Autor: Bondor Cosmina-Ioana,

Sumarizarea si prezentarea datelor

A ALWAYS

S SEEK

K KNOWLEDGE

Obiective

- Măsuri de centralitate
- Măsuri de dispersie
- Grafice
- Exerciţii

Scenariu

- Colectăm datele despre 20 de pacienți dintr-un cabinet de ultrasonografie.
- Vrem să extragem informații din date.

Exemple

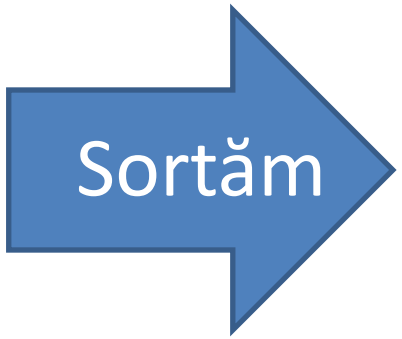
	A	B	C	D	E	F	G	H	I	J	K	L
1	Grup	Gen	Varsta mamei	Greutate copil la nastere	Scor Apgar	Etnie	Numar saptamani de sarcina	Numar nasteri mama	Numar sarcini	Cezariana	Perimetru cranian	Inaltime
2	1	M	22	Normala	10	maghiara	40	2	2	Nu	21	54
3	1	M	25	Normala	9	maghiara	41	0	3	Nu	21	55
4	1	F	32	Supraponderal	10	romana	39	0	2	Da	19	60
5	1	F	28	Supraponderal	10	romana	41	0	0	Nu	23	50
6	1	F	25	Subponderal	5	romana	34	2	3	Nu	17	45
7	1	F	26	Supraponderal	9	romana	41	0	4	Nu	21	60
8	1	M	31	Normala	10	romana	41	0	0	Da	21	59
9	1	F	35	Subponderal	6	roma	36	3	3	Nu	18	45
10	1	M	26	Normala	10	maghiara	41	1	1	Nu	21	60
11	1	M	24	Normala	10	romana	39	0	0	Nu	21	62
12	1	F	25	Normala	10	romana	41	1	1	Da	21	57
13	1	F	27	Normala	10	romana	41	0	8	Nu	22	59
14	1	F	29	Normala	8	romana	41	1	1	Nu	21	55
15	1	F	30	Normala	10	romana	40	1	1	Nu	17	57
16	1	F	26	Supraponderal	7	romana	41	2	2	Da	22	60
17	1	M	21	Normala	8	romana	39	1	4	Nu	21	56
18	1	F	29	Normala	10	romana	41	1	1	Nu	20	59
19	1	M	33	Subponderal	7	romana	28	0	1	Nu	18	48
20	0	F	41	Subponderal	10	romana	29	0	1	Nu	17	46
21	0	F	28	Normala	9	romana	40	0	0	Nu	21	60

Indicatori pentru variabile ordinale

E
Scor Apgar
10
9
10
10
5
9
10
6
10
10
10
10
8
10
7
8
10
7
10
9

?

E
Scor Apgar
10
9
10
10
5
9
10
6
10
10
10
10
8
10
7
8
10
7
10
9



Scor Apgar
5
6
7
7
8
8
9
9
9
10
10
10
10
10
10
10
10
10
10
10
10

Indicatori pentru variabile ordinale

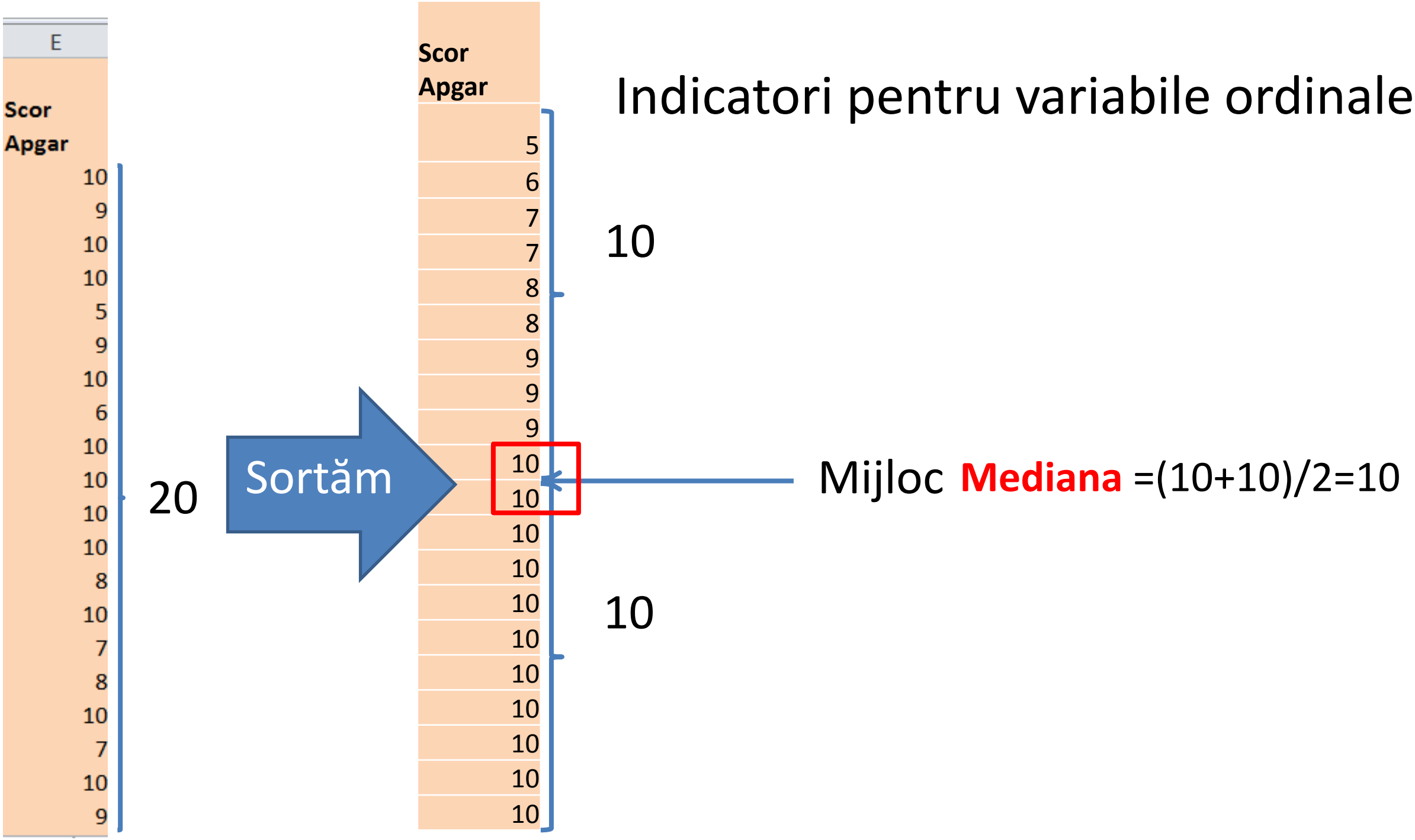
Minim = 5

Maxim = 10

Amplitudinea = Maxim – Minim =

= 10 - 5 = 5

Indicatori pentru variabile ordinale



10

10

Mijloc **Mediana** =(10+10)/2=10

Indicatori pentru variabile ordinale

Mediana – mijlocul valorilor observate,

Valoarea de la care jumătate din observații sunt mai mici și jumătate sunt mai mari.

Calcul:

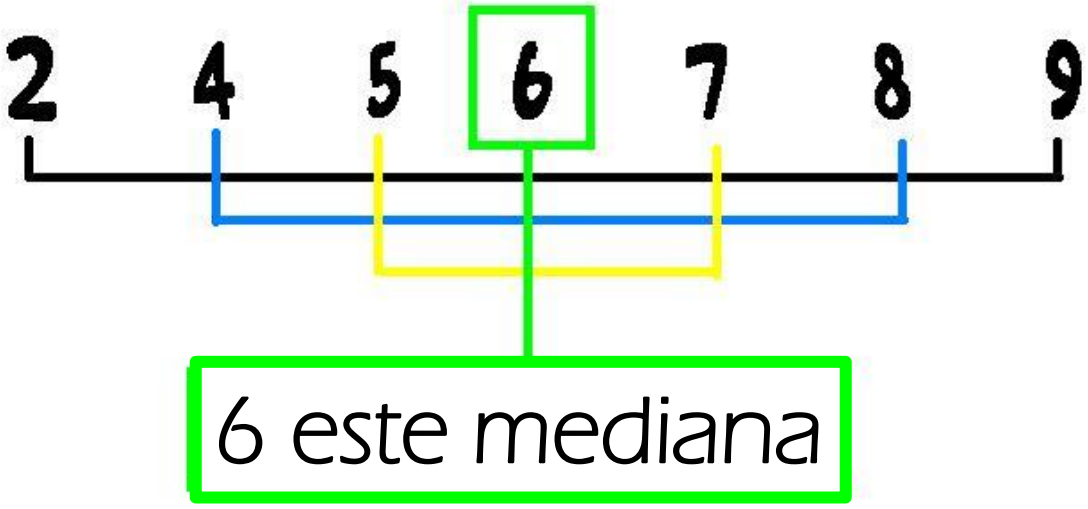
Aranjați observațiile de la cel mai mic la cel mai mare

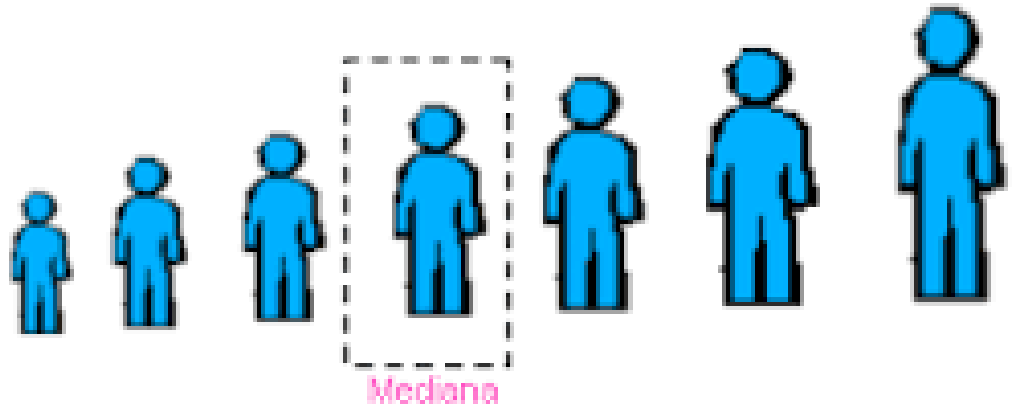
Găsiți mijlocul prin numărare:

număr impar de observații - mediana = valoarea aflată la mijlocul setului de date

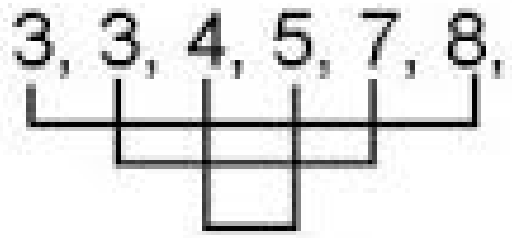
numărul par de observații – mediana = media celor două valori aflate la mijlocul setului de date.

Număr impar de observații





Număr par de observații

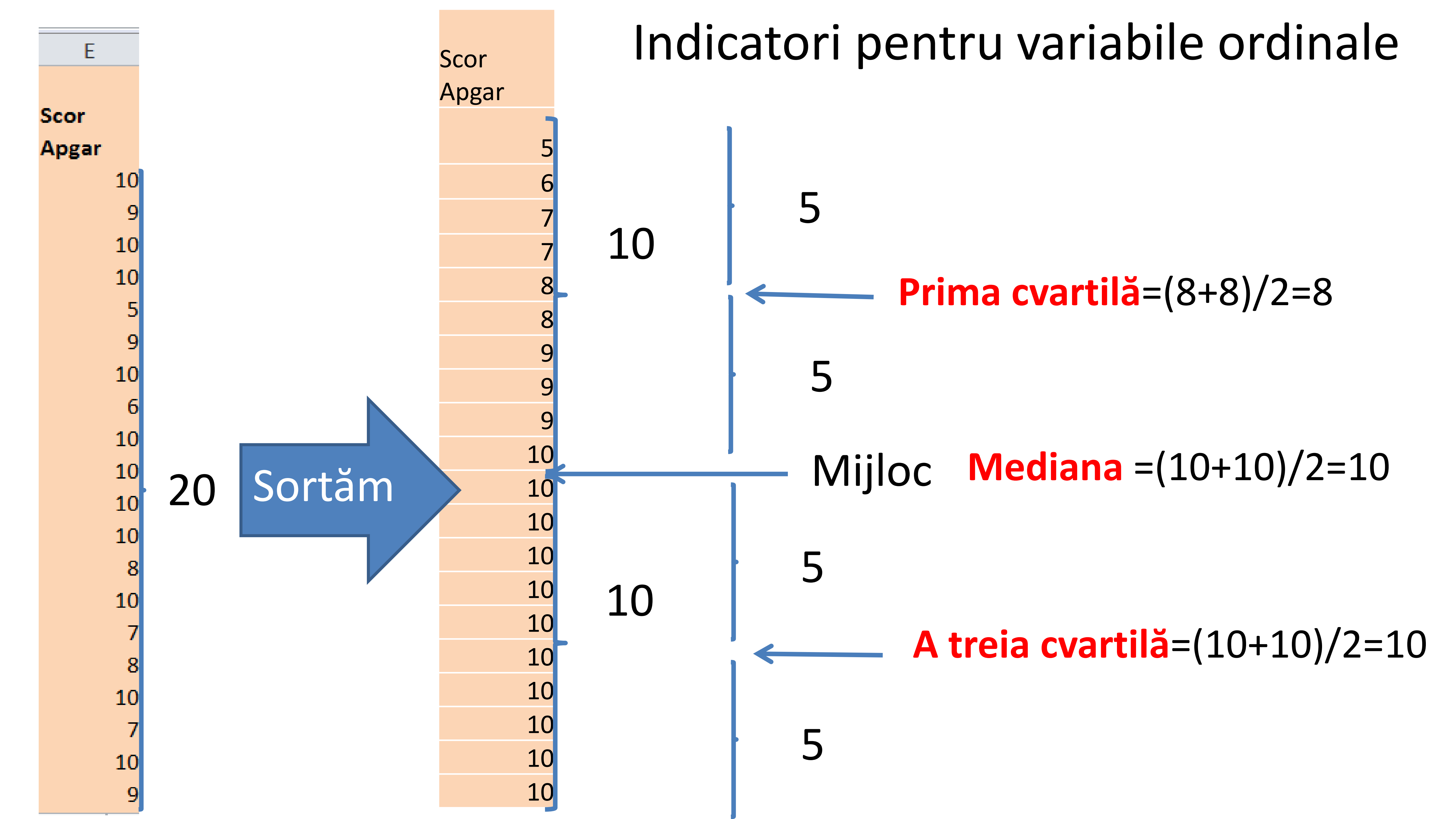


4+5=9

9/2=4,5

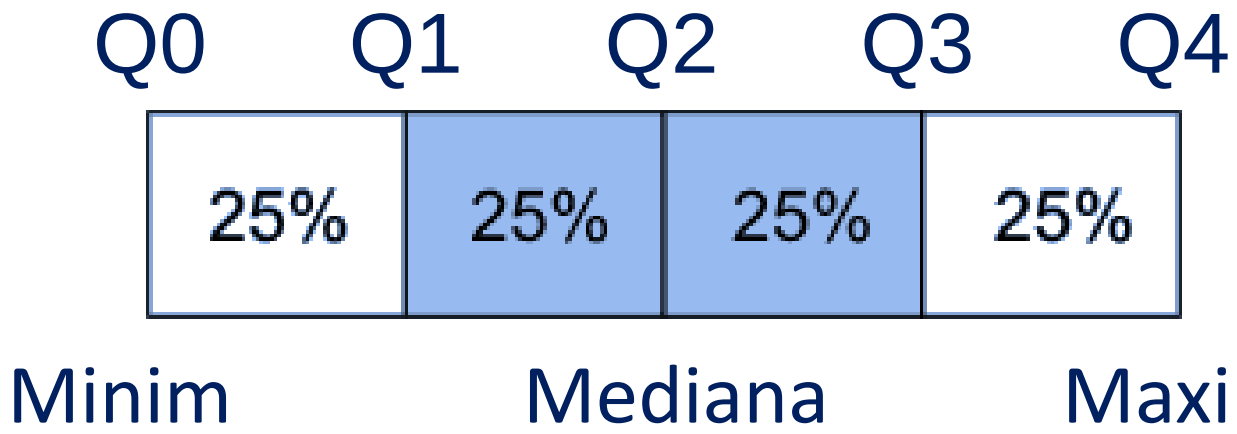
4,5 este mediana

$$MED + IAN \over 2$$


Indicatori pentru variabile ordinale
E
Scor Apgar
10
9
10
10
5
9
10
6
10
10
10
10
8
10
7
8
10
7
10
9
20
Sortăm
Scor Apgar
5
6
7
7
8
8
9
9
9
10
10
10
10
10
10
10
10
10
10
10
10
5
5
Prima cvartilă=(8+8)/2=8
Mijloc
Mediana =(10+10)/2=10
10
5
5
A treia cvartilă=(10+10)/2=10

Cvartile

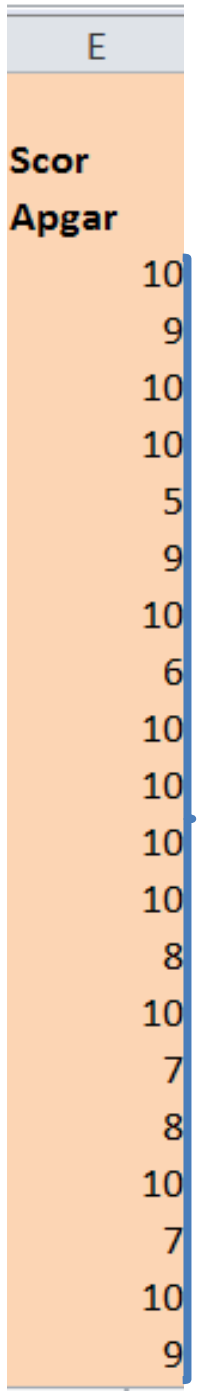
- Obs. A doua cvartilă (Q2) = Mediana
- Q = cvartile



Cvartile

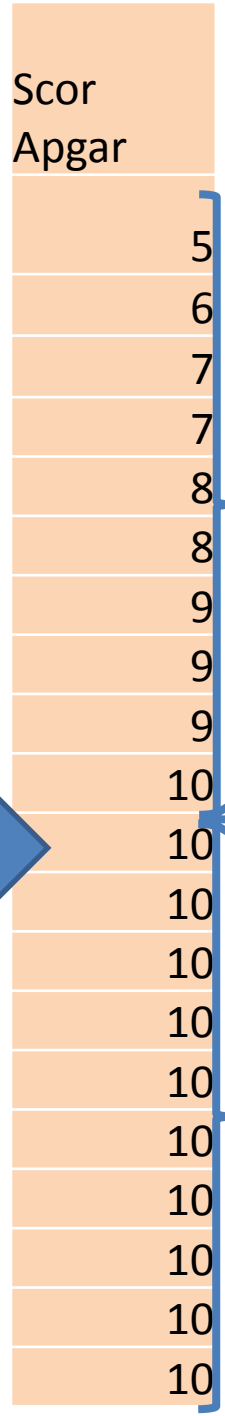
- **Cvartila 0** = minim
- **Cvartila 1**
- **Cvartila 2** = mediana
- **Cvartila 3**
- **Cvartila 4** = maxim

Indicatori pentru variabile ordinale



20

Sortăm



10

10

Q0 = minim=5

5

25%

Prima cvartilă=(8+8)/2=8

5

25%

Mijloc **Mediana** =(10+10)/2=10

5

25%

A treia cvartilă=(10+10)/2=10

5

25%

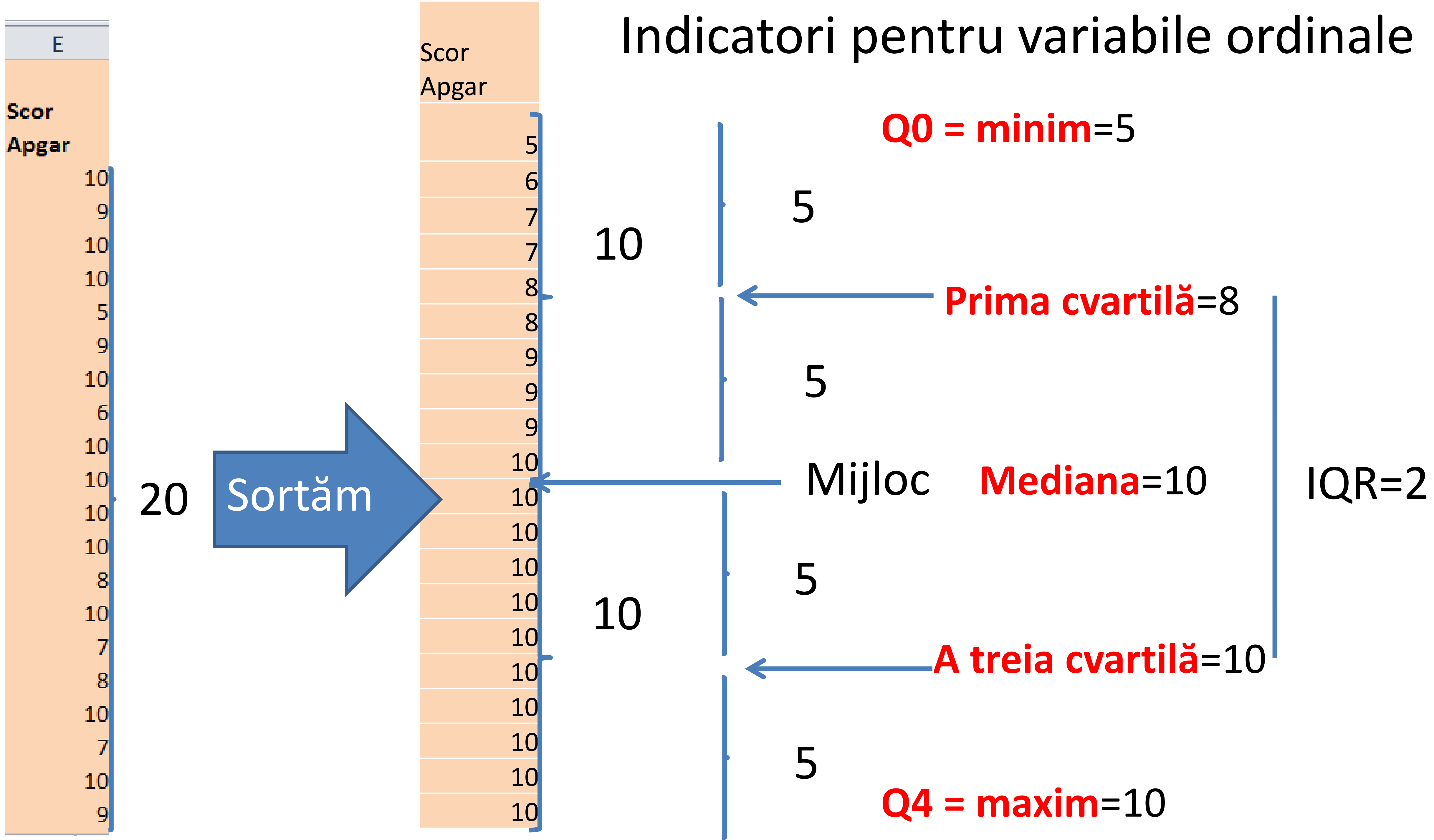
Q4 = maxim=10

Alte măsuri

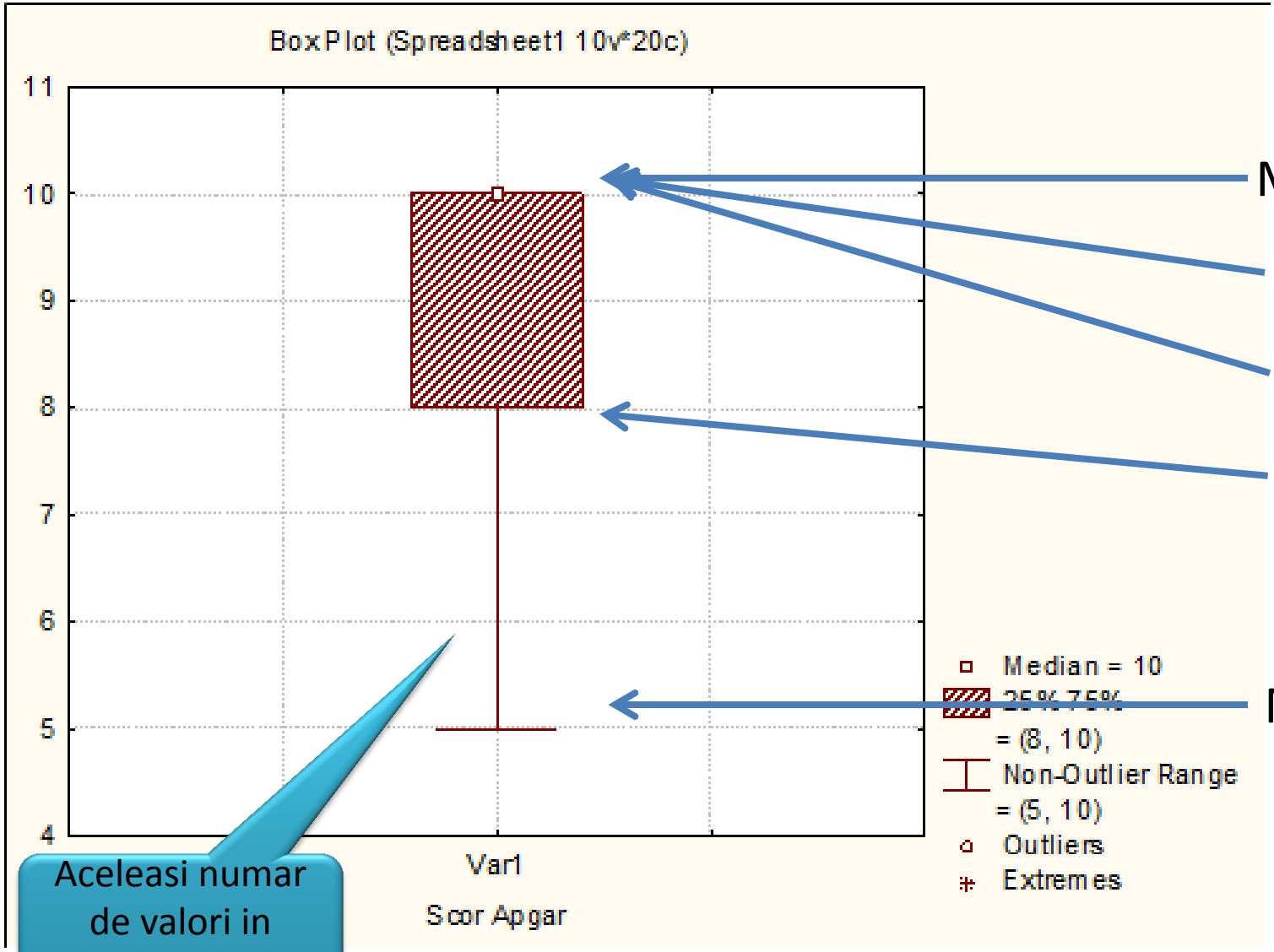
- **Interval intercvartilic** – diferenta dintre a treia si prima cvartilă

$$IQR = Q3-Q1$$

Indicatori pentru variabile ordinale



Indicatori pentru variabile ordinale – Chart – Box-whiskers Plot



Maxim = Cvartila 4 = 10

Cvartila 3 = 10

Mediana = Cvartila 2 = 10

Cvartila 1 = 8

Minim = Cvartila 0 = 5

Aceleasi numar de valori in fiecare interval

Percentila

- O valoare astfel încât un anumit procent de date dintr-un set de date sunt sub ea

Ex. Percentila 25 = 7,

25% din date sunt sub 7

- 0-100 percentile

Scor Apgar
5
6
7
7
8
8
9
9
9
10
10
10
10
10
10
10
10
10
10
10

Modul (de la modă)

- Valoarea cu cea mai mare frecvenţă
- Ex. Scor Apgar 10 apare 11 ori ⟶ Modul=10
- Dacă sunt 2 valori ⟶ bimodală
- Ex. 1 3 4 5 5 6 6 7 8 9 ⟶ Modul = 5 şi 6

Exerciţiu

- Notele la examenul de informatică pentru grupa 4

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

Sortăm ascendent: 4, 5, 6, 6, 7, 7, **7**, 7, 8, 8, 9, 9, 10

Q0 = Minim =

Q1 =

Q2 = Mediana =

Q3 =

Q4 = Maxim =

Amplitudinea = Maxim-Minim =

Interval intercvartilic=

Modul =

Exerciţiu

- Notele la examenul de informatică pentru grupa 4

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

Sortăm ascendent: 4, 5, 6, 6, 7, 7, **7**, 7, 8, 8, 9, 9, 10

Q0 = Minim = 4

Q1 = 6

Q2 = Mediana =7

Q3 = 8,5

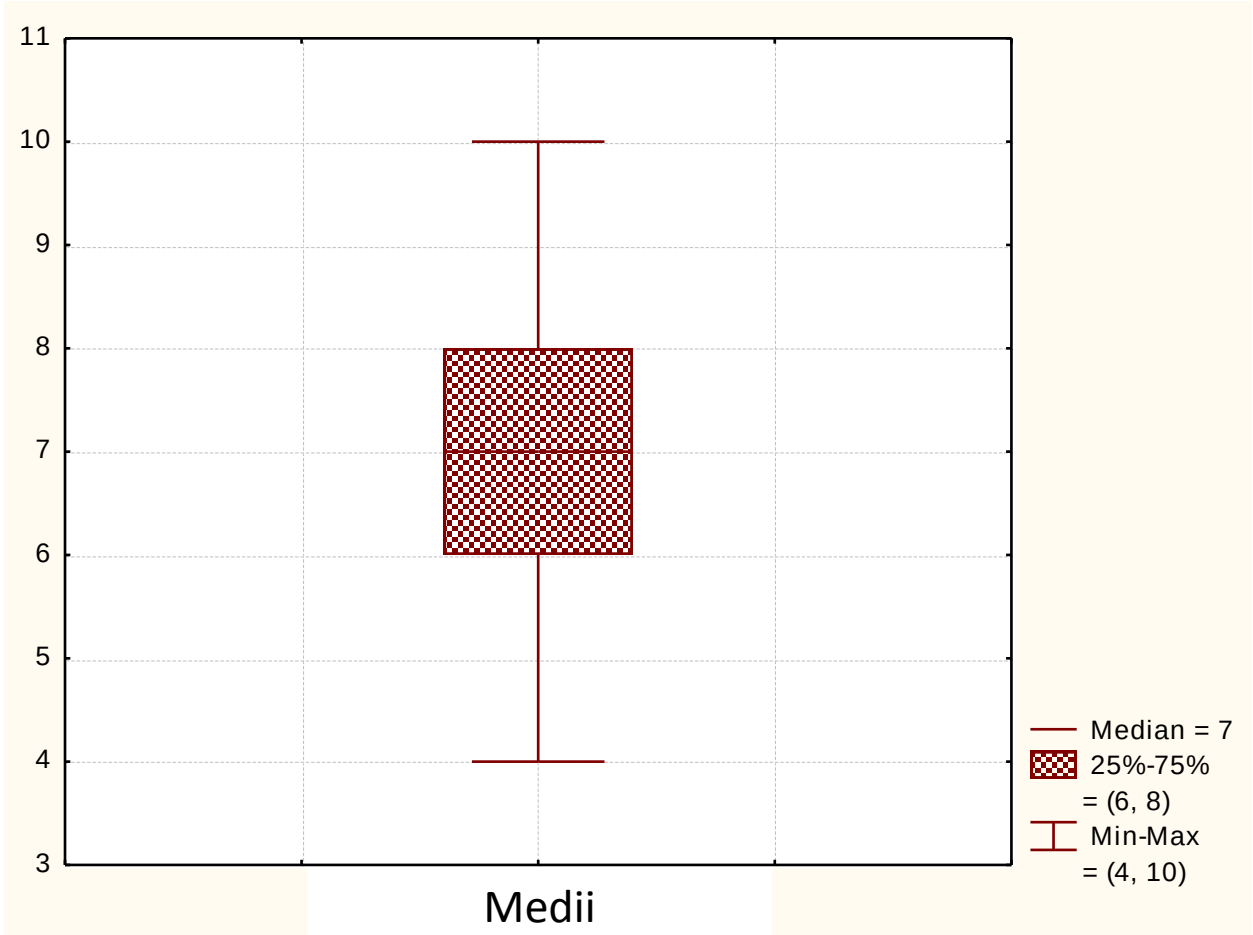
Q4 = Maxim =10

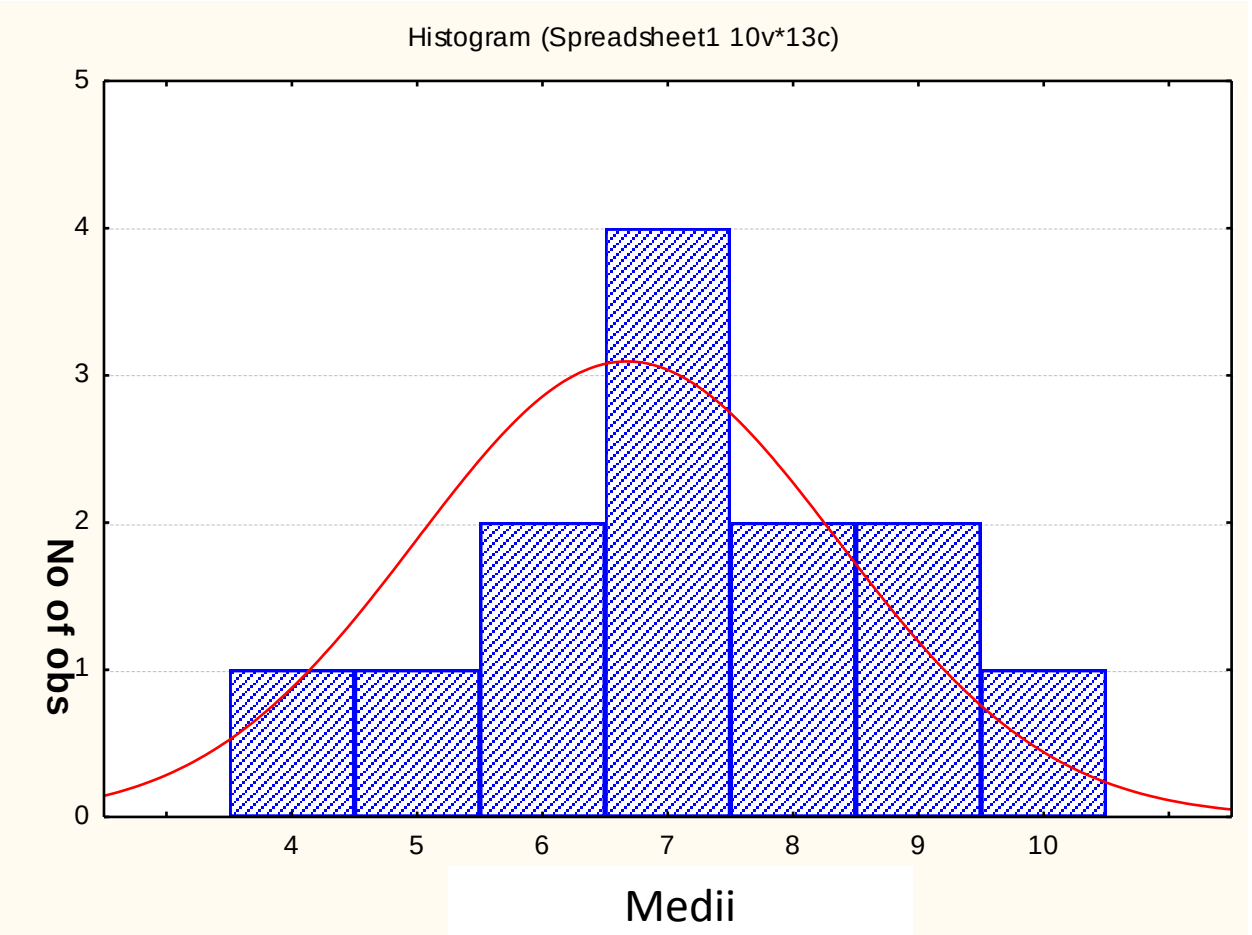
Amplitudinea = Maxim-Minim = 6

Interval intercvartilic = Q3-Q1 = 2,5

Modul = 7

Notele la examenul de informatică pentru grupa 4





Indicatori pentru variabile ordinale

- Indicatori
  - Amplitudinea
  - Mediana
  - Cvartile 0-4
  - Intervalul intercvartilic
  - Modul
- Grafic
  - Box-Plot
  - Coloane

Descrierea datelor numerice (Continue, Discrete)

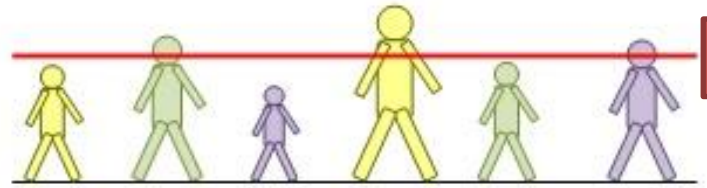
C
Varsta mamei
22
25
32
28
25
26
31
35
26
24
25
27
29
30
26
21
29
33
41
28

G	H	I
Numar saptamani de sarcina	Numar nasteri mama	Numar sarcini
40	2	2
41	0	3
39	0	2
41	0	0
34	2	3
41	0	4
41	0	0
36	3	3
41	1	1
39	0	0
41	1	1
41	0	8
41	1	1
40	1	1
41	2	2
39	1	4
41	1	1
28	0	1
29	0	1
40	0	0

- Măsuri ale tendinţei centrale
  - Media aritmetică
  - Mediana
  - Modul
- Măsuri de dispersie (împrăştiere)
  - Varianţa
  - Deviaţia Standard
  - Coeficientul de variaţie
  - Eroarea Standard

- Alte măsuri
  - Asimetria
  - Boltirea
  - Cvartile
  - Percentile
- Grafice
  - Histograma
  - Box-plots
  - Medie/Error plot

Măsurarea tendinței centrale



Media aritmetică $\overline{X}$ este media observaţiilor.

Mod de calcul: adunați observațiile pentru a obține suma și apoi împărțiți-o la numărul de observații.

Formula pentru medie: $$\overline{X} = \frac{1}{n}\sum X$$

Σ înseamnă adunare, X reprezintă observaţiile individuale, n este numărul de observaţii.

Mean is a measure of the middle? Not all the time

- Ex. 6 dentists earn in a month 2400, 2500, 2900, 2900, 3000, 3100 Euro.

$$\bar{X} = \frac{2400+2500+2900+2900+3000+3100}{6} = 2800 \text{ Euro}$$

Venitul personalului medical

2400€ 2500 €

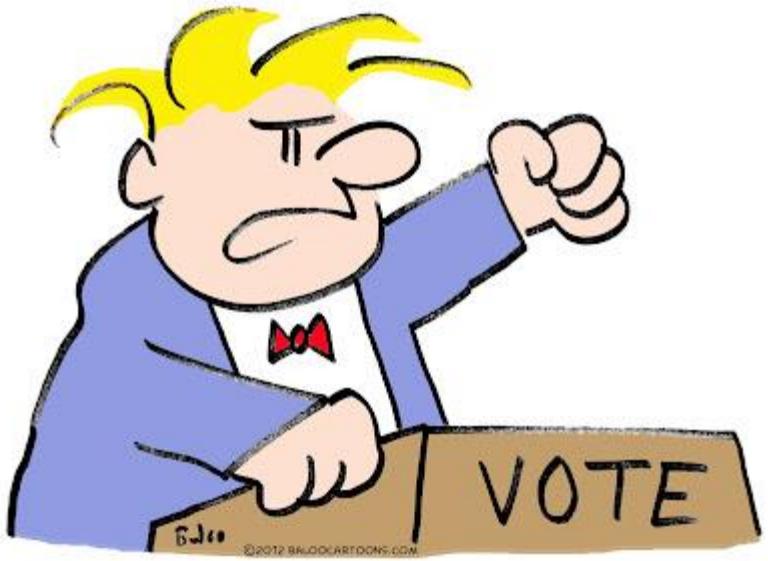
2900 € 2900 €

3000 € 3100 €

Media aritmetică = 2800 €

Media veniturilor medicilor 2800 €

Se schimba politicile sau apare o criză



"Desperate times call for desperate politicians!"

După 12 luni politicienii anunţă că:

S-au dublat veniturile personalului medical după 12 luni

2400€ 2500 €

2900 € 2900 €

3000 € 19900 €

Media aritmetică = 5600 €

a fost 2800 €

- Concluzie: cayurile aberante pot influenţa media

Mediana

2400€ 2500 €

2900 € 2900 €

3000 € 3100 €

Mediana = 2900 €

Veniturile personalului medical după 12 luni

2400€ 2500 €

2900 € 2900 €

3000 € 19900 €

Mediana = 2900 €

a fost 2900 €

Media este o măsură a mijlocului?

Folosim media dacă datele sunt numerice fără cazuri aberante

Utilizăm valoarea mediană dacă datele sunt numerice cu valori extreme sau datele sunt ordinale

Alte măsuri de centralitate

- Media geometrică $\left(\prod_{i=1}^{n} a_i\right)^{1/n} = \sqrt[n]{a_1\, a_2\, a_3 \ldots . a_n}$
- Media armonică $H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_1} + \ldots + \frac{1}{x_1}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$
- Valoarea centrală = (maxim+minim)/2
- Media ponderată $\bar{x} = \frac{w_1 x_1 + w_2 x_2 + \ldots + w_n x_n}{w_1 + w_2 + \ldots + w_n}$

Seria 1	Seria 2	Seria 3
1	1	1
1	44	8
2	45	11
3	46	14
5	48	28
6	48	30
6	49	37
7	50	48
93	50	52
94	51	62
94	52	70
95	52	72
97	54	84
98	55	91
98	55	92
100	100	100
800	800	800

Măsuri de dispersie

De ce avem nevoie de măsurile de dispersie când vrem să descriem datele?

Numărul de observaţii = 16

Media = 50

Mediana = 50

Distribuţia definiţie

- Arată cât de des apare o valoare (frecvenţa) sau un grup de valori (clase de frecvenţă)

Seria1

1

1

2

3

5

6

6

7

93

94

94

95

97

98

98

100

Împărţim amplitudinea = maxim-minim = 99 în 4-10 clase

Clase de frecvenţă = intervale egale

Seria 1 –

Clase

0-20

21-40

41-60

61-80

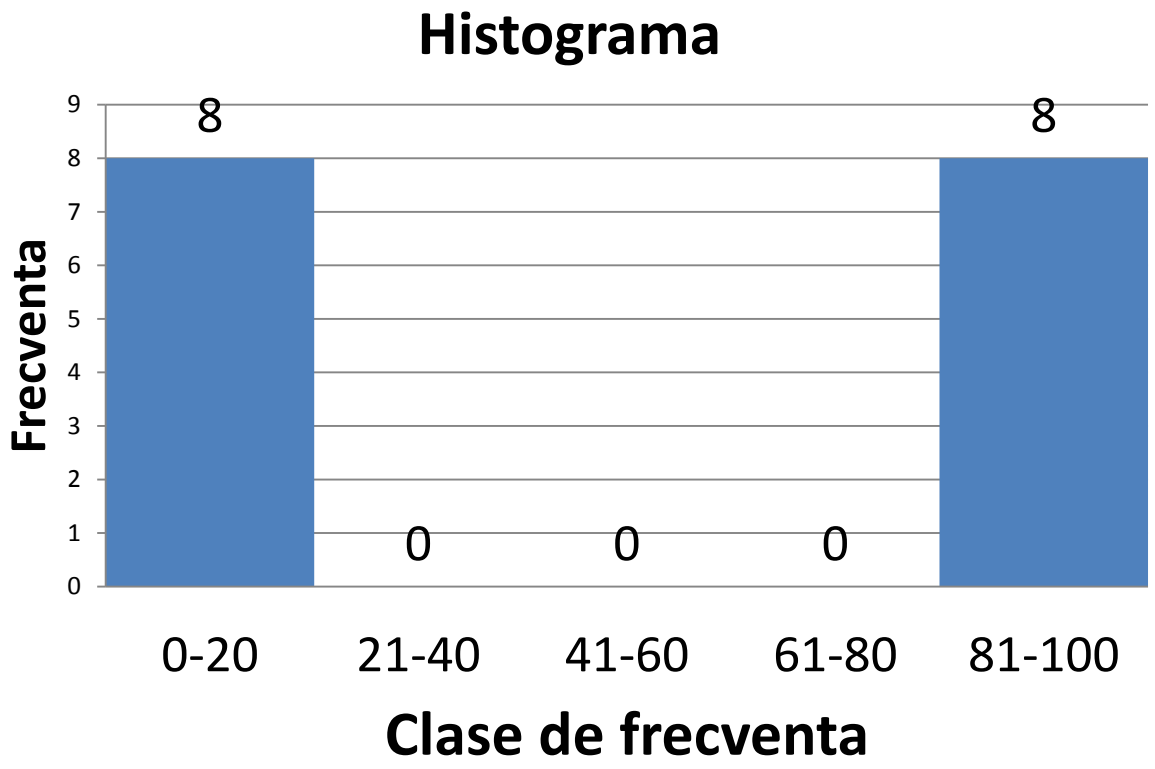
81-100

Seria 1

1
1
2
3
5
6
6
7
93
94
94
95
97
98
98
100

Distribuţia

Seria 1 – Clase	Frecvenţa absolută	Frecvenţa relativă %
0-20	8	50
21-40	0	0
41-60	0	0
61-80	0	0
81-100	8	50



Series 2

1

44

45

46

48

48

49

50

50

51

52

52

54

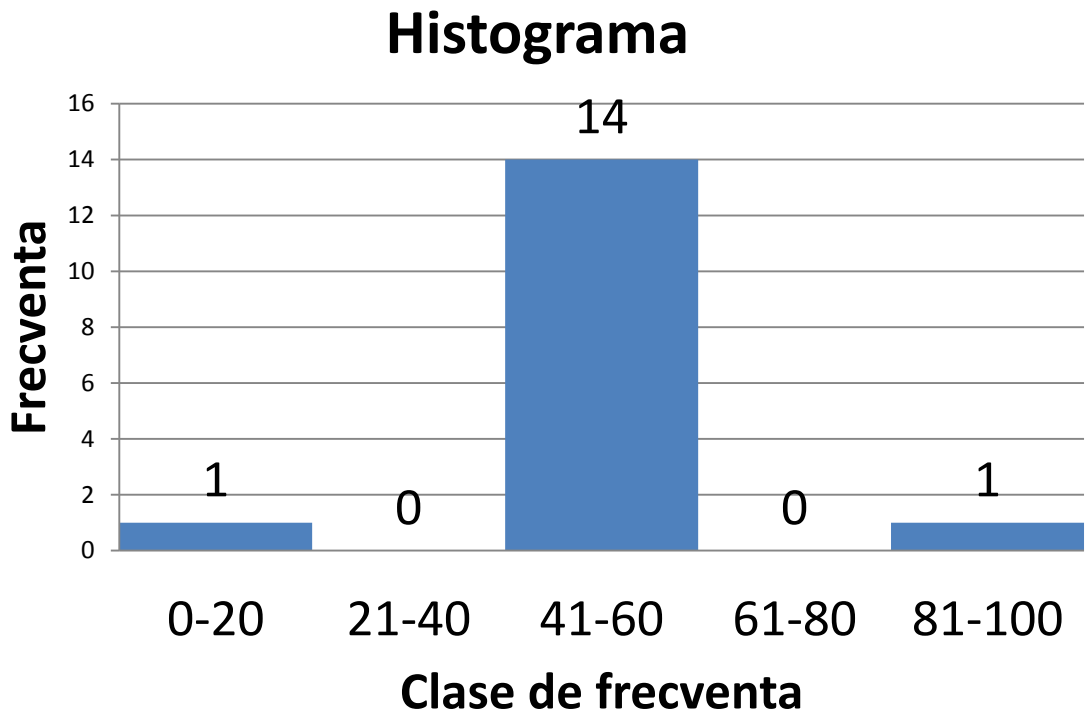
55

55

100

Distribuţia

Seria 2 – Clase	Frecvenţa absolută	Frecvenţa relativă %
0-20	1	6.25
21-40	0	0
41-60	14	87.50
61-80	0	0
81-100	1	6.25

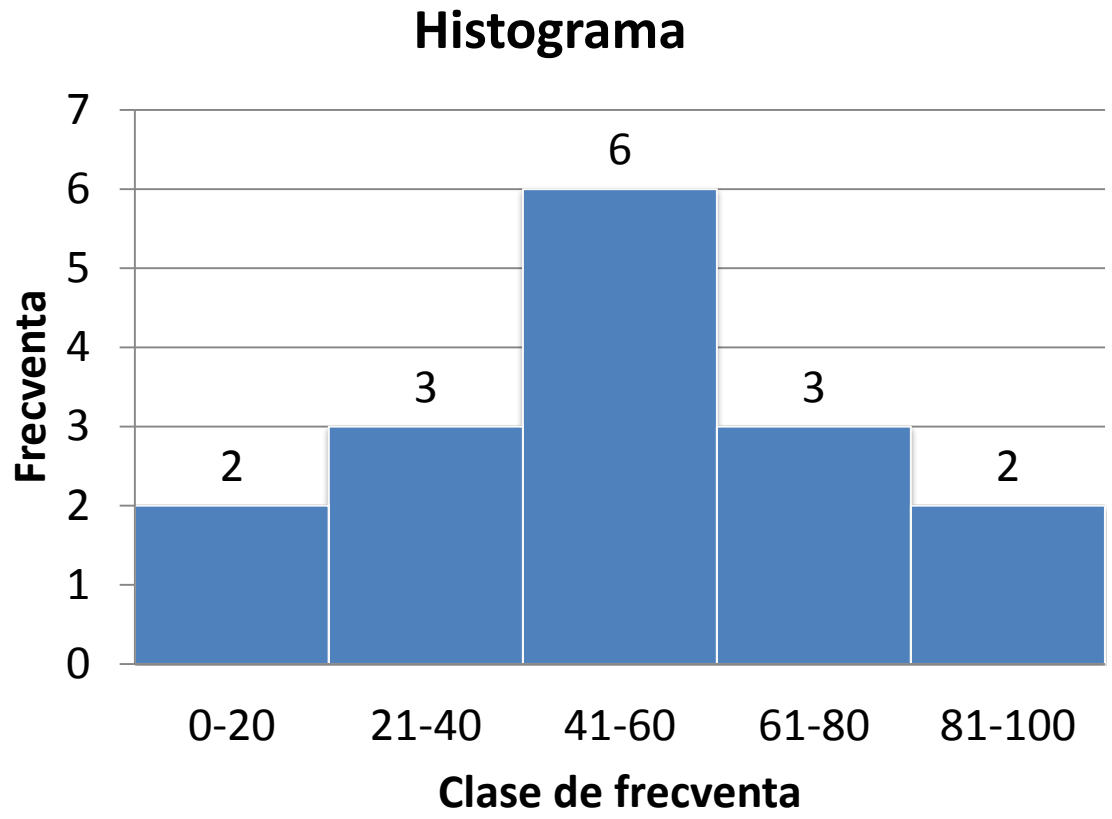


Series 3

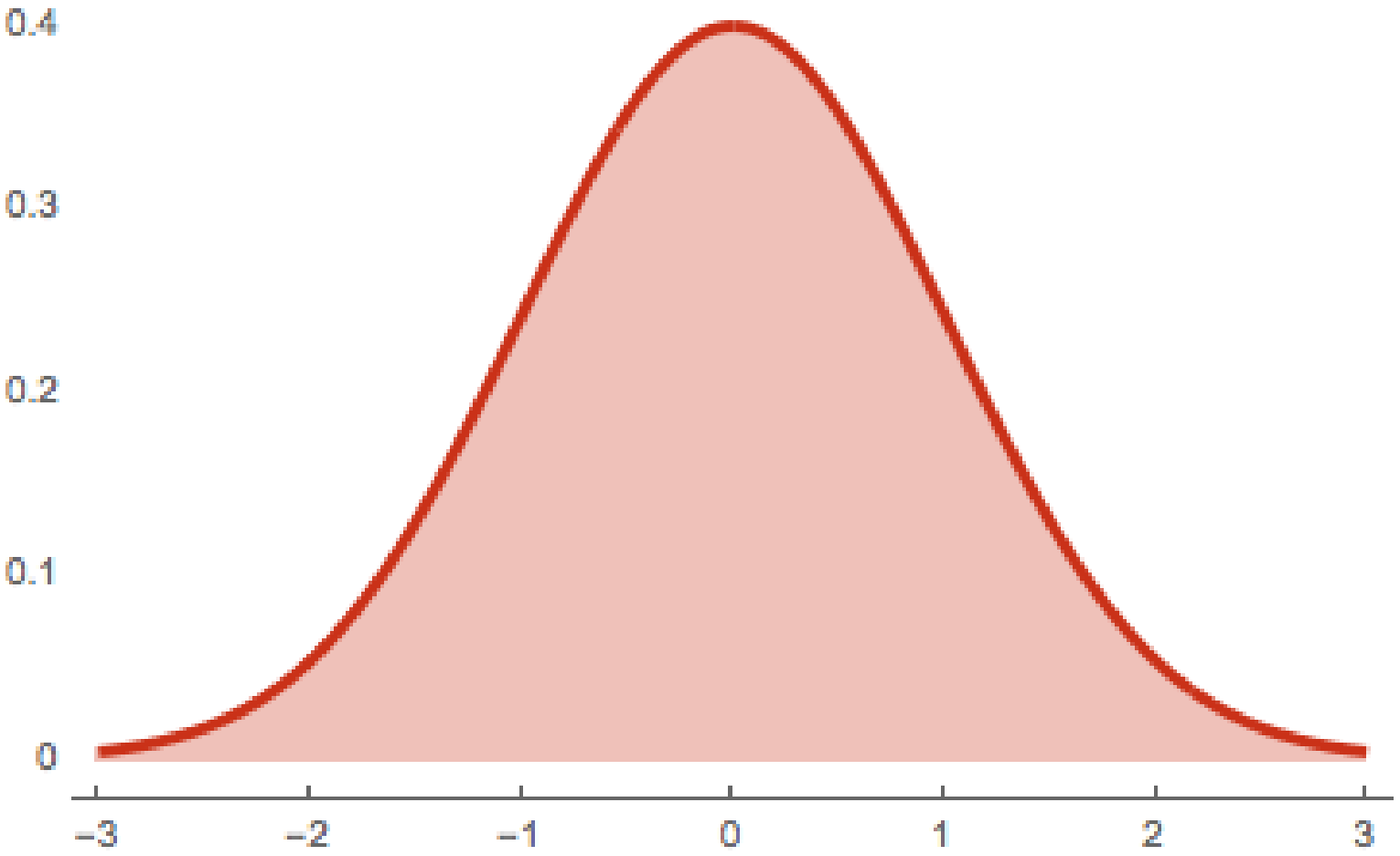
1
11
24
29
36
41
45
49
51
55
59
64
71
76
88
100

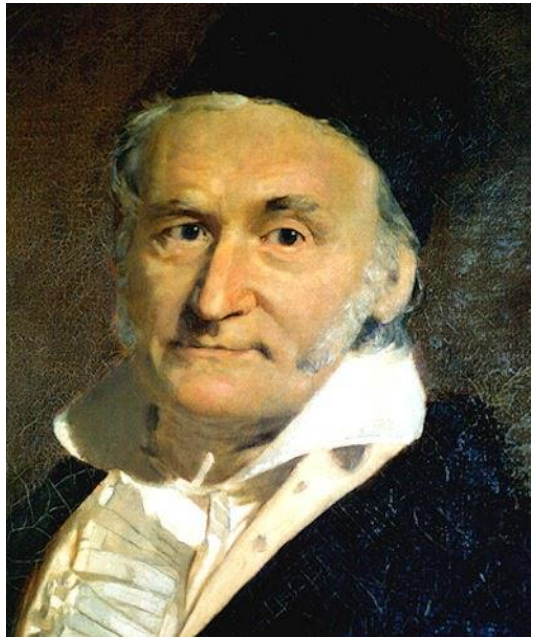
Distribuţia

Seria 3 – Clase	Frecvenţa absolută	Frecvenţa relativă %
0-20	2	12.50
21-40	3	18.75
41-60	6	37.50
61-80	3	18.75
81-100	2	12.50



Distribuţia normală (sub formă de clopot, Gauss)

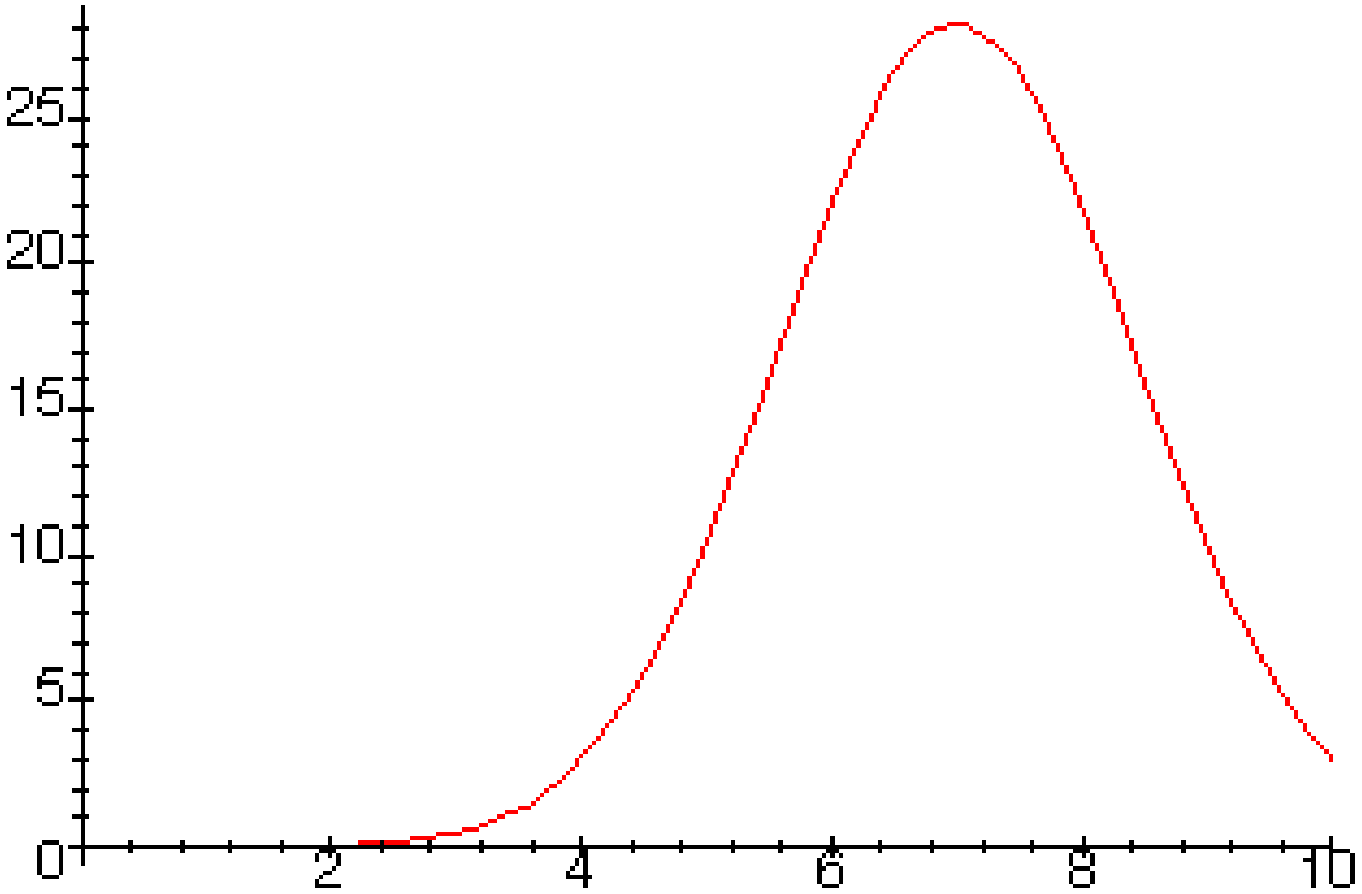


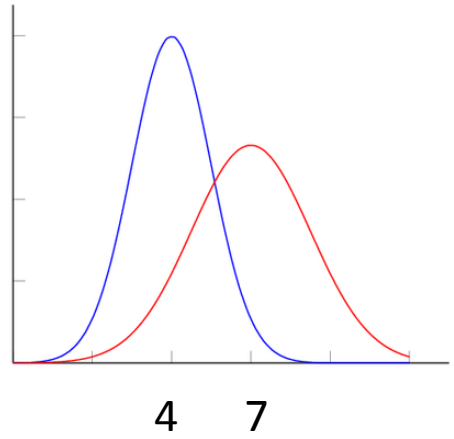


Carl Friedrich Gauss
(30 Aprilie 1777 – 23 Februarie 1855)

Normalitatea

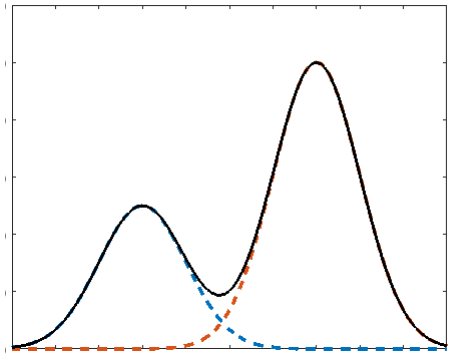
Notele a 100 de studenţi - Distribuţia





Subiecte prea grele

Asimetrie



Frauda la examen–
distributie bimodala

Măsuri de dispersie

Deviația standard – Media deviației de la medie

Formula pentru populație (toate observațiile):

$$\sigma = \sqrt{\frac{\sum(X - \mu)^2}{N}}$$

Σ înseamnă adunare, X reprezintă observaţiile individuale, μ este media aritmetică a întregii populaţii, N este numărul de observaţii.

Formula pentru un subset al populaţiei (eşantion)

- **Deviația standard pentru eşantion** (cu corecție)

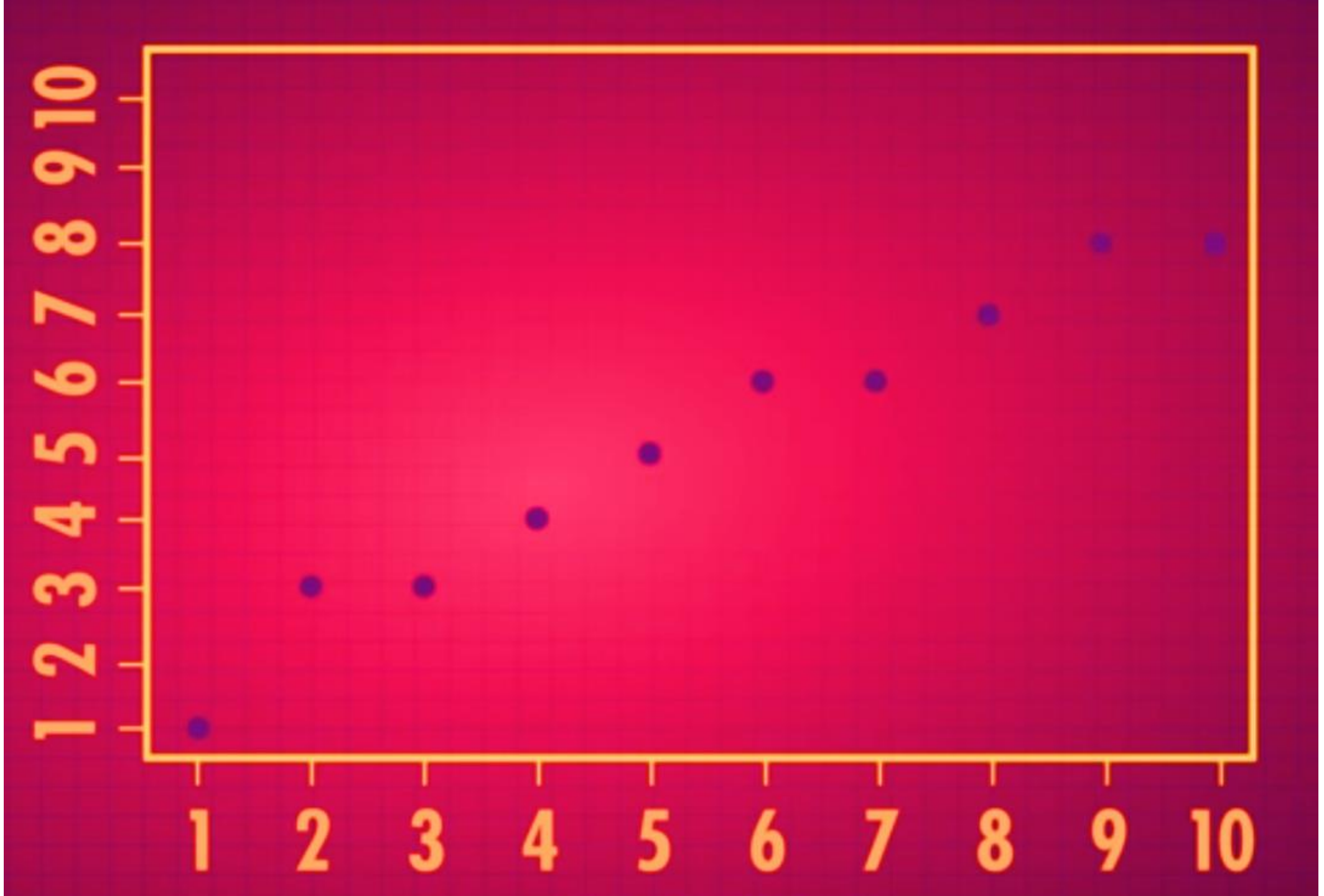
$$S = \sqrt{\frac{\sum(X-\bar{X})^2}{N-1}} = \sqrt{\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_N-\bar{x})^2}{N-1}}$$

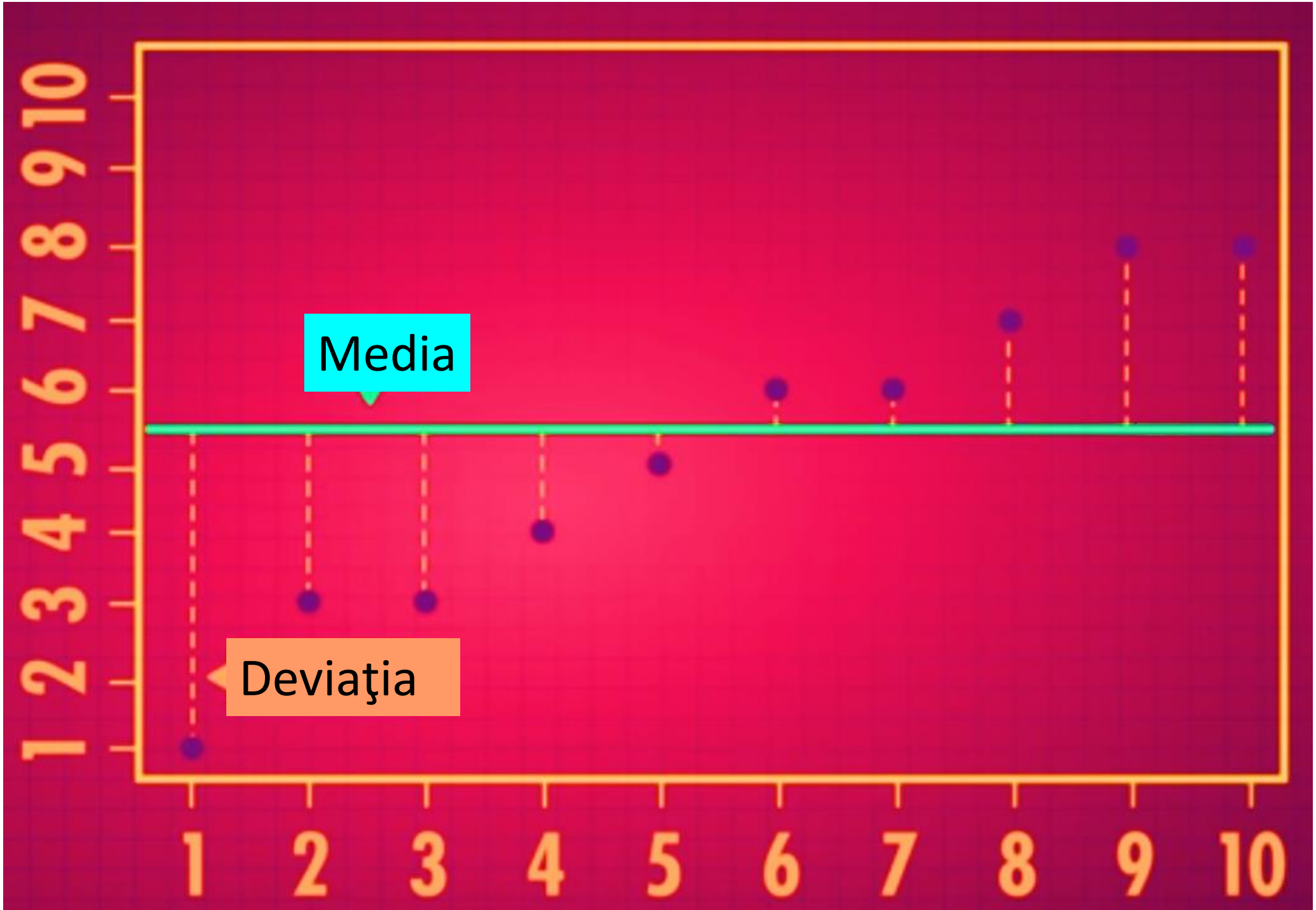
unde

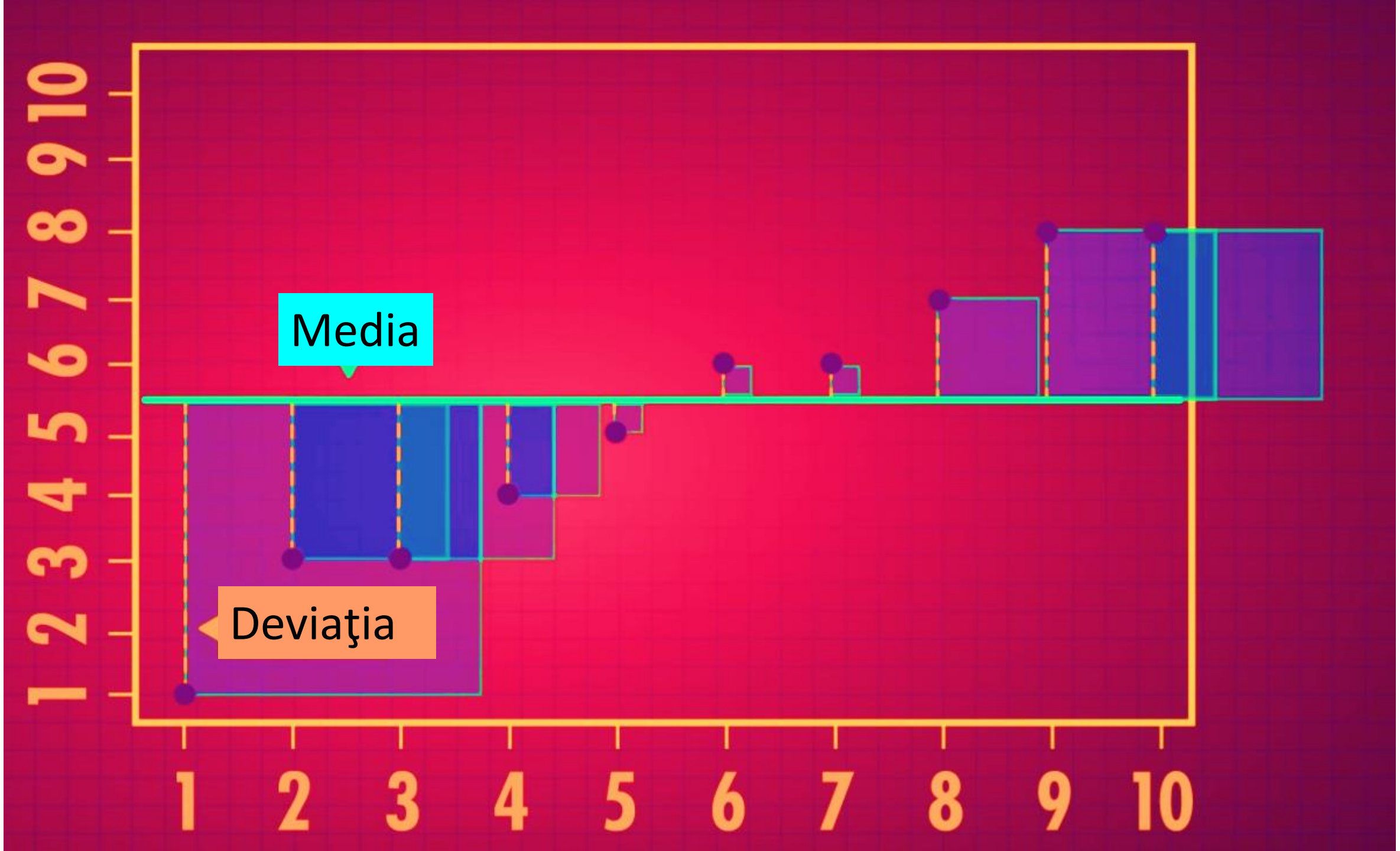
N – numărul totalde observaţii

$\bar{X}$ - media aritmetică

X_1, … X_N - observaţiile


1 2 3 4 5 6 7 8 9 10
1 2 3 4 5 6 7 8 9 10


Media
Deviaţia
1 2 3 4 5 6 7 8 9 10
1 2 3 4 5 6 7 8 9 10


Media
Deviaţia
1 2 3 4 5 6 7 8 9 10
1 2 3 4 5 6 7 8 9 10

Exemplu

- Nr. de dinţi pentru 6 pacienţi: 20, 24, 26, 25, 25, 30
- Media aritmetică $\bar{X}$ =(20+24+26+25+25+30)/6=25
- Deviaţia standard:
- $S = \sqrt{\frac{(20-25)^2+(24-25)^2+(26-25)^2+(25-25)^2+(25-25)^2+(30-25)^2}{6-1}}=$
- $=\sqrt{\frac{(-5)^2+(-1)^2+1^2+0^2+0^2+5^2}{5}} = \sqrt{\frac{25+1+1+25}{5}} = \sqrt{\frac{52}{5}} = 10{,}4$

Deviaţia standard - **exemplu**

Ex. Vârsta pe un eşantion de 6 persoane: 24, 25, 29, 29, 30, 31 zile.

X = 24, 25, 29, 29, 30, 31 – datele individuale, $\bar{X}$ = 28

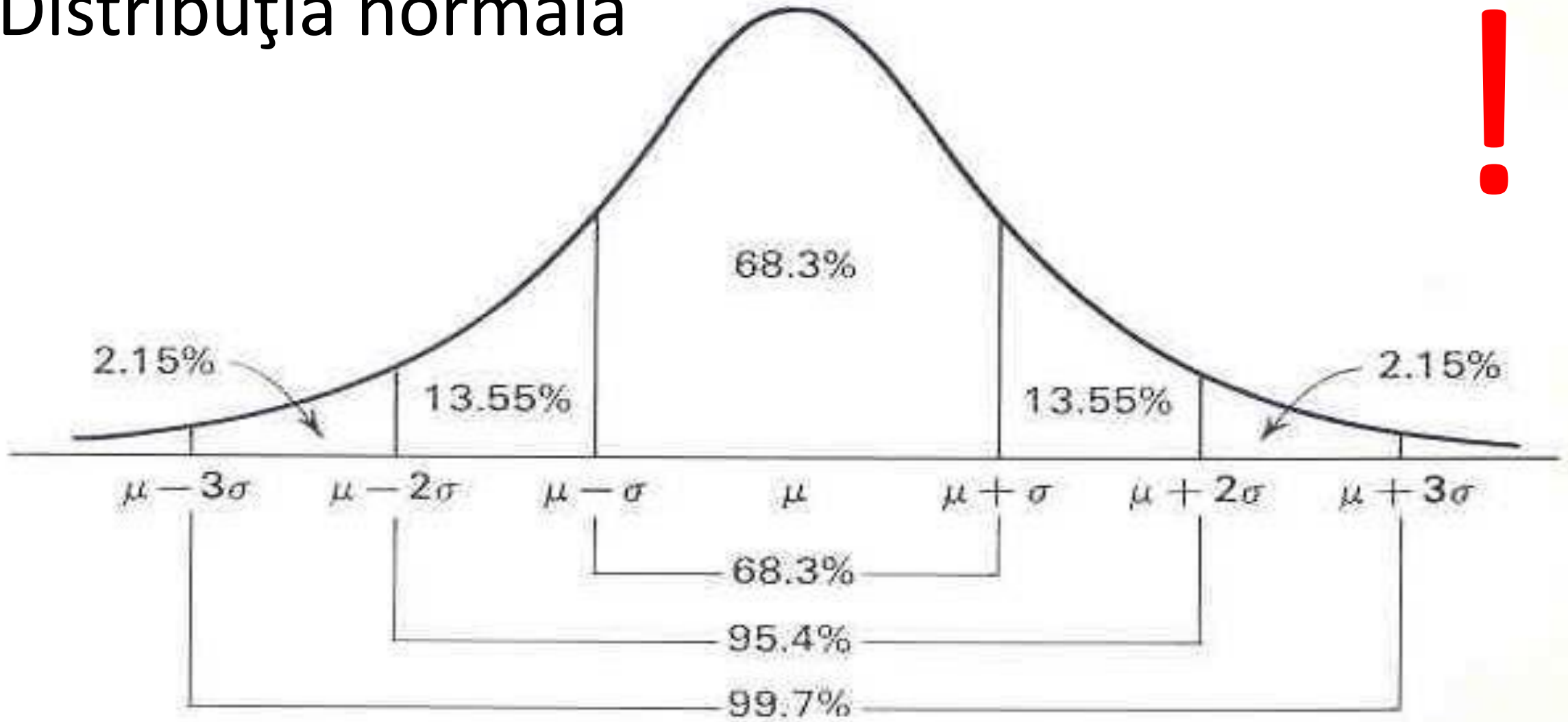
1. $X - \bar{X}$ = -4, -3, 1, 1, 2, 3.
2. $(X - \bar{X})^2$= 16, 9, 1, 1, 4, 9.
3. $\Sigma(X - \bar{X})^2$ = 16+9+1+1+4+9 = 40
4. $\Sigma(X - \bar{X})^2 /(n-1)$ = 40/5 = 8.
5. $\sqrt{\sum (X - \bar{X})^2 /(n-1)}$ = 2,83 Această valoare este deviaţia standard

Deviaţia standard- exemplu

Colesterolu pentru 10 pacienţi: 200, 180, 140, 160, 180, 150, 170, 110, 230, 170

	Colesterol *(X)*	**X-$\overline{X}$**	**(X-$\overline{X}$)²**
	200	31	961
	180	11	121
	140	-29	841
	160	-9	81
	180	11	121
	150	-19	361
	170	1	1
	110	-59	3481
	230	61	3721
	170	1	1
Suma	1690	**0**	9690
$\overline{X}$	**169**	**(X-$\overline{X}$)²/(n-1)**	**1076,6**
		Dev. St.	**32,8**

Distribuţia normală

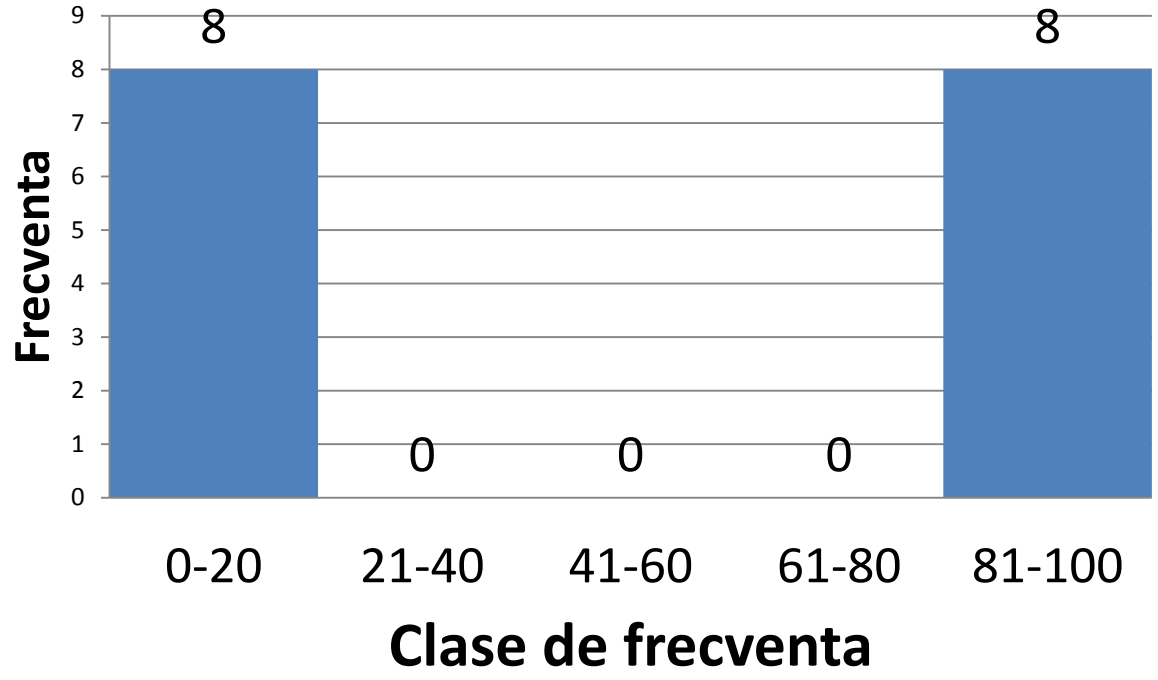


Aplicaţii ale deviaţiei standard

- Dacă distribuţia este normală (clopot, Gauss), atunci:
  - în intervalul: media ± dev.st. există minim 68.3% din date
  - În intervalul: media ± 2*dev.st. există minim 95.4% din date
  - În intervalul: media ± 3*dev.st. există minim 99.7% din date

Seria 1

1
1
2
3
5
6
6
7
93
94
94
95
97
98
98
100



Minim 68.3% din date

Minim 95.4% din date

Minim 99.7% din date

Deviaţia standard foarte mare, aproape de medie, concluzie: există date extreme

- Media aritmetică = 50
- Deviaţia standard = 47.70
- Media±dev.st.=[2.30;97.70] sunt **62.5%** din date
- Media ±2*dev.st. = **[-45.39**;145.39] sunt 100% din date
- Media ±3*dev.st. = [-93.09;193.09] sunt 100% din date

Seria 2

1

44

45

46

48

48

49

50

50

51

52

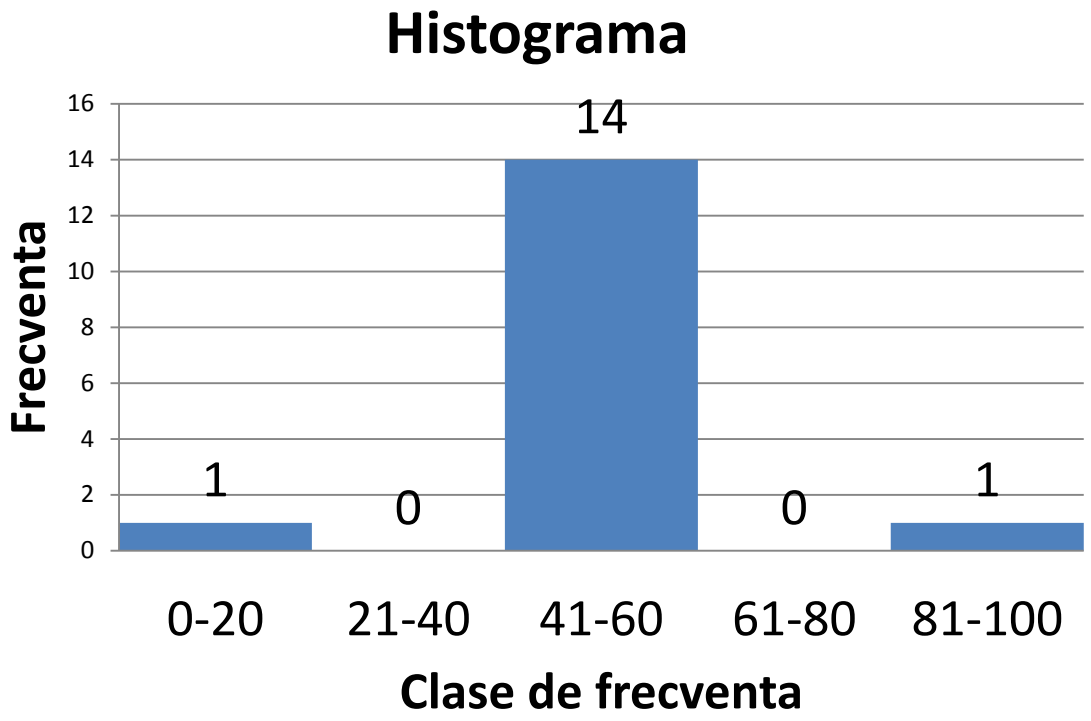
52

54

55

55

100



Minim 68.3% din date

Minim 95.4% din date

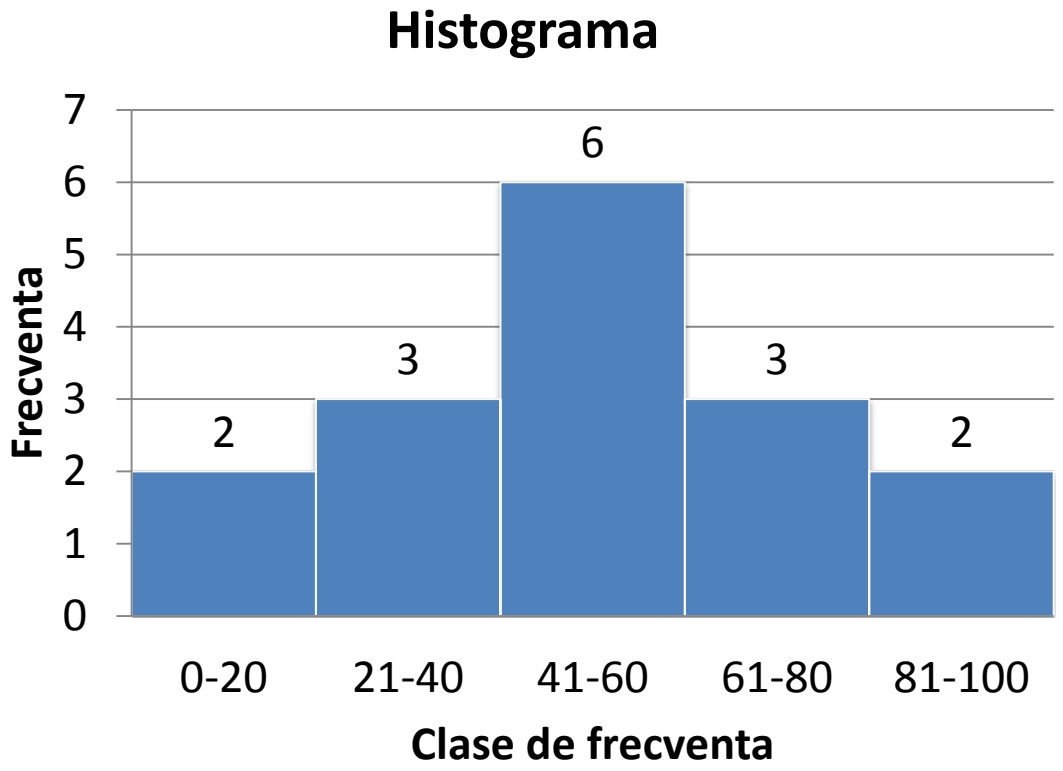
Minim 99.7% din date

Deviaţia standard este mică, concluzie: nu sunt cayuri aberante

- Media aritmetică = 50
- Deviaţia standard = 18.37
- Media±dev.st= [31.63;68.37] sunt 87.5% din date
- Media±2*dev.st.= [13.26;86.74] sunt **87.5%** din date
- Media±3*dev.st.= [-5.11;105.11] sunt 100% din date

Seria 3

1
11
24
29
36
41
45
49
51
55
59
64
71
76
88
100



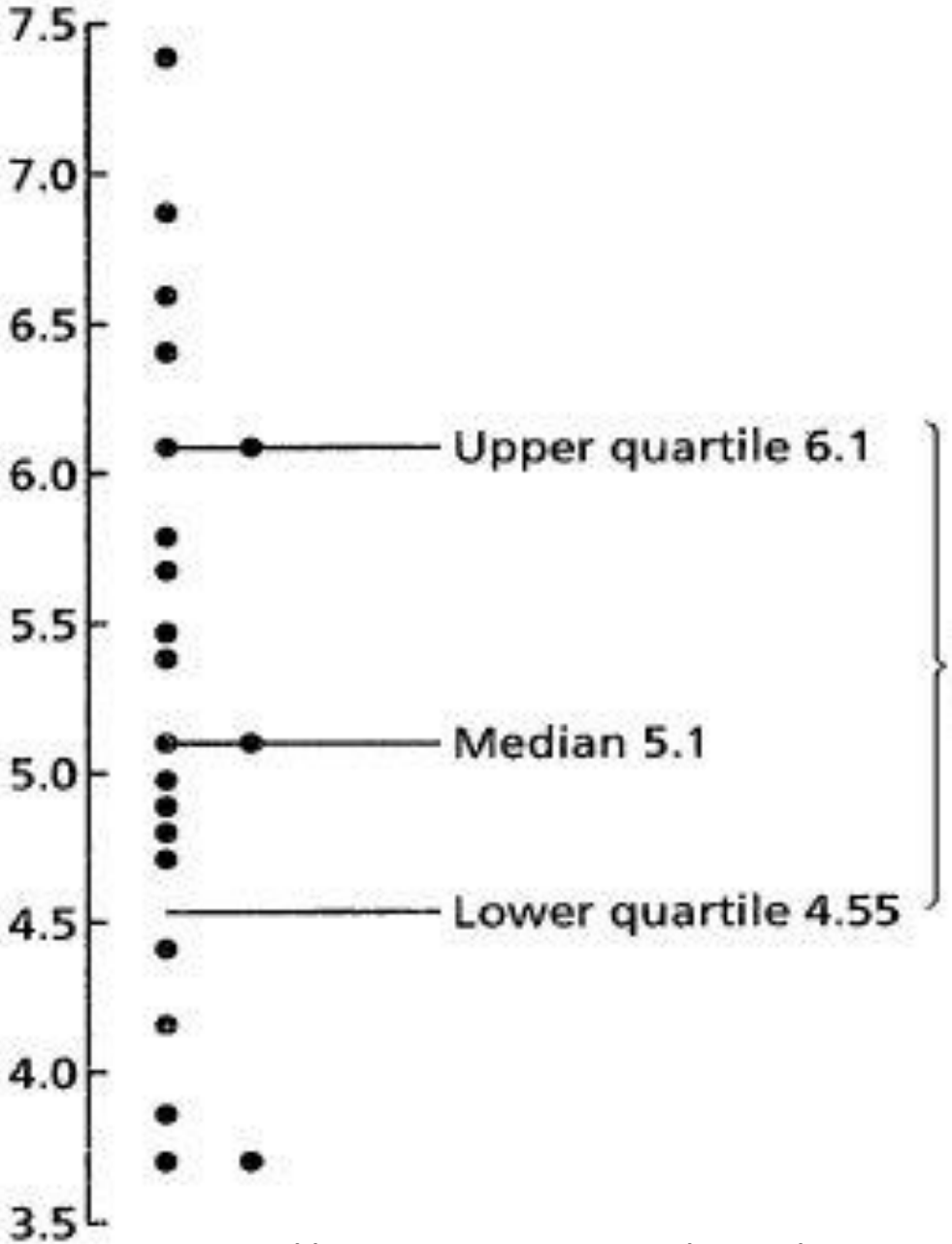
Minim 68.3% din date
Minim 95.4% din date
Minim 99.7% din date

Distribuţia normală

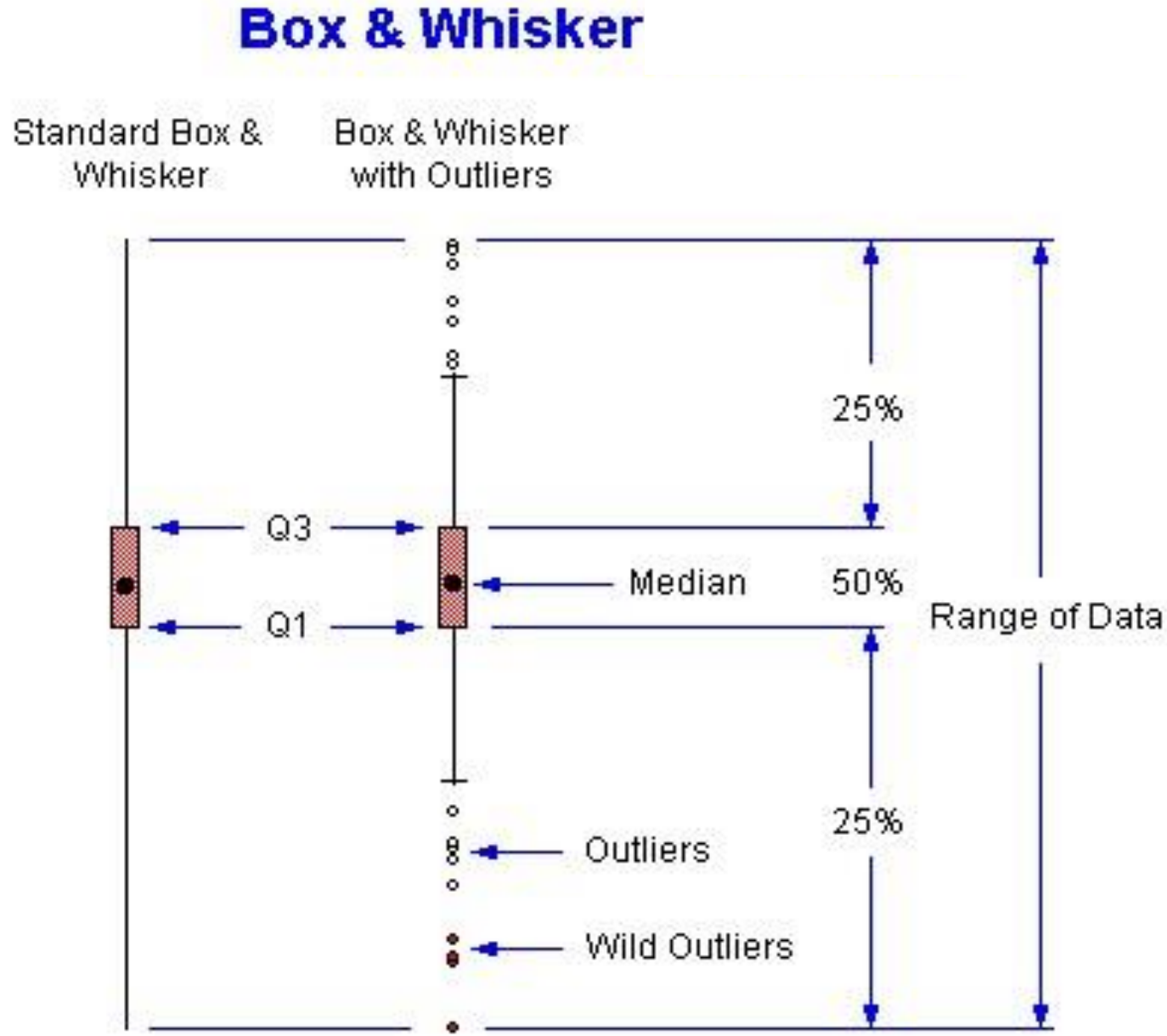
- Media aritmetică = 50
- Deviaţia standard = 26.71
- Media±dev.st. = [23.28;76.72] sunt 87.5% din date
- Media±2*dev.st. = [-3.43;103.43] sunt 100% din date
- Media±3*dev. st. = [-30.15;130.15] sunt 100% din date

Alte măsuri

- **Minim**
- **Maxim**
- **Amplitudinea** = maxim-minim
- **Coeficientul de variaţie** (CV=s/$\bar{X}$)
- **Eroarea standard** (E_s=s/$\sqrt{n}$)
- **Percentile**
- **Cvartile 0-4**
- **Interval intercvartilic**



http://www.answers.com/topic/interquartile-range

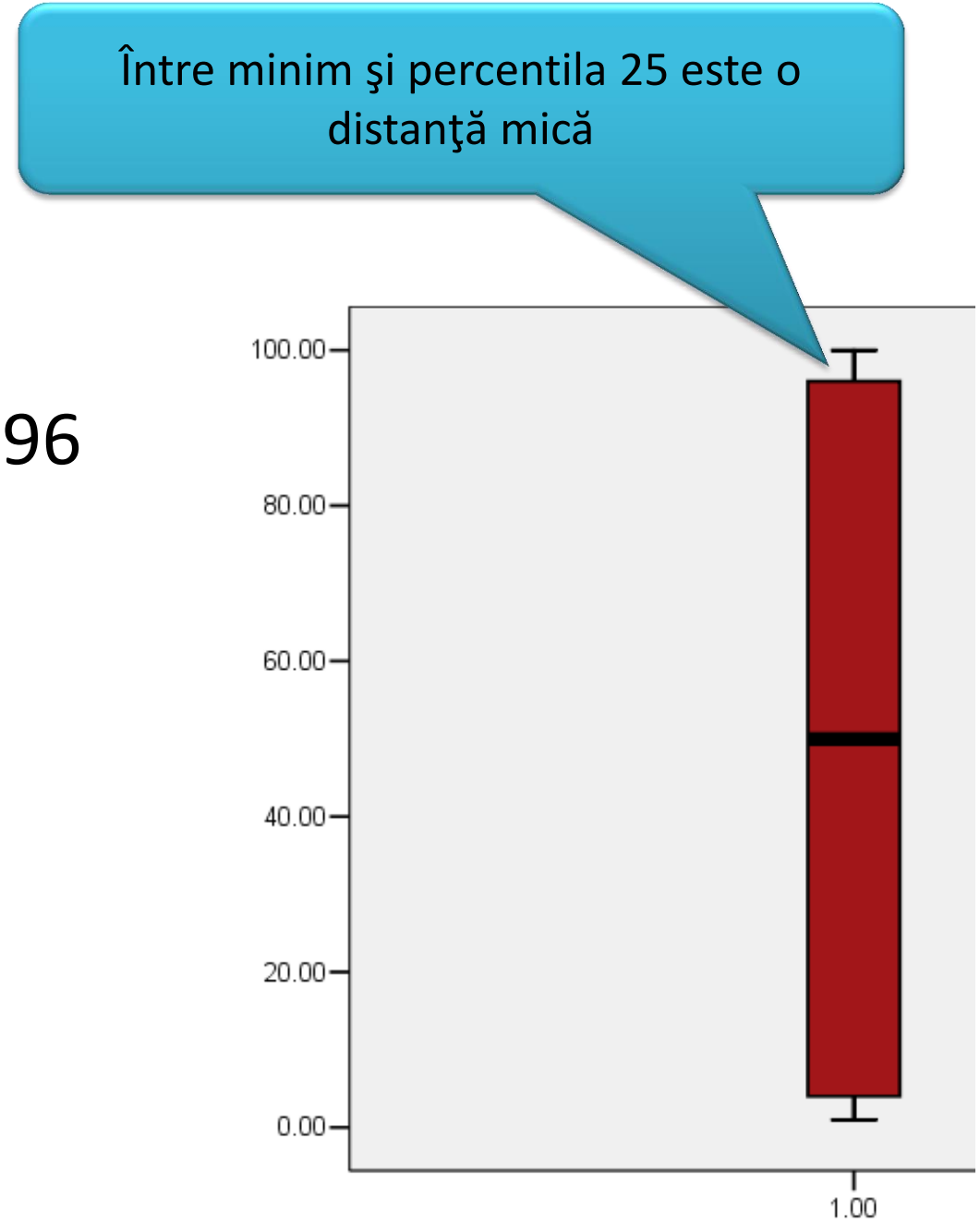


http://mvpprograms.com/help/mvpstats/graphics/WhatAreBoxAndWhiskerPlots

Seria 1
1
1
2
3
5
6
6
7
93
94
94
95
97
98
98
100

Mediana =(7+93)/2=50

- Percentile 25 =(3+5)/2= 4
- Percentile 75 =(95+97)/2=96

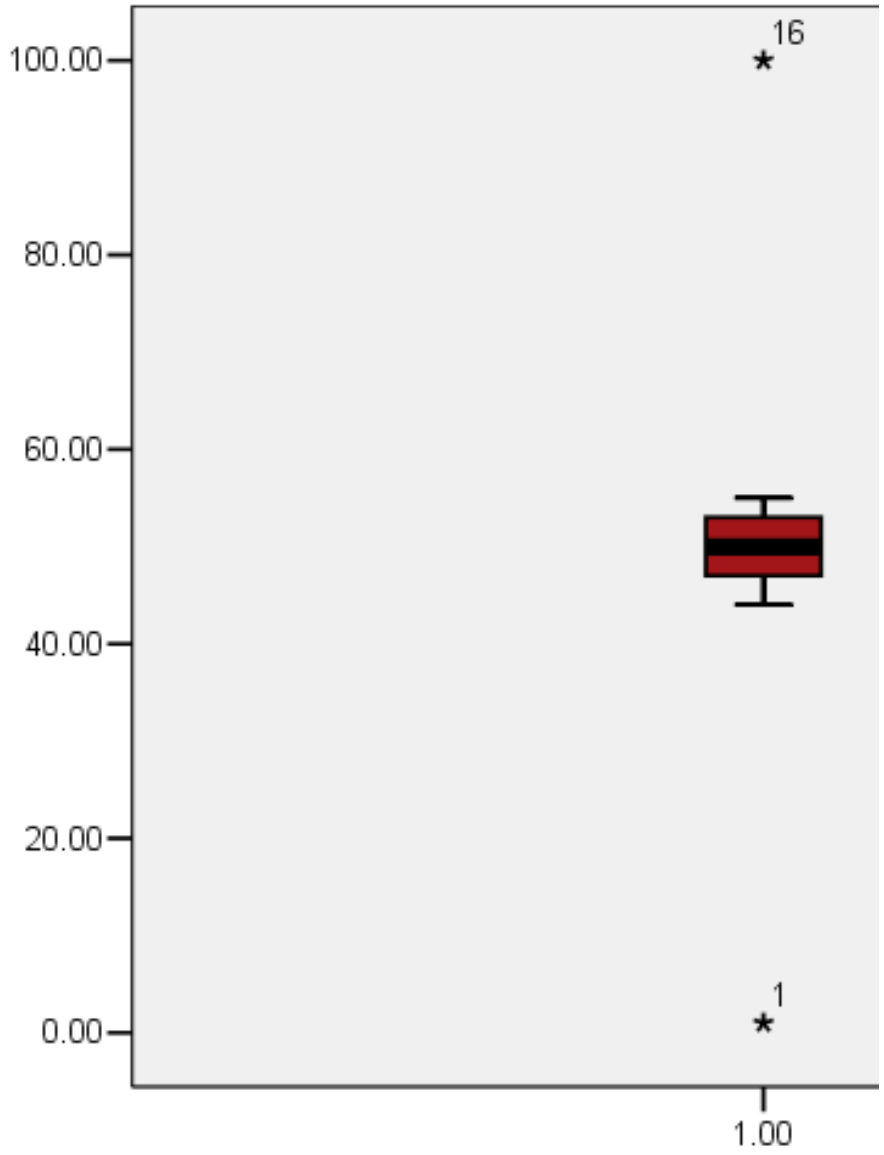


Seria 2

1
44
45
46
48
48
49
50
50
51
52
52
54
55
55
100

Mediana =(50+50)/2=50

- Percentila 25 =(46+48)/2= 47
- Percentila 75 =(52+54)/2=53

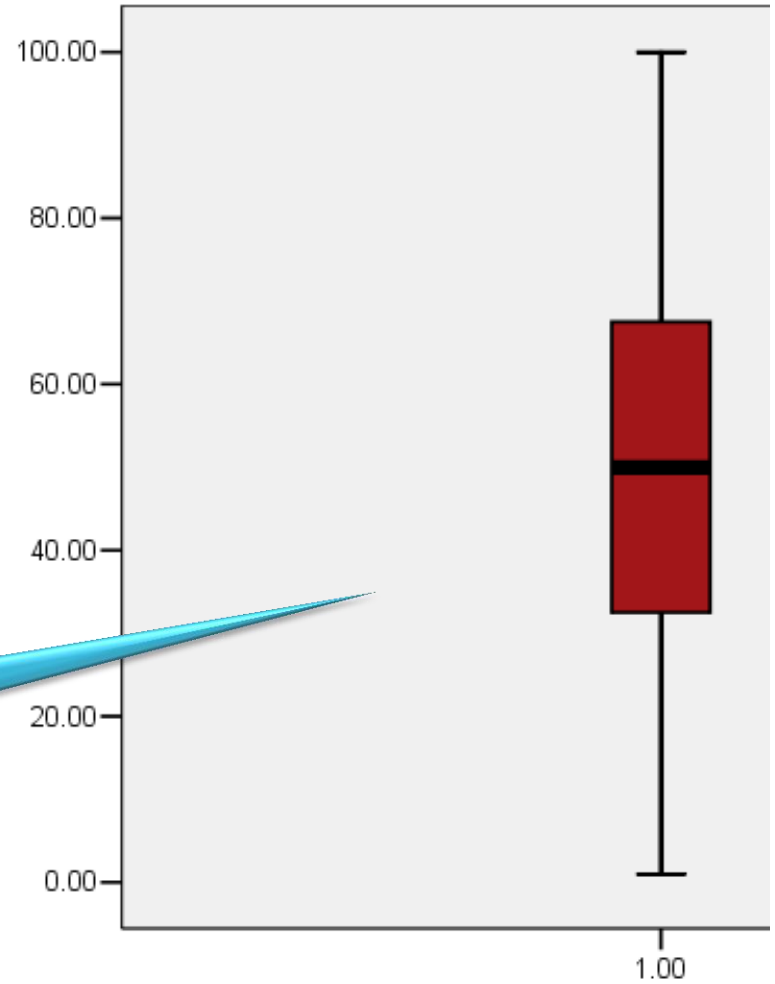


Seria 3

1
11
24
29
36
41
45
49
51
55
59
64
71
76
88
100

Mediana =(49+51)/2=50

- Percentile 25 =(29+36)/2= 32,5
- Percentile 75 =(64+71)/2=67,5

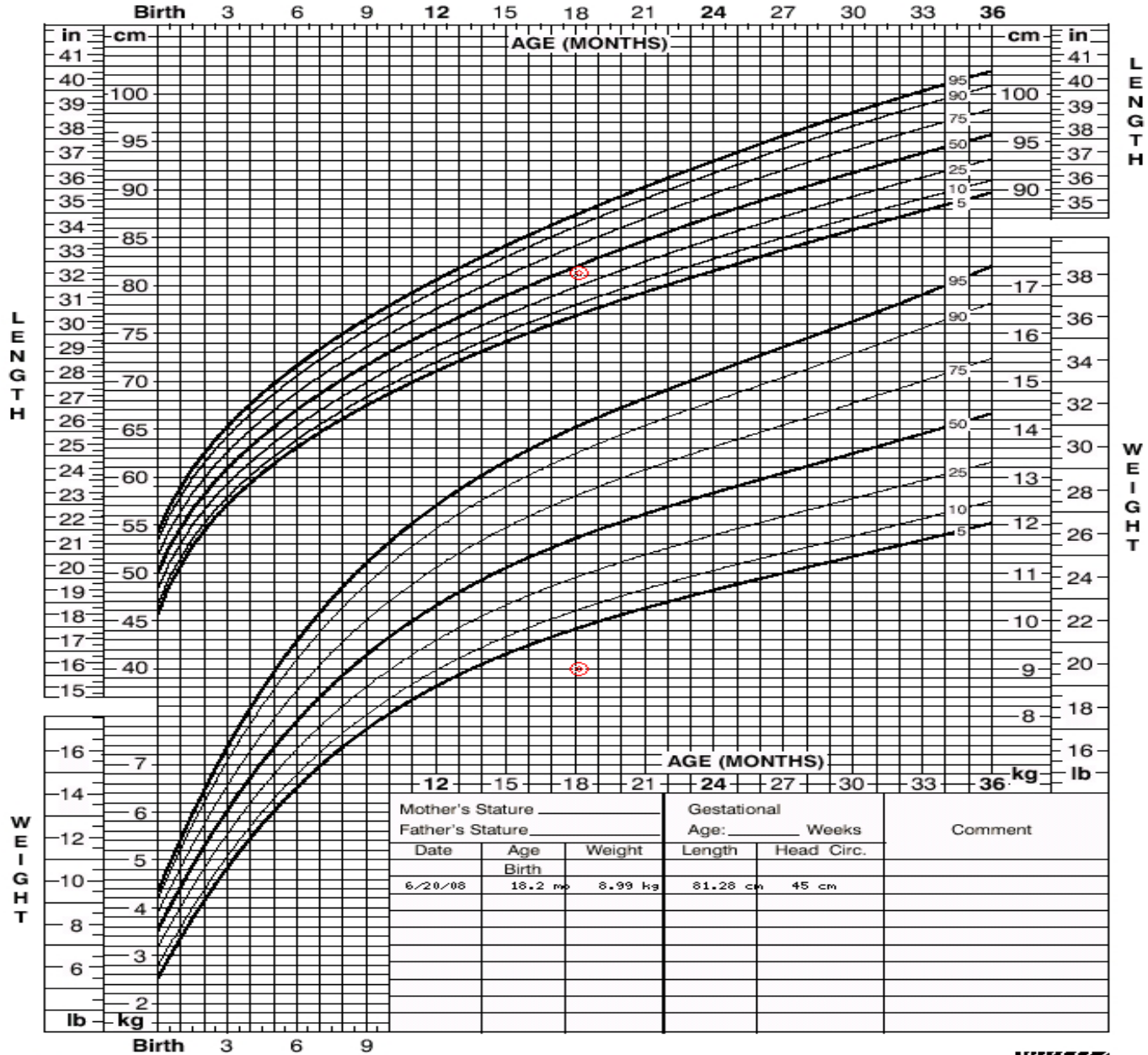


Distribuţie normală

Birth to 36 months: Boys
Length-for-age and Weight-for-age percentiles

NAME _____

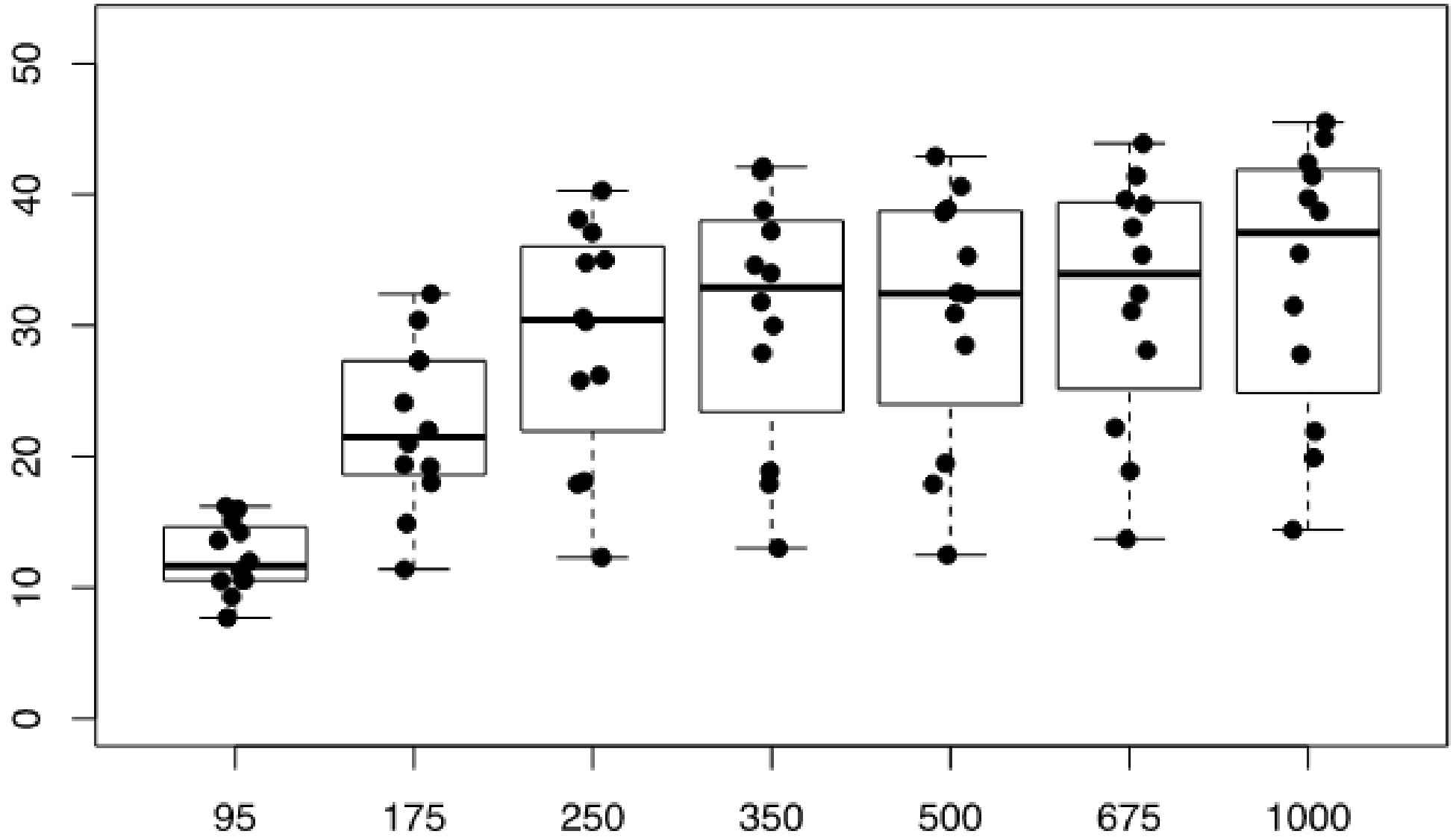
RECORD # _____



Mother's Stature _____ Father's Stature _____			Gestational Age: _____ Weeks		Comment
Date	Age	Weight	Length	Head Circ.	
	Birth				
6/20/08	18.2 mo	8.99 kg	81.28 cm	45 cm	

Revised April 20, 2001.

SOURCE: Developed by the National Center for Health Statistics in collaboration with the National Center for Chronic Disease Prevention and Health Promotion (2000).
http://www.cdc.gov/growthcharts

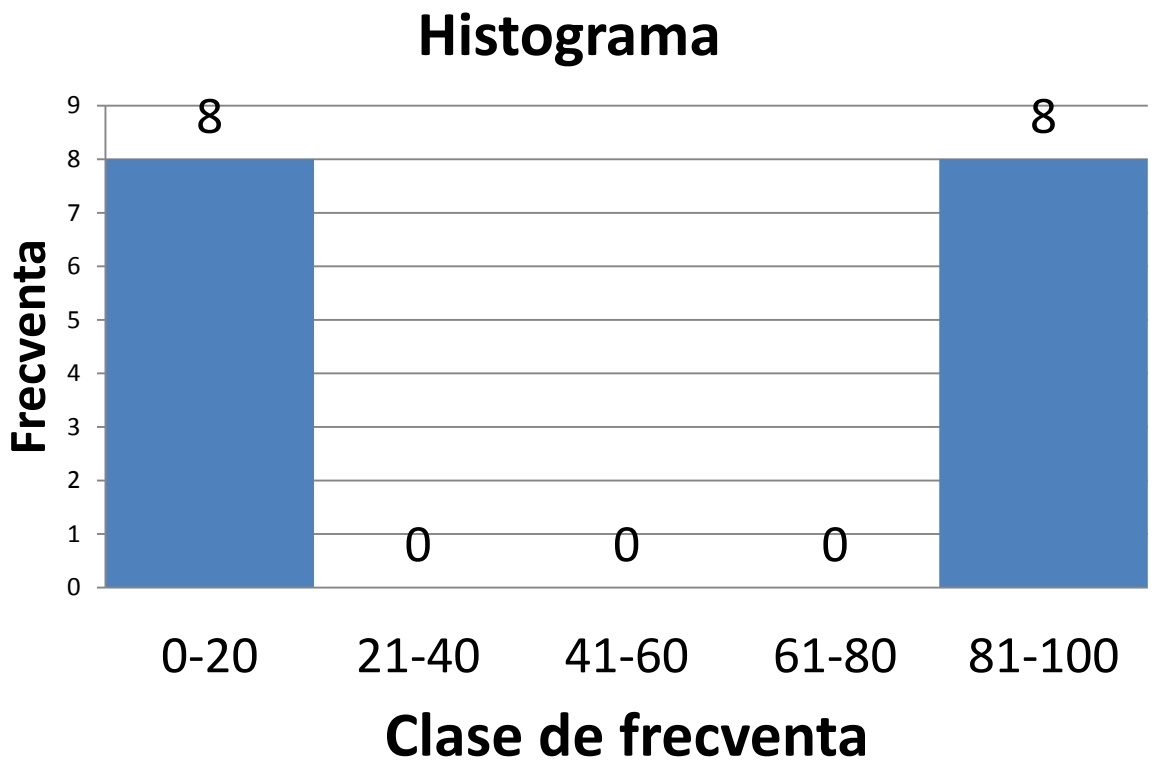

0
10
20
30
40
50
95
175
250
350
500
675
1000

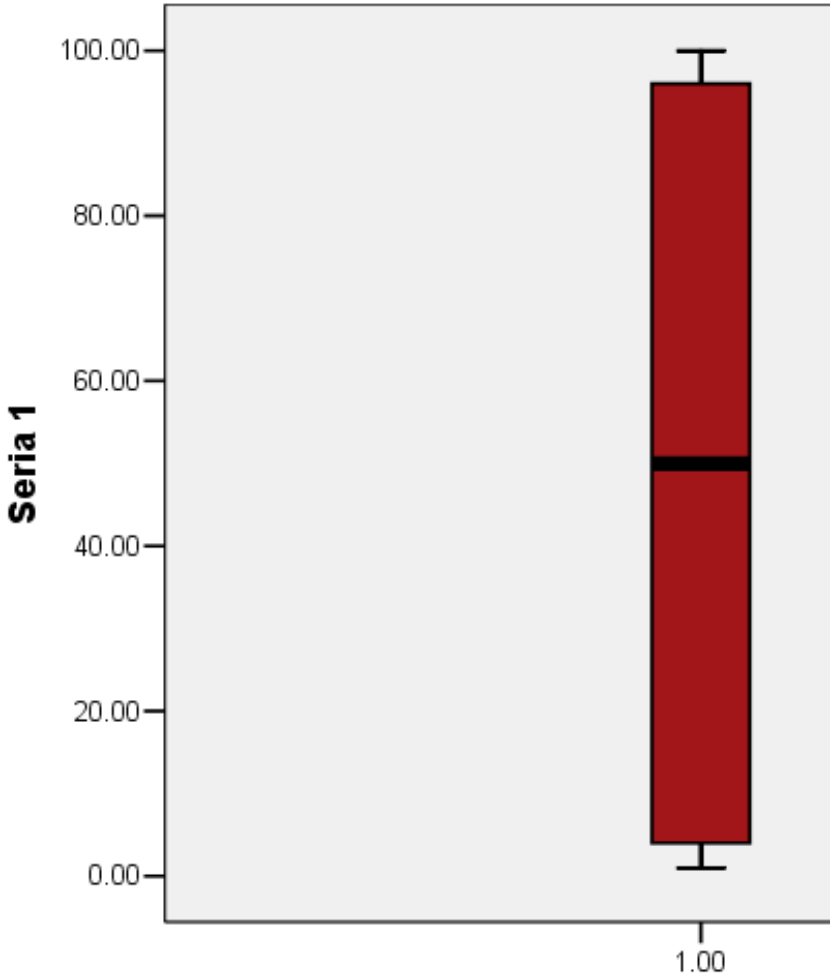
Cum prezentăm datele?

- Datele sunt:
  - Normal distribuite: media±dev. standard
  - Nu sunt normal distribuite: mediana (IQR)

Seria 1

1
1
2
3
5
6
6
7
93
94
94
95
97
98
98
100

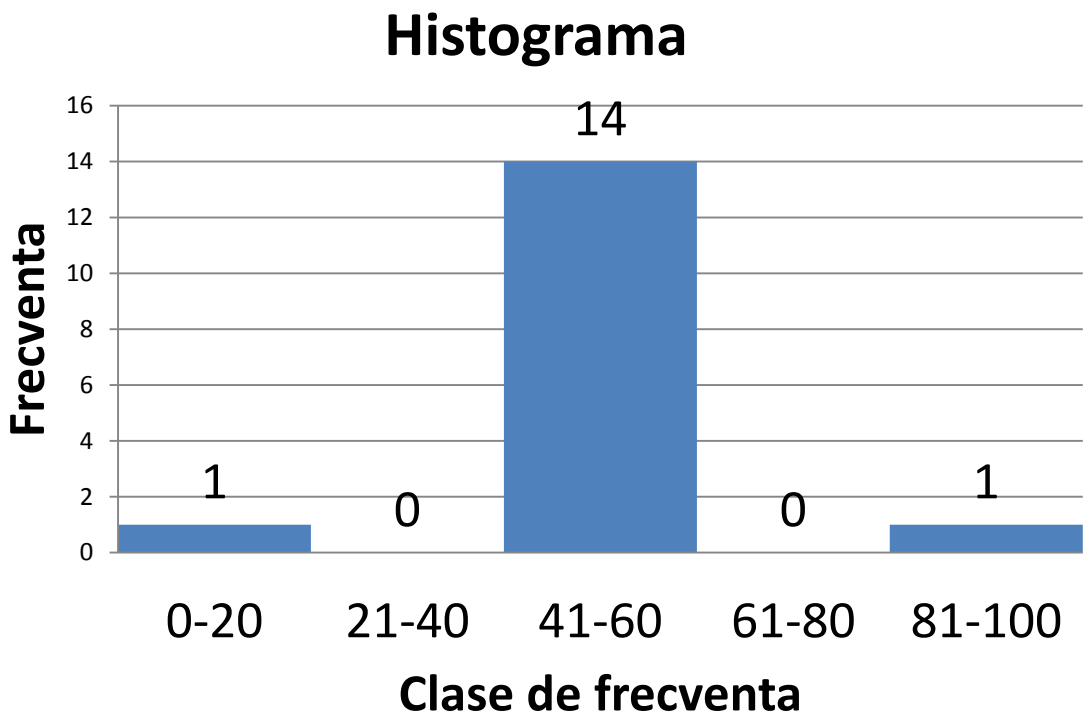


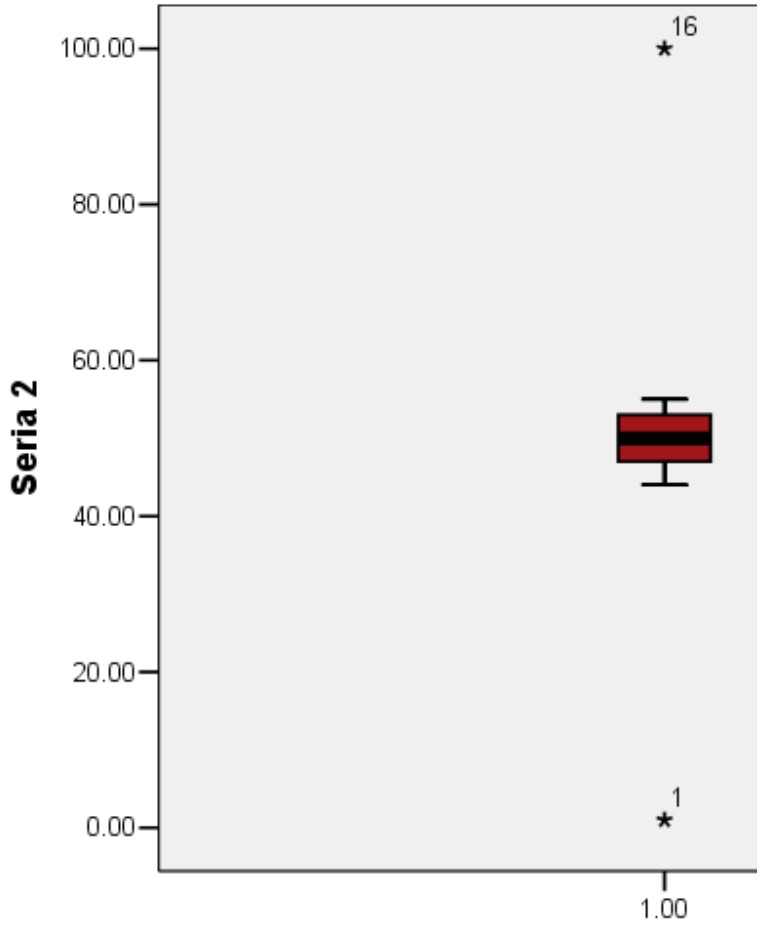


- Mediana (25 percentile-75 percentile)
- 50 (4-96)

Seria 2

1
44
45
46
48
48
49
50
50
51
52
52
54
55
55
100

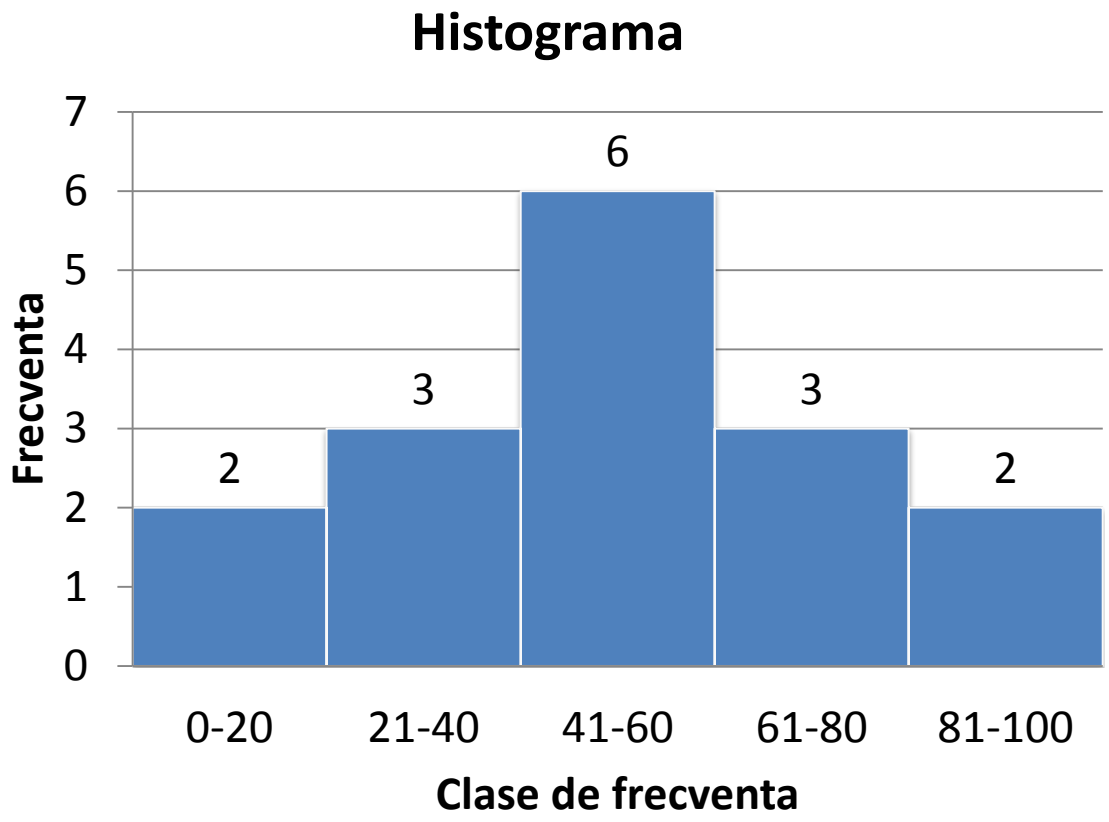


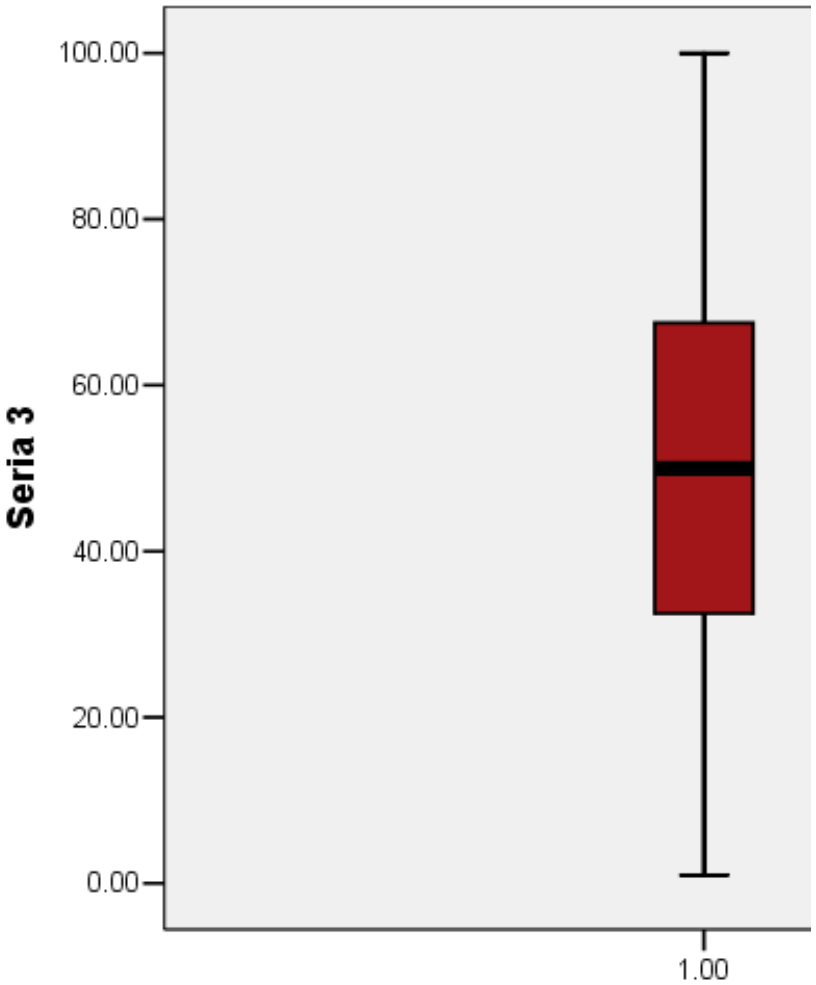


- Mediana (25 percentile-75 percentile)
- 50 (47-53)

Seria 3

1
11
24
29
36
41
45
49
51
55
59
64
71
76
88
100

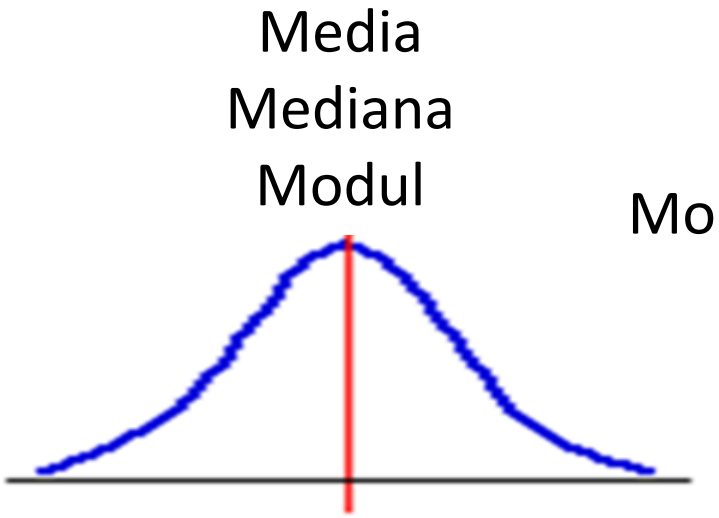




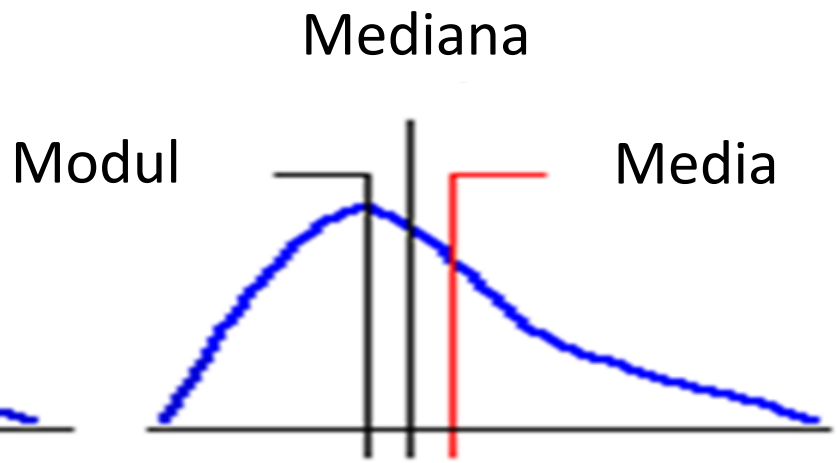
- Medie ± Deviaţia standard
- 50.00±26.71

Măsuri de centralitate - aplicabilitate

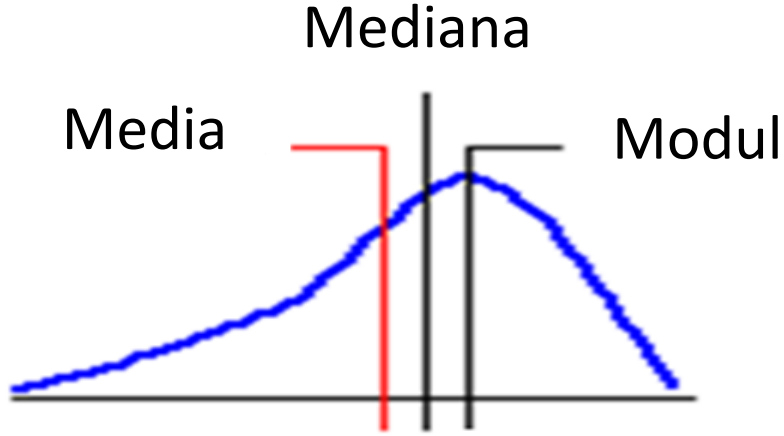
Cum să interpretăm media, mediana şi modulul:



Simetrică



Asimetrică la dreapta



Asimetrică la stânga

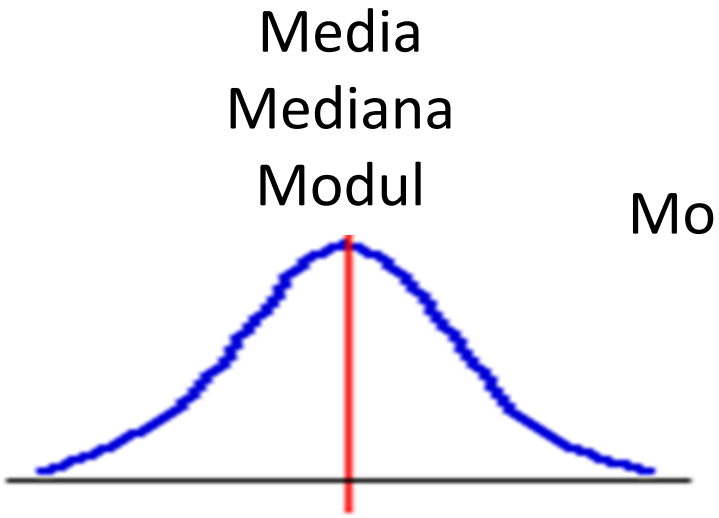
Distribuţie normală

Asimetria - Boltirea

- Asimetria – Asimetria datelor
- Boltirea – Platitudinea

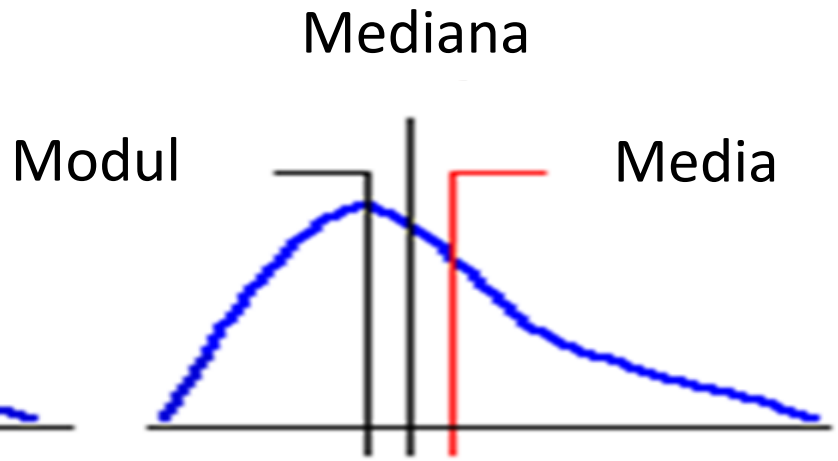
Asimetria (skewness)

Asimetria este o măsură care descrie distribuţia unei variabile cantitative continue



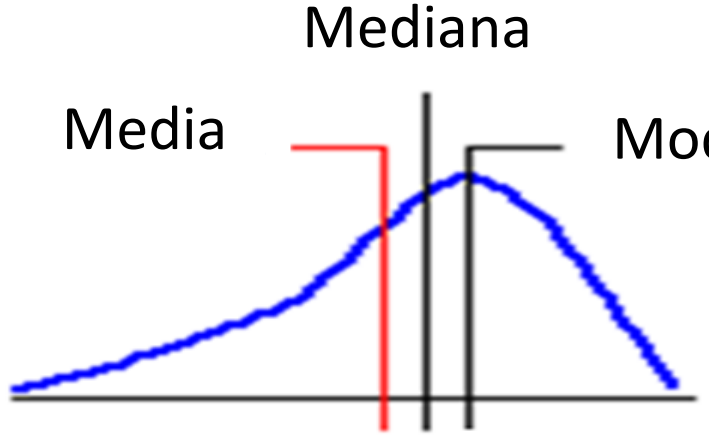
Simetrică

Asimetria=0



Asimetrică la dreapta

Asimetria>0

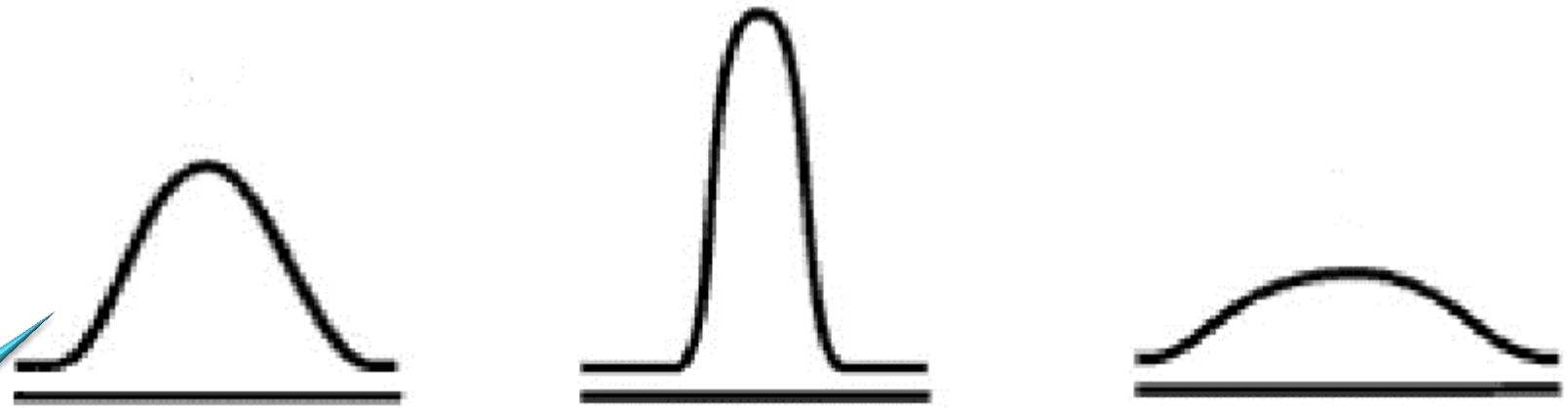


Asimetrică la stânga

Asimetria<0

Boltirea (kurtosis)

Boltirea este o măsură care descrie distribuţia unei variabile cantitative continue



Distribuţia normală

Mesocurtică
Boltirea=0

Leptocurtică
Boltirea>0

Platicurtică
Boltirea<0

Mulţumesc!