

Bondor Cosmina

Variabile aleatoare, distributii de probabilitate

A ALWAYS

S SEEK

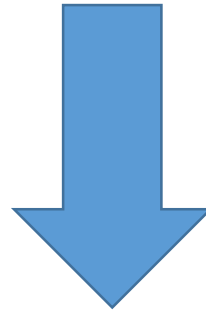
K KNOWLEDGE

Obiectivele cursului

- eșantionare
 - un eșantion reprezentativ al populației țintă într-un studiu
- distribuția de probabilitate a unei variabile aleatoare
- caracteristicile distribuției normale

Obiectiv: Prevalența obezității în populația România

- Cum realizăm studiul? cântărim toată populația țării
 - nu putem, costuri mari, timp îndelungat, lipsă de personal etc.
 - dar putem să cântărim 2000 de persoane



Eșantion



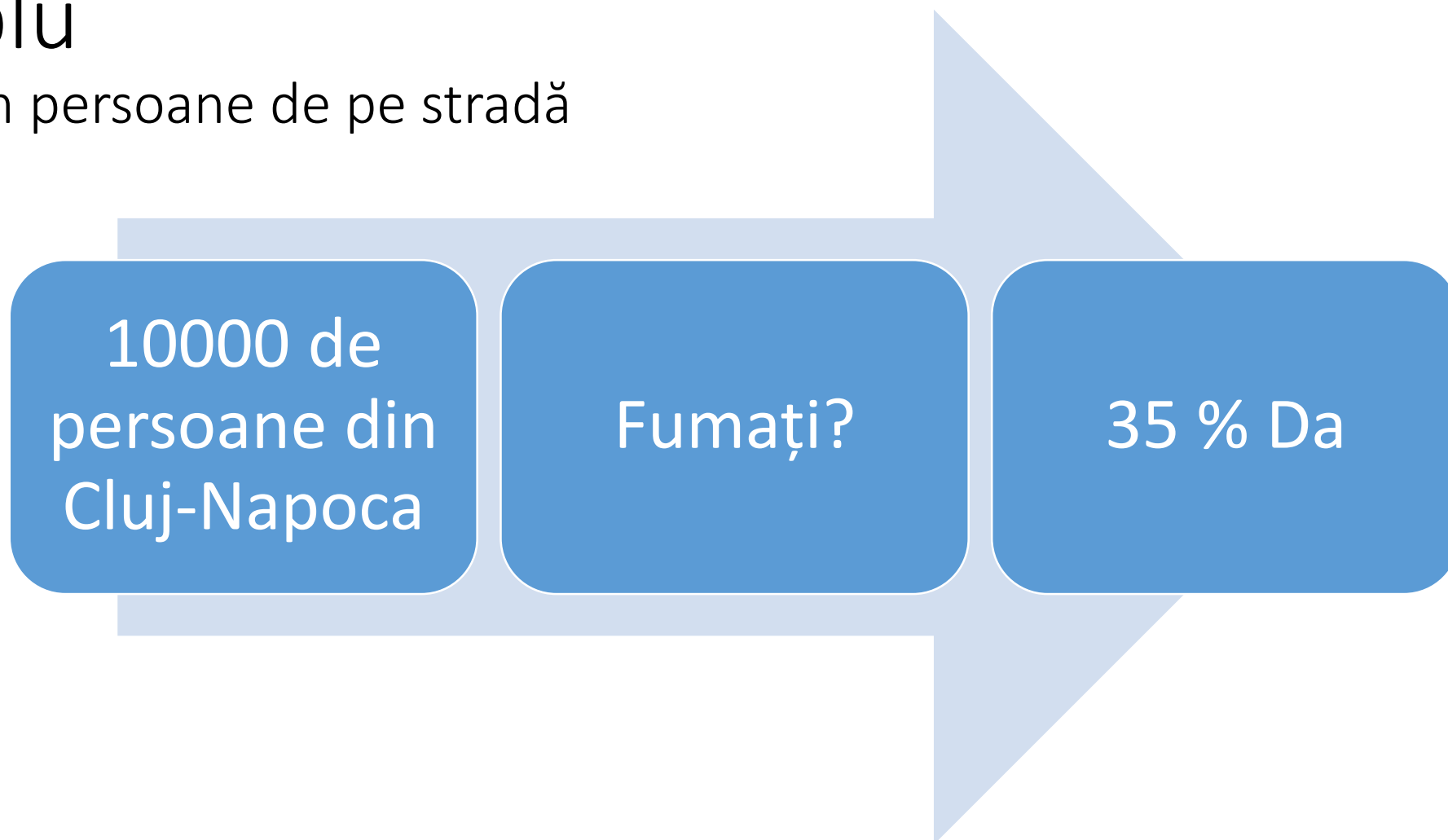
Populația



- din punct de vedere
 - statistic
 - o colecție de elemente care au aceeași caracteristică
 - in domeniul sănătății
 - pacienți
 - unități spitalicești

Exemplu

– selectăm persoane de pe stradă



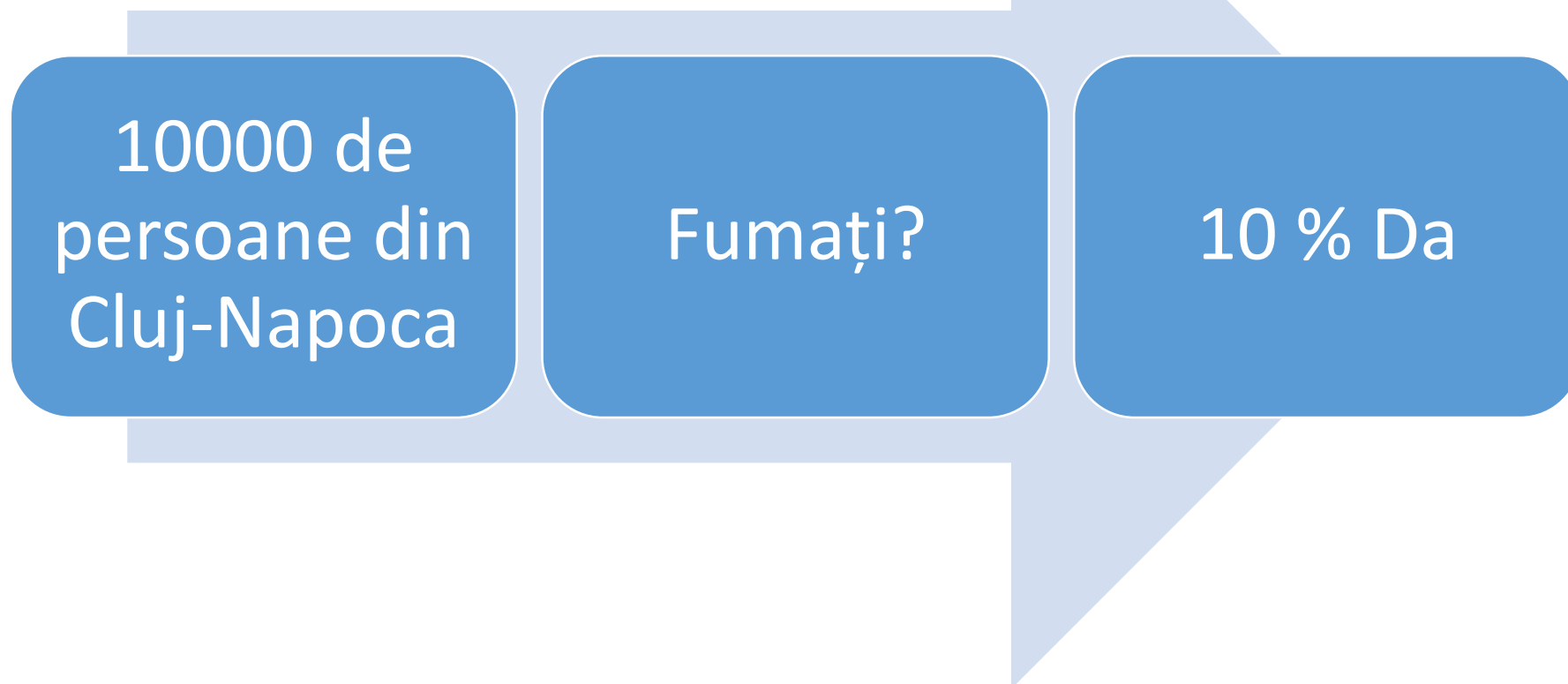
Generalizarea: În Cluj-Napoca probabilitatea ca o persoană să fumeze este 0,35

Avem 35% fumători?

Când se poate realiza?

Exemplu

– selectăm persoane de la sala de gimnastică



Selecția influențează rezultatul!

**Ca să realizăm o aproximație corectă a frecvenței fumatului în populația țintă -
selectăm un eșantion reprezentativ pentru populația din Cluj-Napoca**

Eroare (bias) de selecție

- Ex.
- obiectivul - numărul de fracturi în populația generală într-un an
 - selecție de indivizi de la clubul de ski
- obiectivul – numărul de persoane cu infertilitate
 - selecție indivizi care vin la laborator să testeze infertilitatea
 - aceștia suspectează că sunt infertili

Eșantion reprezentativ pentru populație



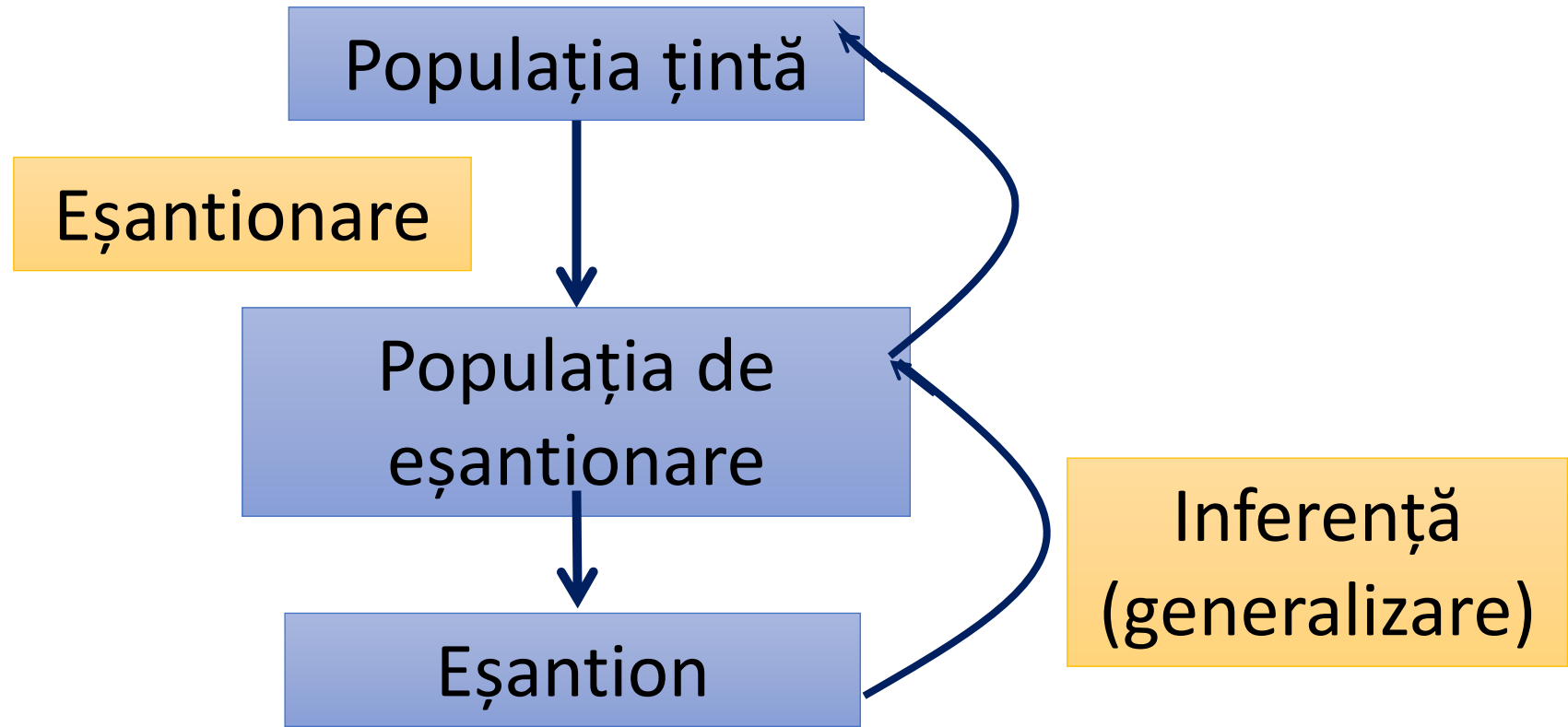
Selecție aleatoare



Selecție aleatoare

- Fiecare individ din populație are aceeași probabilitate de a fi selectat în eșantion
- Ex. Mergeți la primărie. Luați toate CNP-urile – extrageți aleator





Populație țintă – populația la care se dorește generalizarea rezultatelor studiului
Populația de eșantionare – populația din care a fost extras eșantionul

Condiția inferenței eşantion \rightarrow populație



Esantion reprezentativ (Selecția aleatoare)

- Când folosim termenul „eșantion” în contextul cercetării medicale
 - vom presupune că eșantionul a fost selectat aleator într-un mod corect



Metode de eșantionare

Probabilistică: fiecare subiect din populație are o probabilitate cunoscută de a fi selectat

- Eșantionare simplu randomizată
 - Subiecților li se atribuie un număr
 - Se extrag numere **aleatorii** din listă
- Eșantionare sistematică
 - tot **al k-lea individ** se alege pentru a fi inclus în eșantion
- Eșantionare stratificată
 - Populația este împărțită în straturi după **însușiri care nu sunt echiprobabile, dar care pot influența obiectivul studiului**, se extrage aleator din fiecare strat
- Eșantionare de tip cluster
 - Cluster= **arie delimitată geografic**
 - Delimitarea clusterelor, selectarea aleatorie a clusterelor
 - Selectare aleatorie a subiecților din fiecare cluster selectat

Metode de eșantionare

Non-probabilistică: probabilitatea unui individ de a fi selectat este necunoscută

- Convenient:
 - Participanții sunt selectați deoarece sunt accesibili
- Bulgărele de zăpadă:
 - Subiecții incluși în studiu vor aduce alți potențiali participanți
- Deliberat
 - Grup de tehnici de eșantionare care au la bază gândirea cercetătorului

Recensământ

- Recensământ
 - participă toată populația – nu necesită inferență statistică
 - se aplică metode ale statisticii descriptive

Variabila aleatoare,
distribuția de probabilitate

Eșantioane aleatorii

```
graph TD; A[Eșantioane aleatorii] --> B[Măsurători]; B --> C[Rezultatul imprevizibil]; C --> D[Rezultatul = variabilă aleatoare];
```

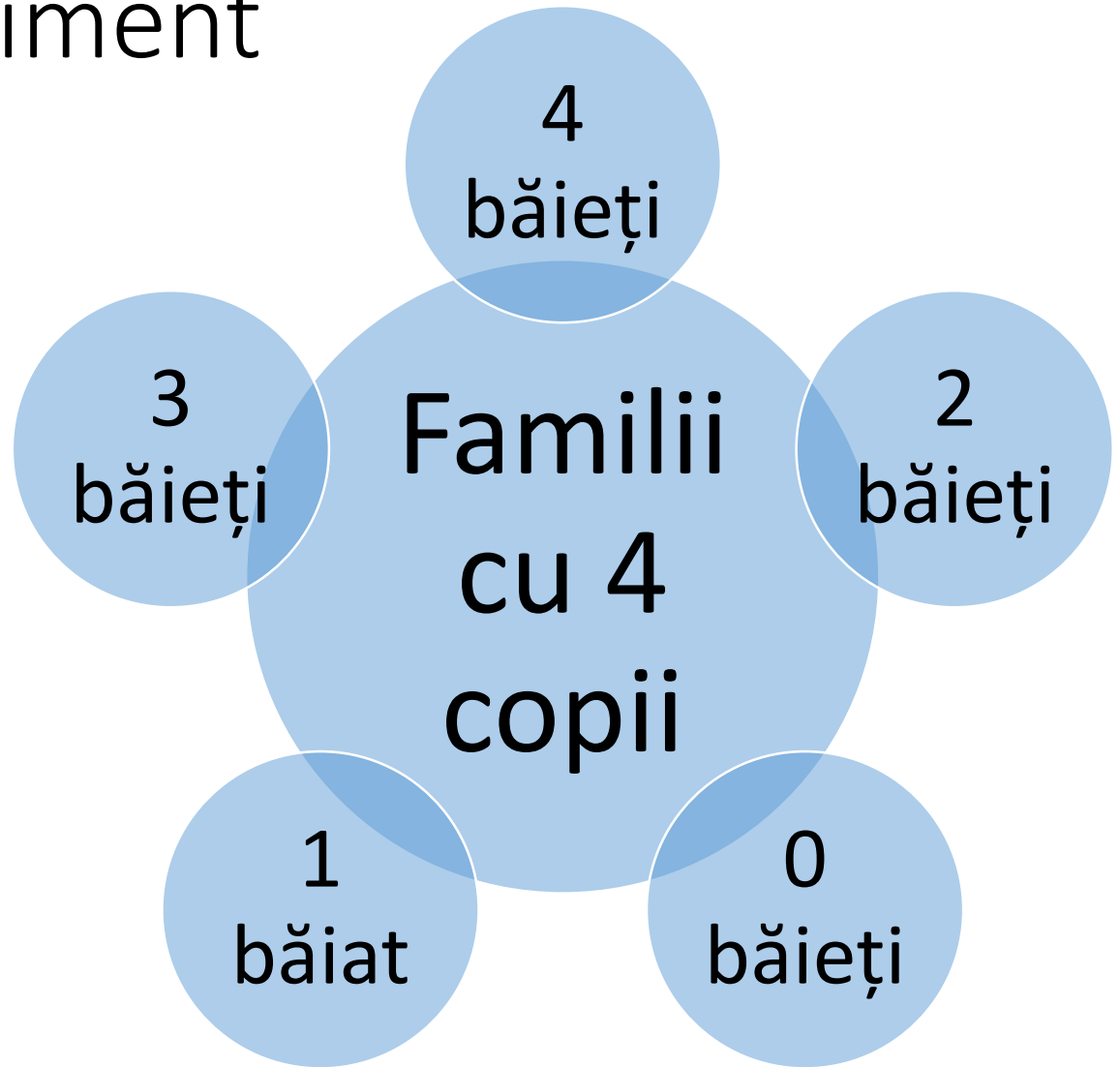
Măsurători

Rezultatul imprevizibil

Rezultatul = variabilă aleatoare

Probabilitatea ca
să se nască un
copil de sex
masculin ≈ 0.5
(50% dintre
cazuri)

Experiment

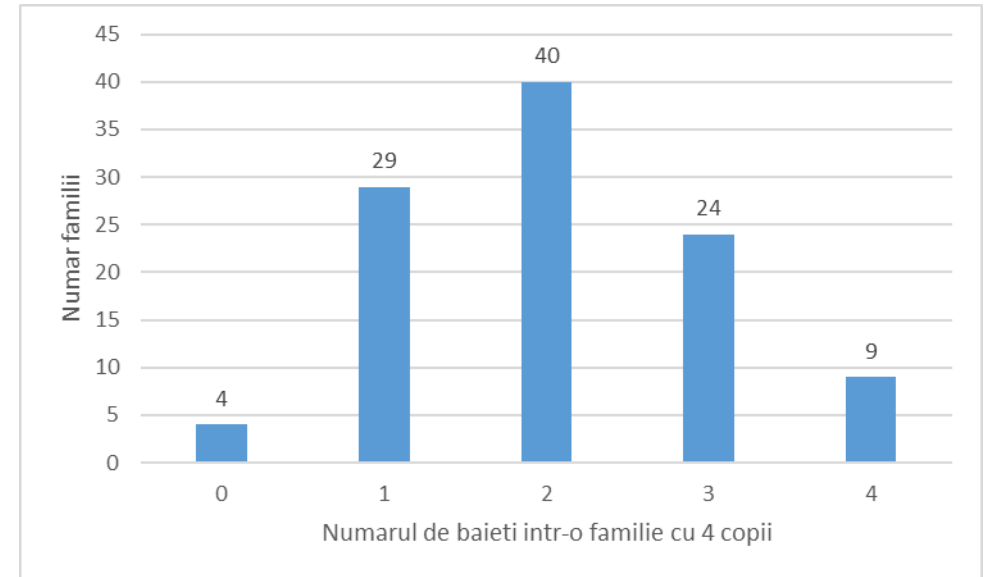


În **100 de familii cu 4 copii?**

In 100 de familii cu 4 copii?

Selectăm aleator **100 de familii cu 4 copii:**

Număr de băieți	0	1	2	3	4	Total
Nr. de familii	4	29	40	24	9	100



- Numarul de băieți într-o familie – **Variabila aleatoare**
- 0 băieți în 4 familii
- 1 băiat în 29 de familii
- 2 băieți în 40 familii
- 3 băieți în 24 de familii
- 4 băieți în 9 familii

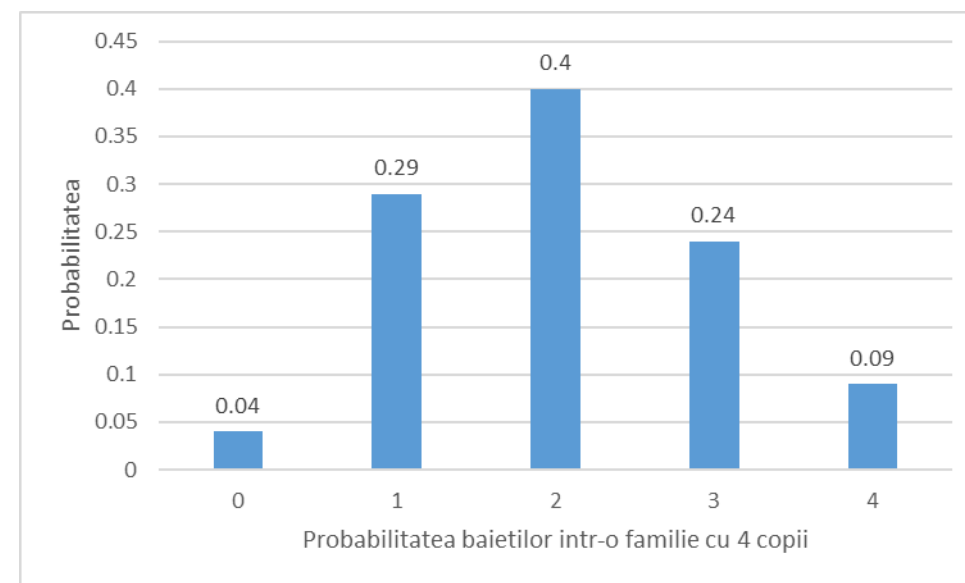
Distribuție de probabilitate

distribuție de frecvențe = distribuție de probabilitate

Distribuția de probabilitate

Numim **distribuție de probabilitate a variabilei X**
numărul de apariții a valorilor posibile a variabilei X

Număr de băieți	0	1	2	3	4	Total
Nr. de familii	4	29	40	24	9	100
Probabilitatea	0,04	0,29	0,40	0,24	0,09	1

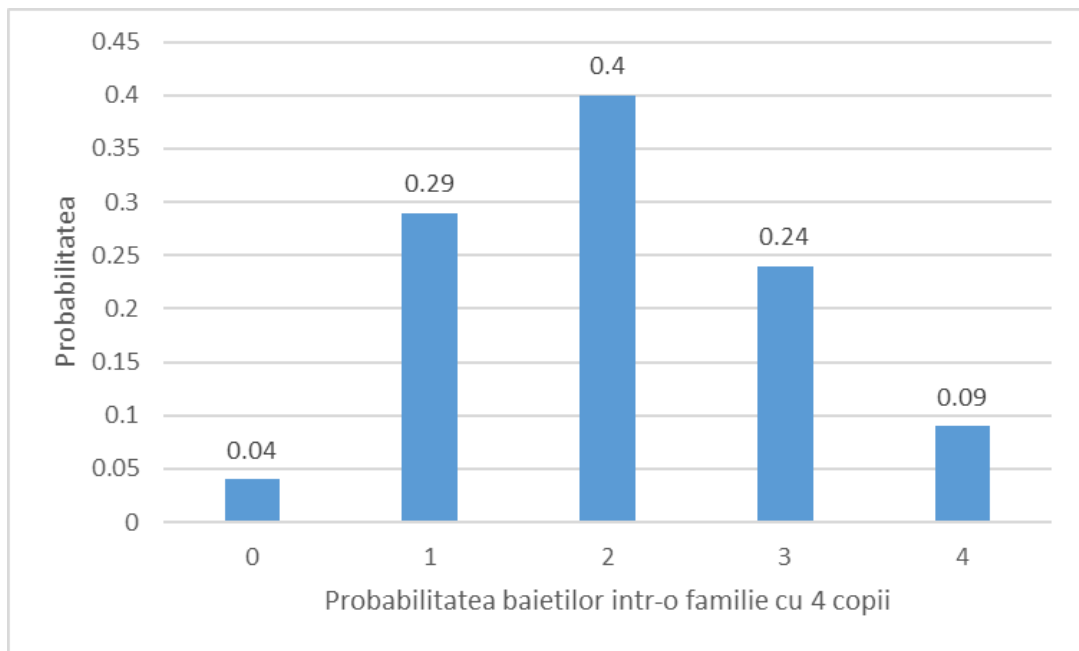


- Numărul de băieți într-o familie – **Variabila aleatoare**

Cum calculăm distribuția de probabilitate?

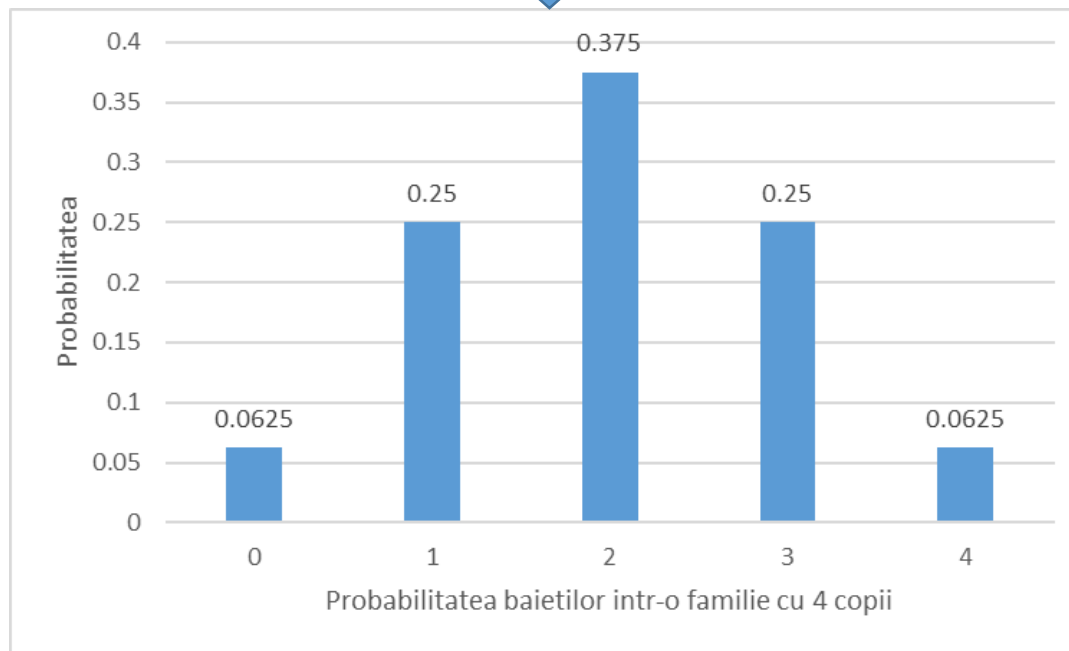
- Empiric

- eșantion, apoi inferență



- Teoretic

- Formulă
 - Aproximare cu o distribuție de probabilitate teoretică cunoscută



Ex. Distribuția de probabilitate teoretică a numărului de băieți în familiile cu 4 copii

Număr băieți	0	1	2	3	4	Total
Probabilitatea	0,0625	0,25	0,375	0,25	0,0625	1,00

- Cum a fost calculată?
 - modelată după așteptări,
 - un comportament “normal” = neinfluențat de diverși factori

Dacă ne interesează același obiectiv la clinica de infertilitate

- Familii cu gemeni:



Numarul de baieti	0	1	2	3	4	Total
	0,50	0	0	0	0,50	100

Cum aflăm distribuția de probabilitate?



Dacă variabila nu e continuă sau infinită

Empiric – experiment – distribuție de frecvențe

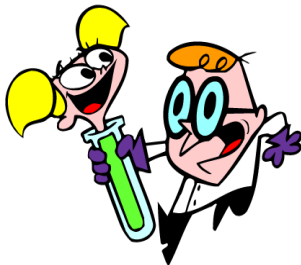


Dacă e ∞ sau continuă?

! suntem norocoși - găsim

Formulă

Regulă



Dacă nu suntem norocoși:

Modelăm (aproximăm) după o distribuție teoretică de probabilitate (cunoscută – una la care am fost norocoși)

Legi de distribuție

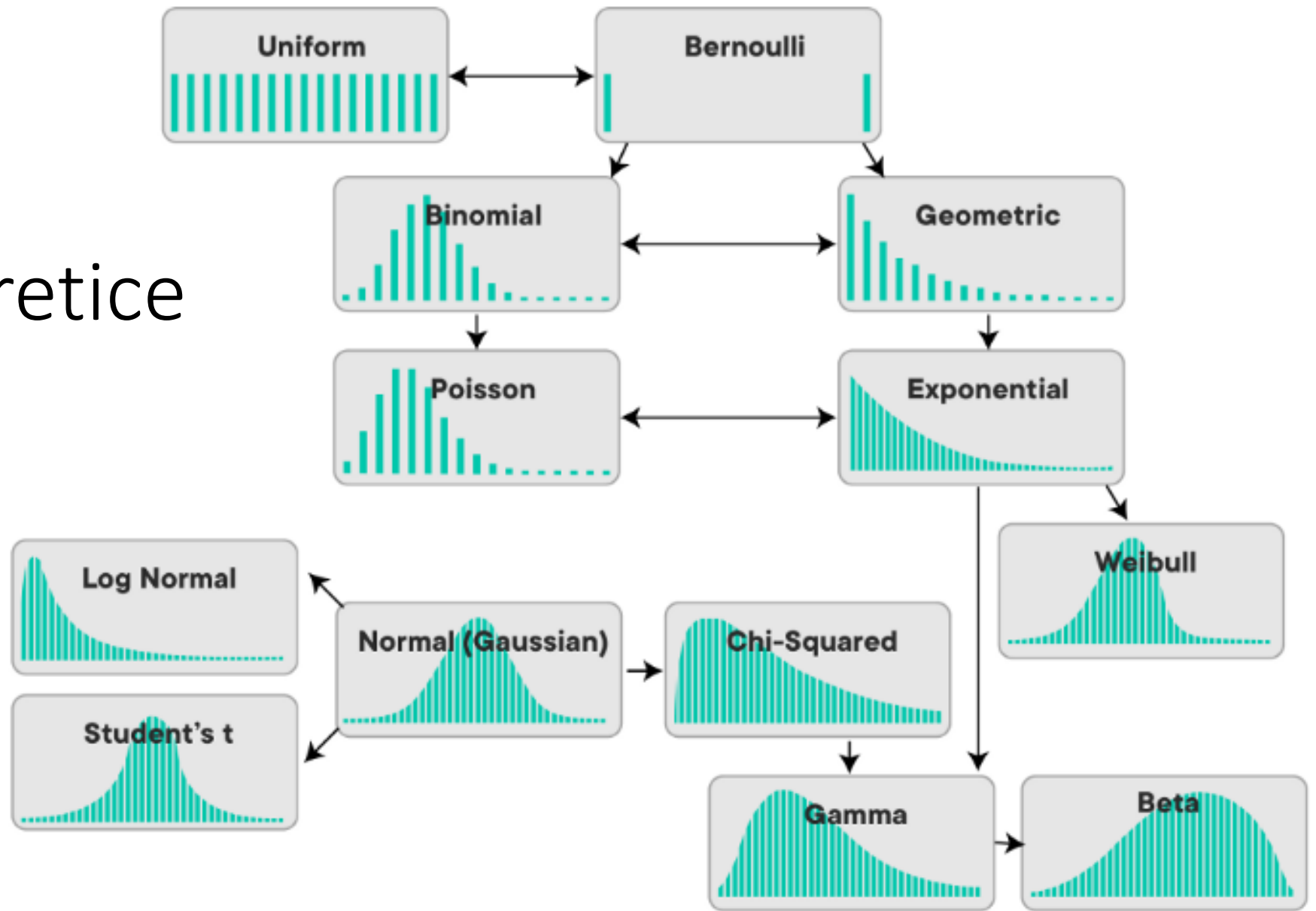
(distribuții de probabilitate)



Cele mai
cunoscute



Distribuții teoretice



- Au o funcție cunoscută, medie și deviație standard deductibilă



LEGEA NORMALĂ

- variabilă aleatoare continuă
- funcție de probabilitate - alură de clopot
 - curba normală
 - curba lui Gauss
- Această distribuție depinde de doi parametri:
 - media aritmetică μ
 - abaterea standard (varianța) σ
- densitate de probabilitate:

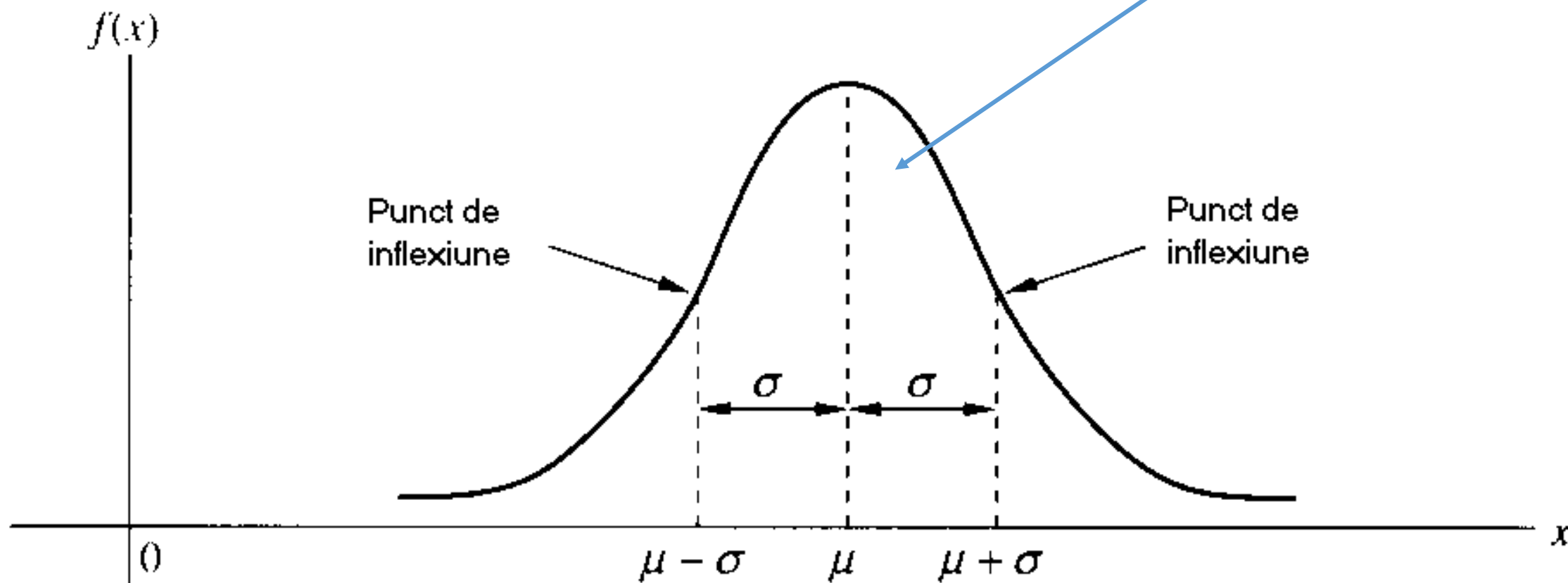
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



1777–1855

LEGEA NORMALĂ

Aria de sub curbă este 1, ca la orice distribuție de probabilitate



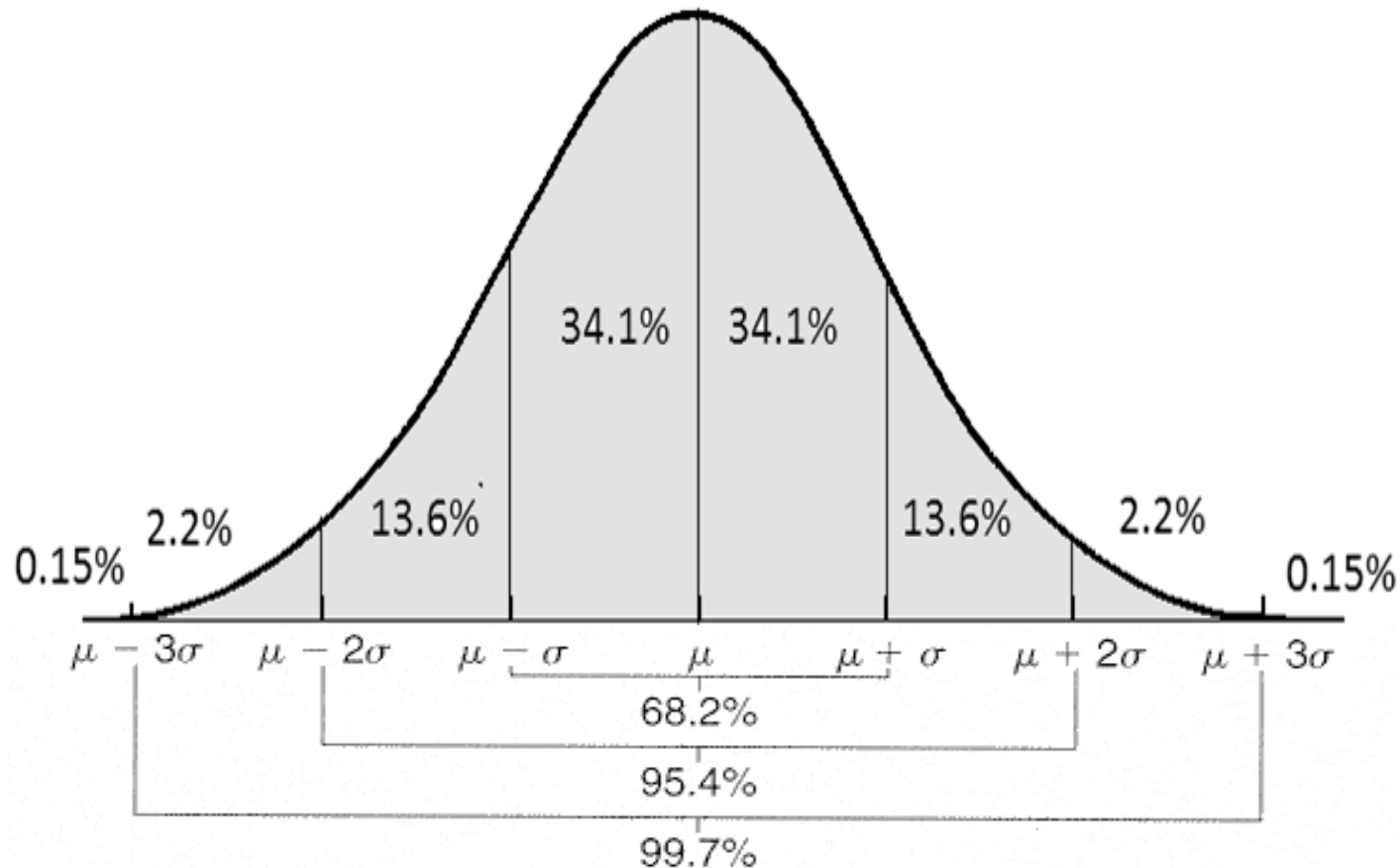
σ deviația standard (varianța) este distanța dintre medie și punctul de inflexiune (acolo unde curba se schimbă din concavă în convexă)

Proprietăți:

În intervalul medie \pm abatere standard - minim 68,2% din observații;

În intervalul medie ± 2 * abatere standard - minim 95,4% din observații;

În intervalul medie ± 3 * abatere standard - minim 99,7% din observații.





Aplicații: Cum este distribuția datelor?

Dacă aceste condiții sunt îndeplinite

- media \approx mediana \approx modulul
- simetria ≈ 0
- boltirea ≈ 0
- cvartilele 1 și 3 simetrice față de media aritmetică
- În intervalul $\text{medie} \pm \text{abatere standard}$ \ni minim 68,2% din observații;
- În intervalul $\text{medie} \pm 2 * \text{abatere standard}$ \ni minim 95,4% din observații;
- În intervalul $\text{medie} \pm 3 * \text{abatere standard}$ \ni minim 99,7% din observații,
- atunci distribuția datelor obținute empiric se apropie de distribuția normală

Exemplu – Seria 1



Seria 1

1

1

2

3

5

6

6

7

93

94

94

95

97

98

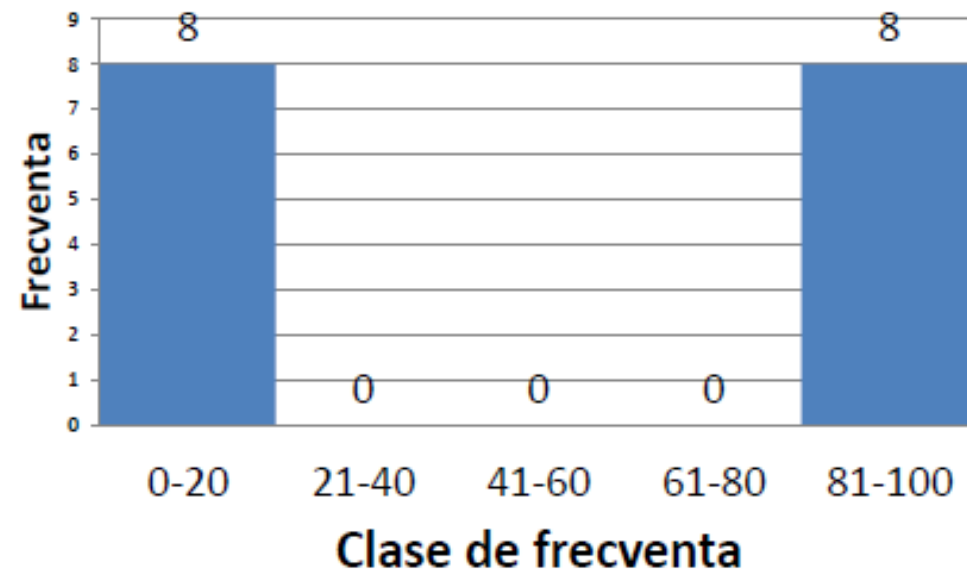
98

100

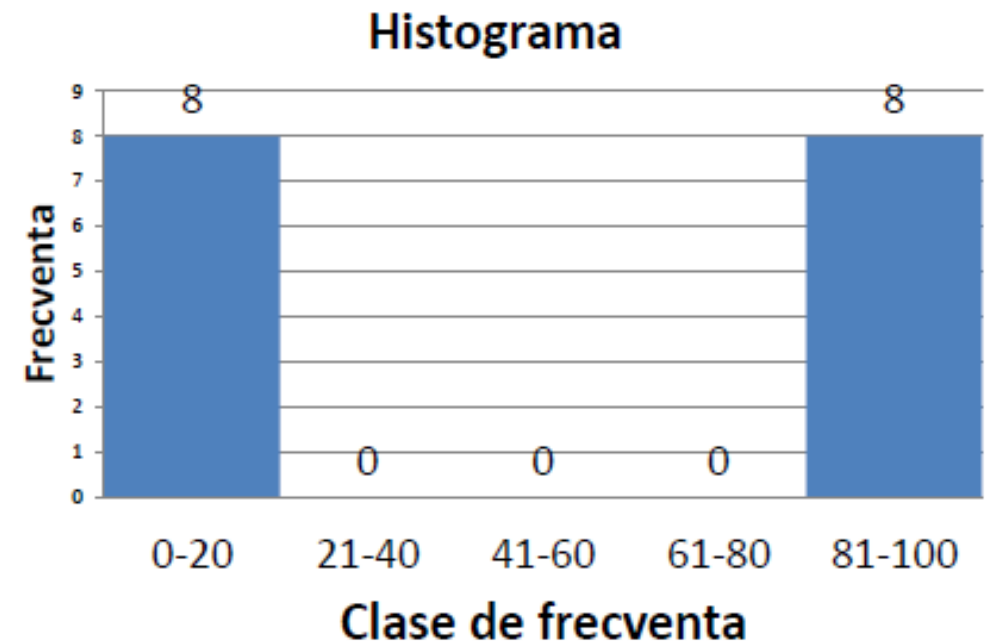
- Media aritmetică = 50
- Mediana = 50
- **Modul – nu are**
- Deviația standard = 47,70
- Cvarțila 1 = 4,5
- Cvarțila 3 = 95,5
- Simetria = 0,0002
- **Boltirea = -2,29**

Ne arată diferențe
mari față de
distribuția normală

Histograma



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



- Media aritmetică = 50
- Deviația standard = 47,70

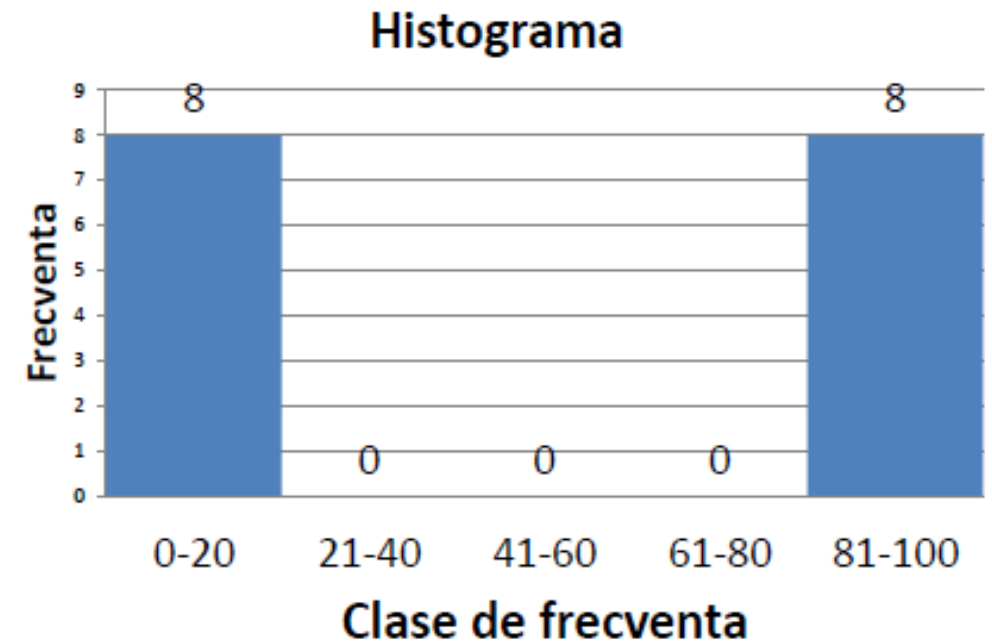
Deviația standard foarte mare,
concluzie: există date în cele două
extreme



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

16

- Media aritmetică = 50
- Deviația standard = 47,70
- Media - deviația standard = $50 - 47,7 = 2,3$
- Media + deviația standard = $50 + 47,7 = 97,7$
- intervalul media \pm deviația standard = $[50 - 47,7; 50 + 47,7] = [2,3; 97,7]$



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

- Media aritmetică = 50
- Deviația standard = 47,70

Media \pm deviația standard = $[50 - 47,7; 50 + 47,7] = [2,3; 97,7]$

16

- In intervalul $[2,3; 97,7]$ sunt 10 date, adica **62,5%** din date

$$10/16 * 100 = 62,5$$



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

Ca să fie distribuție normală:

in intervalul media \pm deviația standard sunt **minim 68,3% din date**

in intervalul media ± 2 *deviația standard sunt minim 95,4% din date

in intervalul media ± 3 *deviația standard sunt minim 99,7% din date

- Media aritmetică = 50
- Deviația standard = 47,70
- Media \pm deviația standard = [2,3; 97,7]
- In intervalul [2,3; 97,7] sunt 10 date, adica **62,5%** din date



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

Ca să fie distribuție normală:

in intervalul media \pm deviația standard sunt **minim 68,3% din date**

in intervalul media ± 2 *deviația standard sunt minim 95,4% din date

in intervalul media ± 3 *deviația standard sunt minim 99,7% din date

- Media aritmetică = 50
- Deviația standard = 47,70
- Media \pm deviația standard = [2,3; 97,7]
- In intervalul [2,3; 97,7] sunt 10 date, adica **62,5%** din date

62,5% < 68,3%, deci distributia nu este normala



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

16

-45,39 – date medicale
negative nu prea are sens

Medie $\pm 2 \cdot$ deviația standard = $[50 - 2 \cdot 47,7; 50 + 2 \cdot 47,7] = [-45,39; 145,39]$

in intervalul $[-45,39; 145,39]$ sunt 16 valori, e.g. $16/16 = 100\%$ dintre date



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

Ca să fie distribuție normală:

in intervalul $\text{media} \pm \text{deviația standard}$ sunt minim 68,3% din date

in intervalul $\text{media} \pm 2 \cdot \text{deviația standard}$ sunt **minim 95,4%** din date

in intervalul $\text{media} \pm 3 \cdot \text{deviația standard}$ sunt minim 99,7% din date

16

Medie $\pm 2 \cdot \text{deviația standard}$ = $[50 - 2 \cdot 47,7; 50 + 2 \cdot 47,7] = [-45,39; 145,39]$

in intervalul $[-45,39; 145,39]$ sunt 16 valori, adica $16/16 = 100\%$ dintre date

100% > 95,4 proprietatea e îndeplinită pentru acest interval



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

Ca să fie distribuție normală:

in intervalul media \pm deviația standard sunt minim 68,3% din date

in intervalul media ± 2 *deviația standard sunt minim 95,4% din date

in intervalul media ± 3 *deviația standard sunt minim **99,7%** din date

16 Media aritmetică = 50

Deviația standard = 47,70

Media \pm deviația standard = [2,3; 97,7] cu 62,5% dintre date

Mean ± 2 *st.dev = [-45,39; 145,39] sunt 16 valori, adica 100% dintre date

Mean ± 3 *st.dev = [50-3*47,7; 50+3*47,7] = [-93,09; 193,09] sunt 16 valori,
adică 16/16 = **100%** dintre date

100% > 99,7 proprietatea e îndeplinită pentru acest interval



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

Ca să fie distribuție normală:

in intervalul media \pm deviația standard sunt minim **68,3%** din date

in intervalul media ± 2 *deviația standard sunt minim 95,4% din date

in intervalul media ± 3 *deviația standard sunt minim 99,7% din date

16 Media aritmetică = 50

Deviația standard = 47,70

Media \pm deviația standard = [2,3; 97,7] cu 10 valori, adică **62,5%** dintre date

Mean ± 2 *st.dev = [-45,39; 145,39] sunt 16 valori, adica 100% dintre date

Mean ± 3 *st.dev = [-93,09; 193,09] sunt 16 valori, adică 16/16 = 100% dintre date

Distribuția nu este apropiată de cea normală



Seria 1

1

1

2

3

5

6

6

7

93

94

94

95

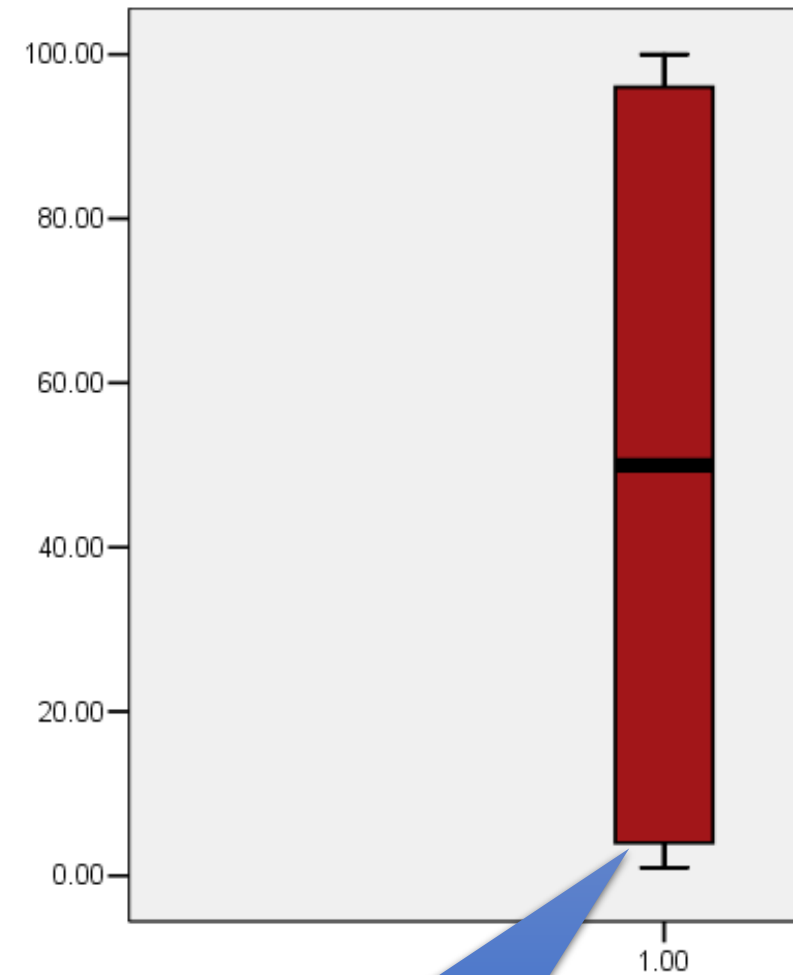
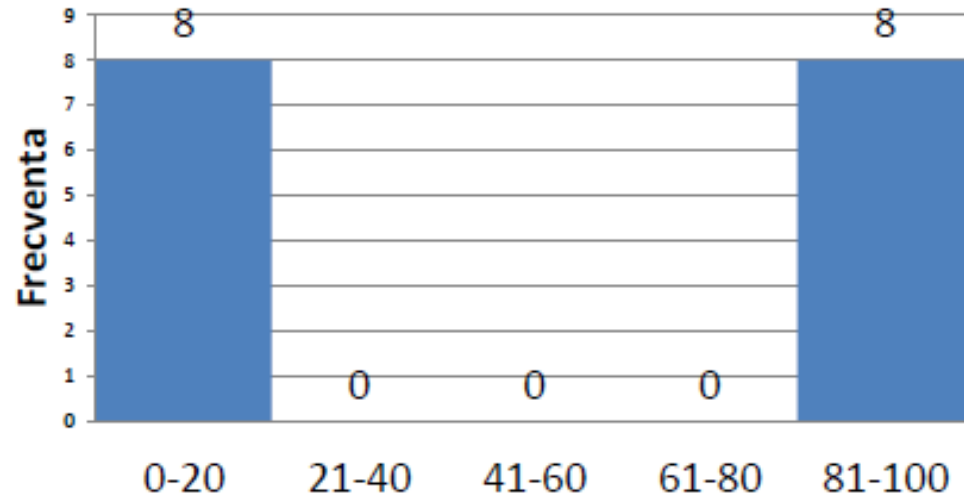
97

98

98

100

Histograma



Între minim și percentila 25 este o distanță mică = în acest interval avem multe date, comparativ cu intervalul următor



Exemplu – Seria 2



Seria 2

1

44

45

46

48

48

49

50

50

51

52

52

54

55

55

100

Media aritmetică = 50

Mediana = 50

Modul = multimodală

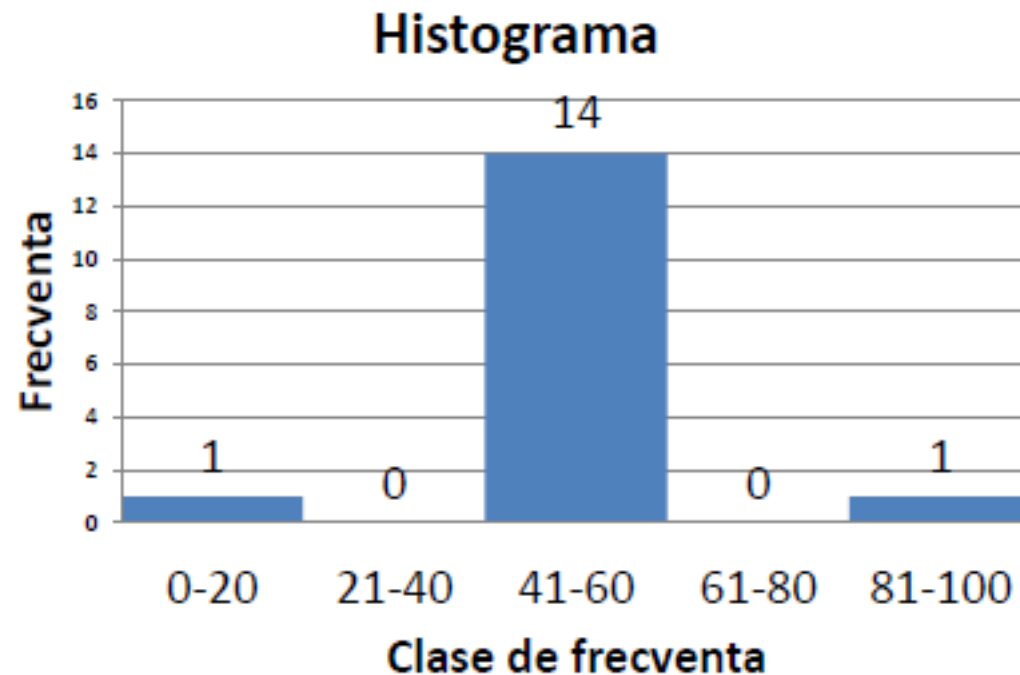
Deviația standard = 18,37

Cvartila 1 = 47,5

Cvartila 3 = 52,5

Simetria = 0,09

Boltirea = 6,81



Ne arată diferențe
mari față de
distribuția normală

Seria 2

1

44

45

46

48

48

49

50

50

51

52

52

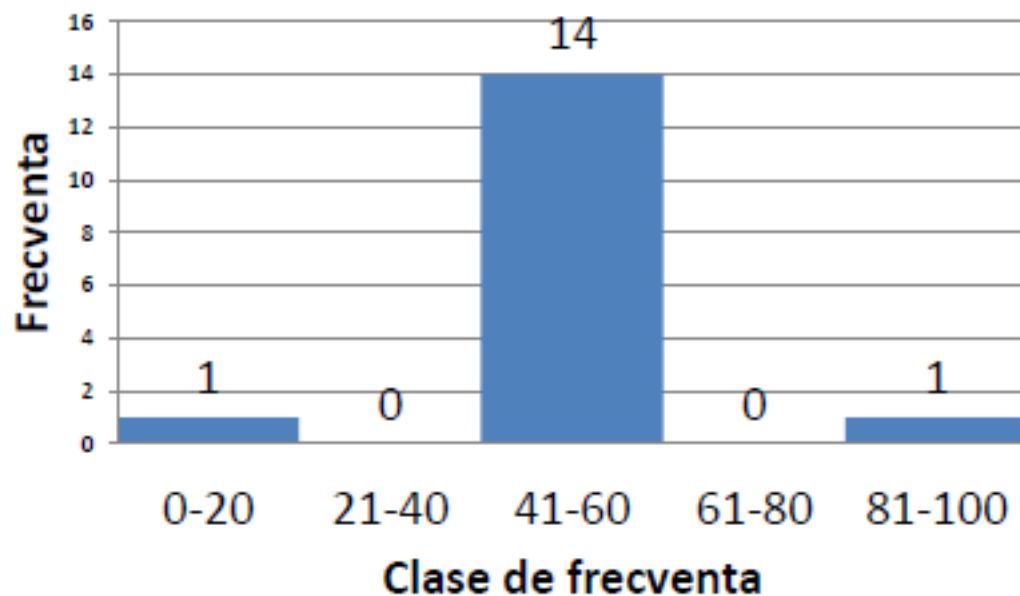
54

55

55

100

Histograma



Media aritmetică = 50
Deviația standard = 18,37

Ca să fie distrib. normală:
Minim 68,3% din date
Minim 95,4% din date
Minim 99,7% din date

Deviația standard este mică,
concluzie: cazurile sunt
aproprite de medie



Seria 2
1
44
45
46
48
48
49
50
50
51
52
52
54
55
55
100

Media aritmetică = 50
 Deviația standard = 18,37

Media \pm dev.st = $[50-18,37; 50+18,37] = [31,63; 68,37]$

16

in intervalul $[31,63; 68,37]$ sunt 14 valori, adica $14/16 = 87,5\%$ din date



Seria 2
1
44
45
46
48
48
49
50
50
51
52
52
54
55
55
100

16

Media aritmetică = 50
 Deviația standard = 18,37

Media \pm dev.st = $[50-18,37; 50+18,37] = [31,63; 68,37]$
 in intervalul $[31,63; 68,37]$ sunt 14 valori, adica $14/16 = 87,5\%$ din date

87,5 > 68,3, deci există minim 68,3% din date

Ca să fie distrib. normală:
 Minim 68,3% din date
 Minim 95,4% din date
 Minim 99,7% din date



Seria 2
1
44
45
46
48
48
49
50
50
51
52
52
54
55
55
100

16

Media aritmetică = 50

Deviația standard = 18,37

Media \pm dev.st = [31,63; 68,37] sunt 87,5% din date

Media ± 2 *dev.st. = [50-2*18,37; 50+18,37] = [13,26; 86,74] sunt tot 14 date,
adica 14/16 = **87,5%** din date, **mai putine** decat 95,4%
deci **seria 2 nu este distribuita normal**

Media ± 3 *dev.st. = [-5,11; 105,11] sunt 100% din date

Ca să fie distrib. normală:

Minim 68,3% din date

Minim 95,4% din date

Minim 99,7% din date



Seria 2
1
44
45
46
48
48
49
50
50
51
52
52
54
55
55
100

16

Media aritmetică = 50

Deviația standard = 18,37

Media \pm dev.st = [31,63; 68,37] sunt 14 valori - 87,5% din date

Media ± 2 *dev.st. = [13,26; 86,74] sunt 14 valori - 87,5% din date

Media ± 3 *dev.st. = [-5,11; 105,11] sunt 16 valori - 100% din date

Ca să fie distrib. normală:

Minim 68,3% din date

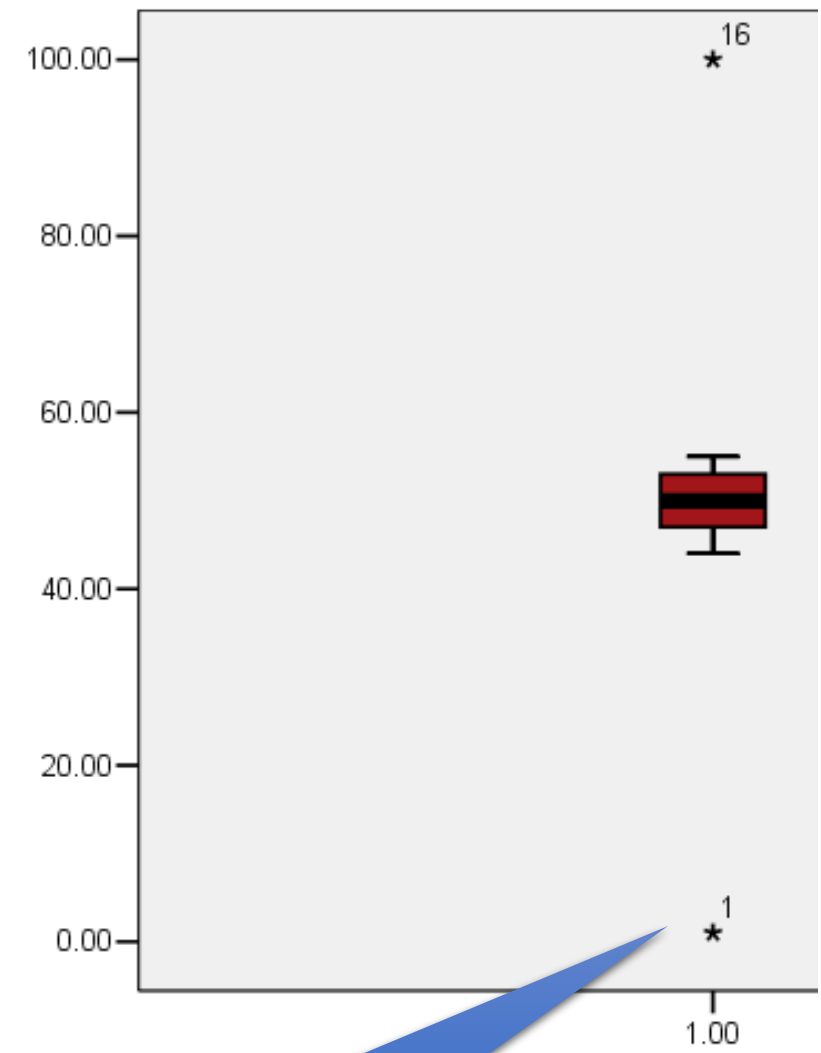
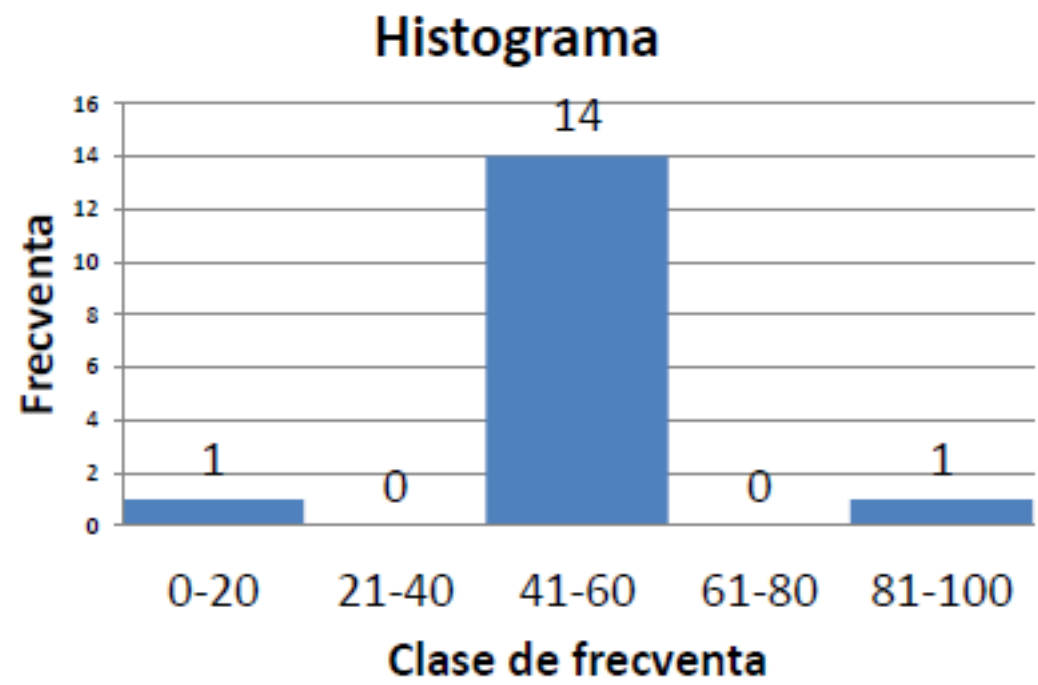
Minim 95,4% din date

Minim 99,7% din date

Distribuția nu este apropiată de cea normală



Seria 2
1
44
45
46
48
48
49
50
50
51
52
52
54
55
55
100



Caz extrem



Exemplu – Seria 3



Seria 3

1

11

24

29

36

41

45

50

50

55

59

64

71

76

88

100

Media aritmetică = 50

Mediana = 50

Modul = 50

Deviația standard = 26,71

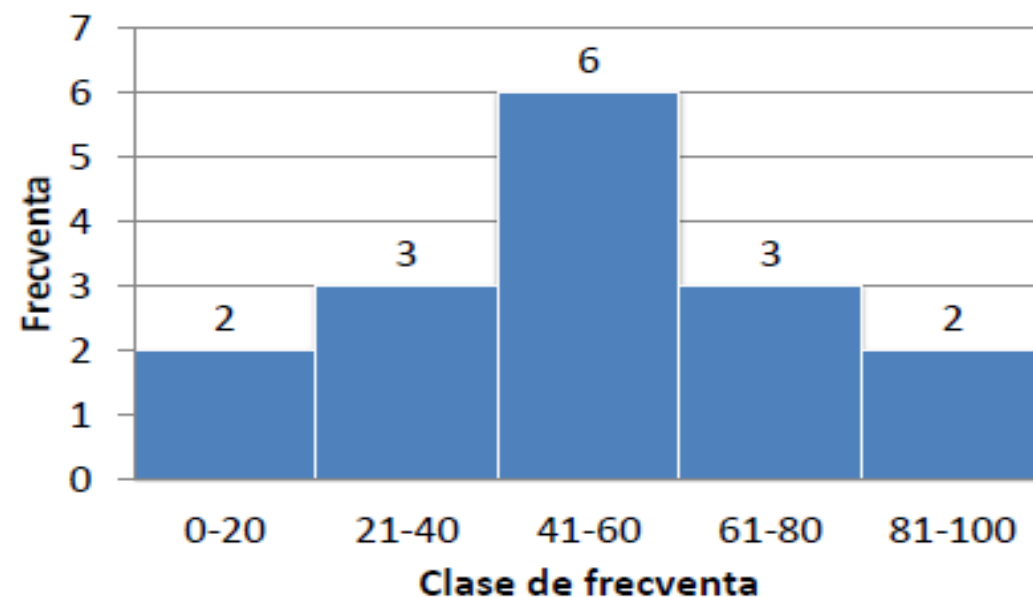
Cvartila 1 = 34,25

Cvartila 3 = 65,75

Simetria = 0,01

Boltirea = -0.23

Histograma



Distribuția este apropiată
de cea normală

Seria 3

1

11

24

29

36

41

45

49

51

55

59

64

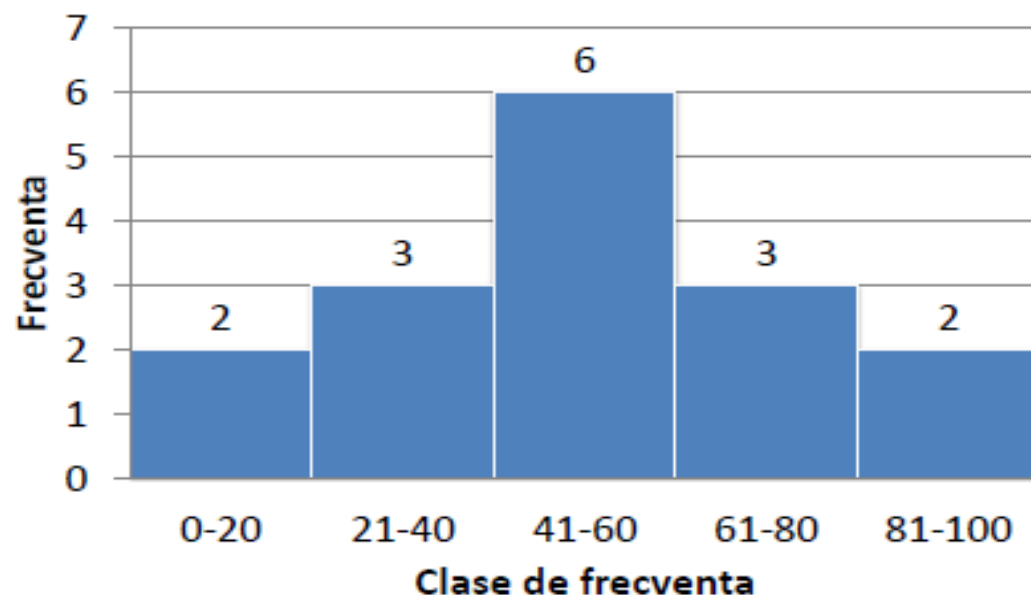
71

76

88

100

Histograma



Ca să fie distrib. normală:

Minim 68,3% din date

Minim 95,4% din date

Minim 99,7% din date

Distribuția este apropiată
de cea normală

Media aritmetică = 50

Deviația standard = 26,71

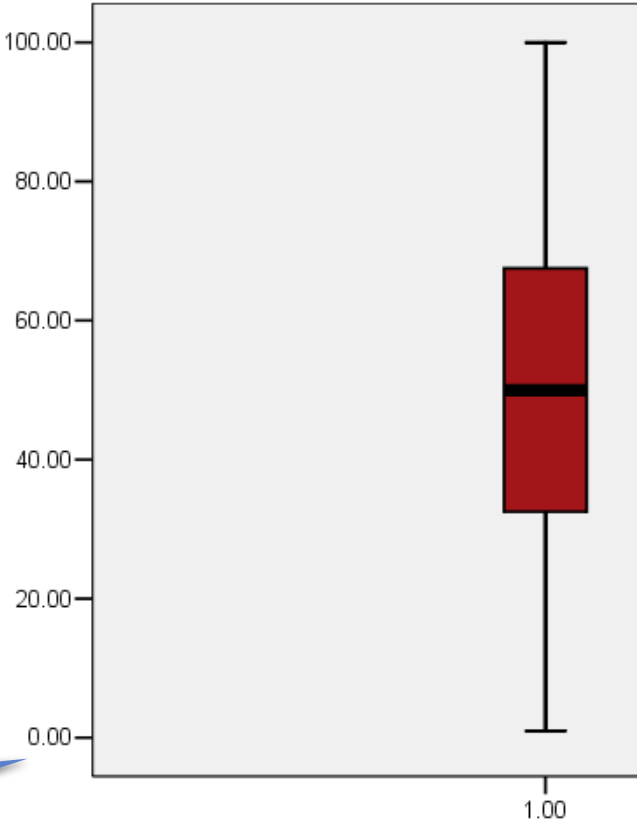
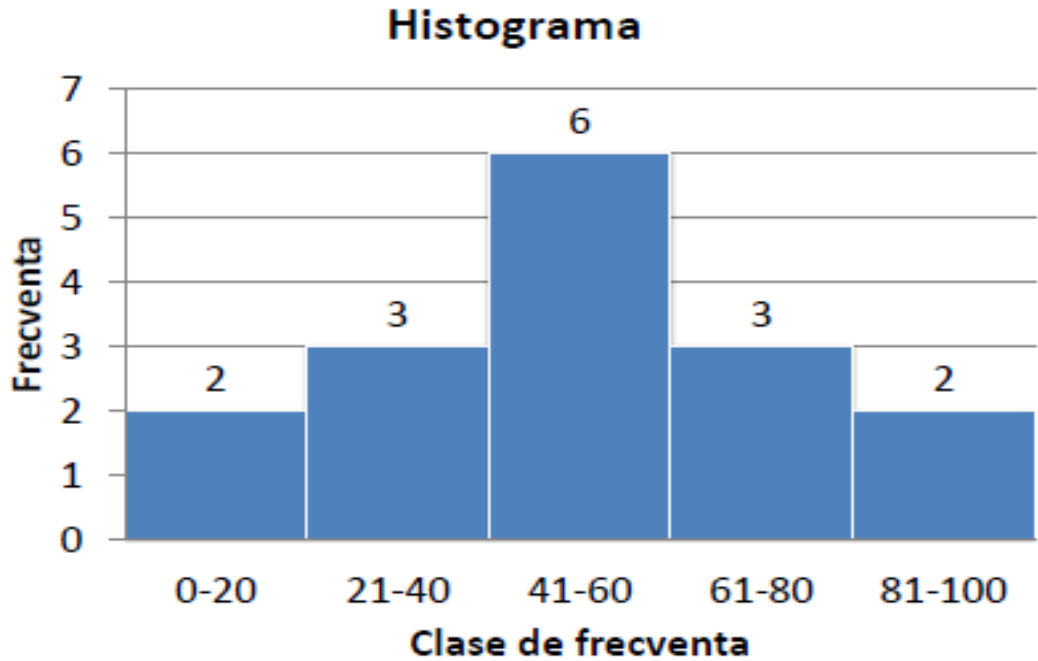
Media \pm dev.st. = [23,28; 76,72] sunt 87,5% din date

Media $\pm 2 \cdot$ dev.st. = [-3,43; 103,43] sunt 100% din date

Media $\pm 3 \cdot$ dev. st. = [-30,15; 130,15] sunt 100% din date



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



Distribuția este apropiată de cea normală



Distribuția paranormală

- Mulțumesc!!!

