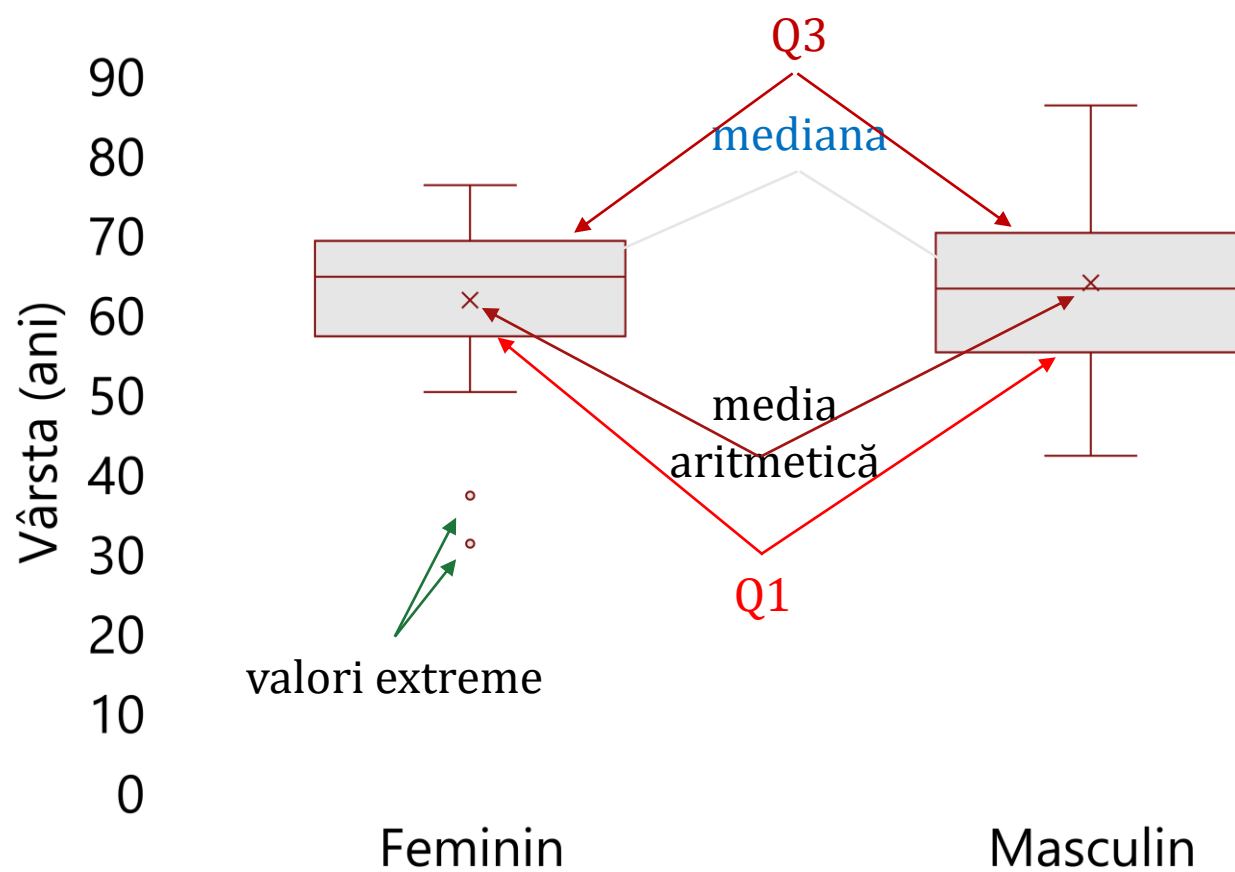
# Histograma & Poligonul de frecvență

Histograma vârstei

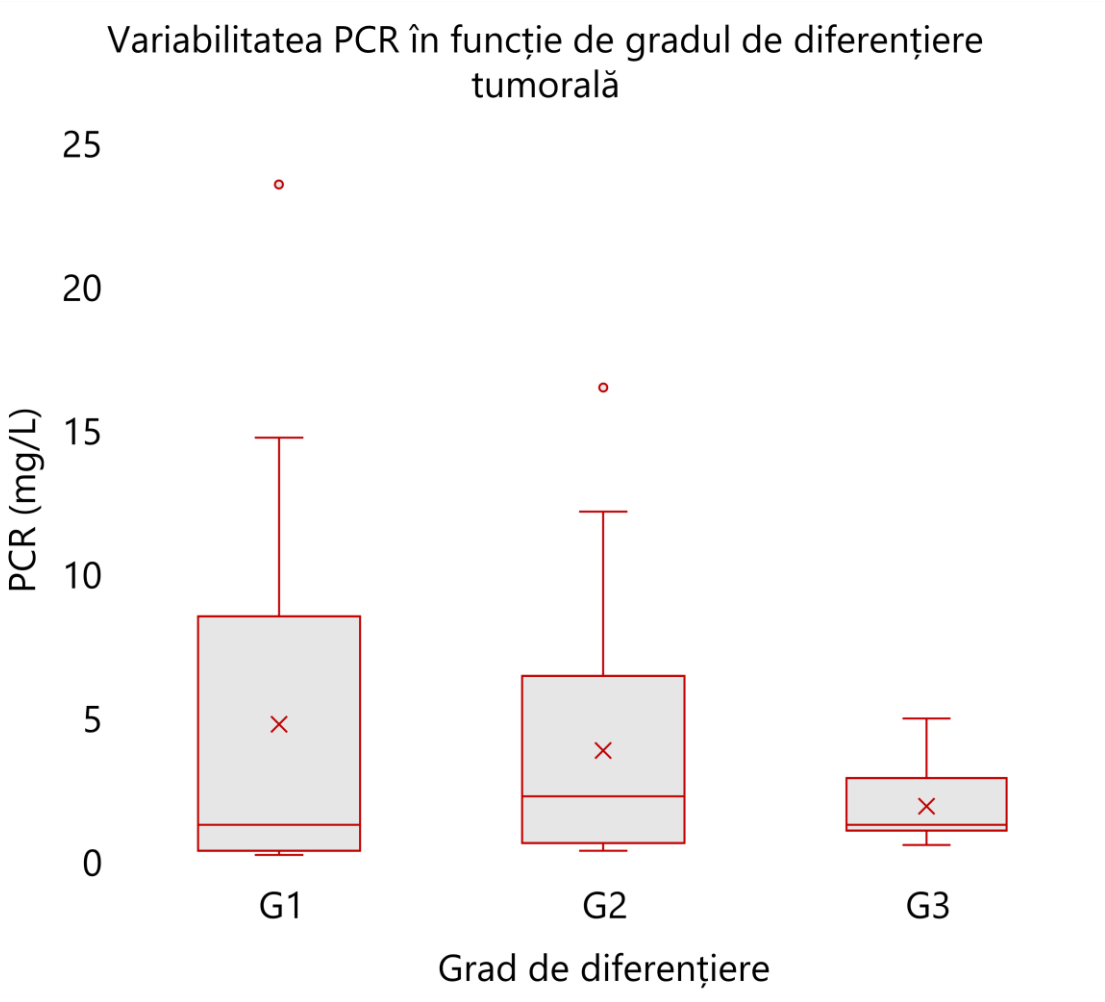


Media	64.5
Mediana	65
Modulul	65
Deviați standard	10.9
Q1	59
Q3	70
min	31
max	86
n	36
CV (%)	16.9
CQV (%)	8.5

# Cutie cu mustăți / Box & Whisker



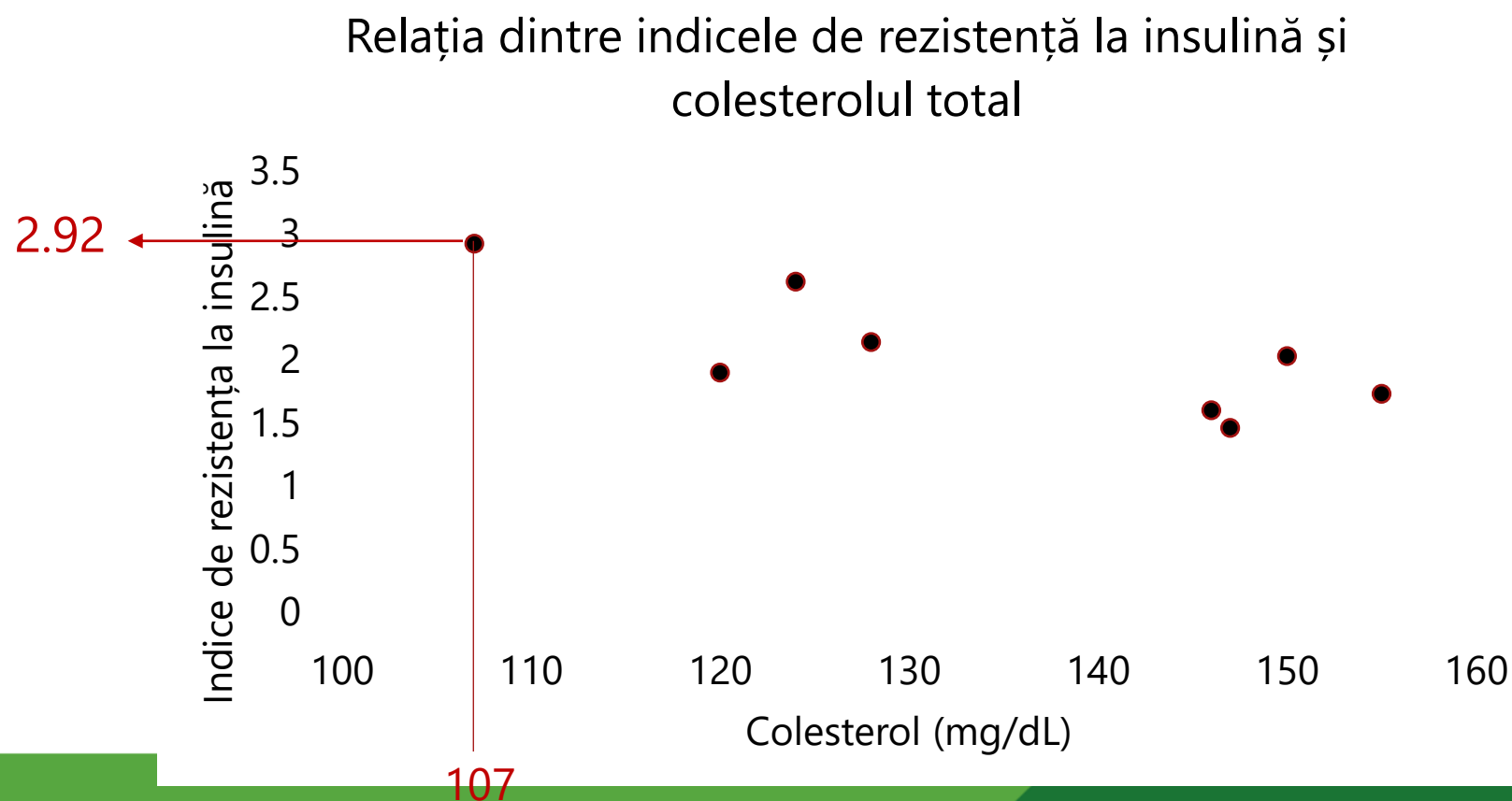
# Cutie cu mustăți / Box & Whisker



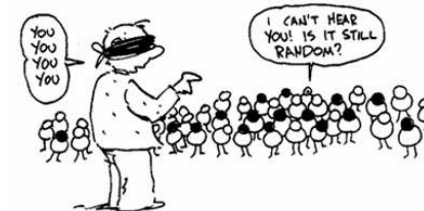
	G1	G2	G3
media	4.7	3.8	2.0
mediana	1.2	2.2	1.25
modulul	0.3	2	#N/A
deviatia standard	7.1	4.3	1.6
Q1	0.31	0.7	1.0
Q3	3.39	6.0	2.5
min	0.16	0.3	0.5
max	23.48	16.41	4.9
n	17	20	6
CV(%)	151.5	114.7	84.1
CQV(%)	83.2	78.3	41.4

# Graficul de tip Nor de puncte (*scatter*)

Colesterol (mg/dL)	Indice de rezistență la insulină
146	1.6
155	1.73
107	2.92
128	2.14
120	1.9
150	2.03
147	1.46
124	2.62



# IMPORTANT



Variabilele calitative se sumarizează ca:

- centralitate → modulul (scala nominală / ordinală -  $\leq 5$  categorii) & mediana [Q1-Q3] (scala ordinală)

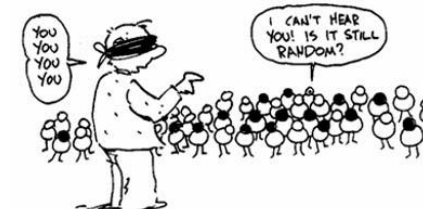
Variabilele cantitative se sumarizează ca

- medie  $\pm$  deviație standard & CV dacă urmează distribuția normală
- mediană și interval cuartilic [Q1-Q3] & CQV dacă nu urmează distribuția normală (aplicat și pentru datele calitative scala ordinală)

Dacă datele urmează distribuția normală, mă aștept ca

- media  $\approx$  mediana  $\approx$  modulul
- asimetria și boltirea  $\in [-1; +1]$

# IMPORTANT



Tip	Scala de măsură	Indicator centralitate	Indicator dispersie
Calitativ	Nominal	Modulul	
	Ordinal	Mediana ( $> 5$ categorii) — Modulul ( $\leq 5$ categorii) —	Intervalul cuartilic [Q1 – Q3] Valoarea minimă și maximă {min – max}
Cantitativ	Interval/rație	Media	Variația & Deviația standard
		Mediana	Intervalul cuartilic
		Modulul	Valoarea minimă și maximă {min – max}

# De reținut!

Variabilelor calitative ordinale se tratează ca variabile cantitative discrete dacă pot lua > 5 categorii distincte

Sumarizăm variabilele cantitative ca medie  $\pm$  deviație standard dacă datele urmează distribuția normală

Sumarizăm variabilele cantitative ca mediană [Q1 – Q3] datele NU urmează distribuția normală.

Nu există rețeta universală pentru reprezentarea tabelară / grafică

Histograma prezintă distribuția unui grup

Graficul de tip cutie cu mustăți permite vizualizarea distribuției mai multor grupuri

Asocierea dintre 2 variabile (ordinale cu peste 5 categorii sau cantitative) se vizualizează cu graficul tip nor de puncte