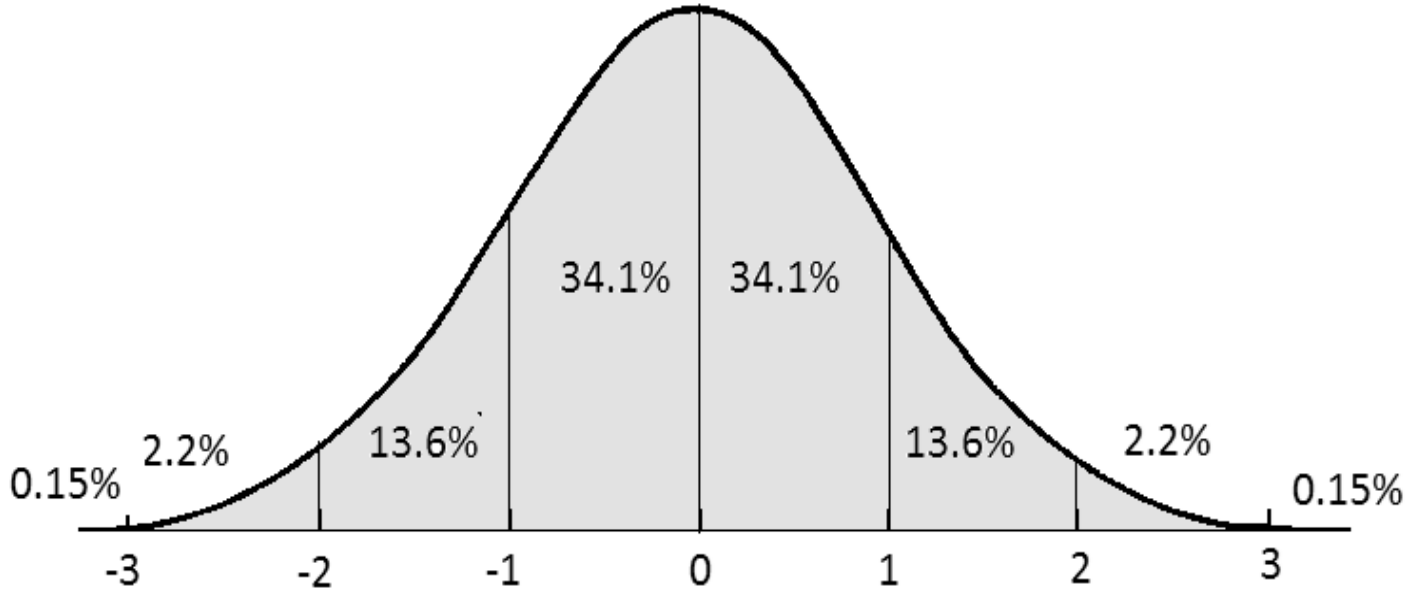


Bondor Cosmina

Variabile aleatoare, distributii de probabilitate

A ALWAYS

S SEEK

K KNOWLEDGE

Obiectivele cursului

După parcurgerea acestui curs studenții vor putea:

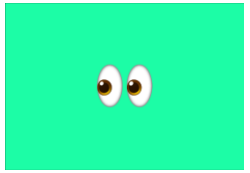
- să realizeze o eșantionare
  - un eșantion reprezentativ al populației țintă
- să identifice erori de selecție
- să determine distribuția de probabilitate a unei variabile
- să identifice caracteristicile distribuției normale
- să calculeze diferite valori numerice asociate distribuției normale

Legendă  de ținut minte  pentru pasionați  important pentru înțelegerea noțiunilor ce urmează a fi prezentate

Populație, eșantion, eșantion reprezentativ

Populația

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



https://app.wooclap.com/CURS7MGRO?from=instruction-slide

Obiectiv:

Frecvența diabetului în populația din România

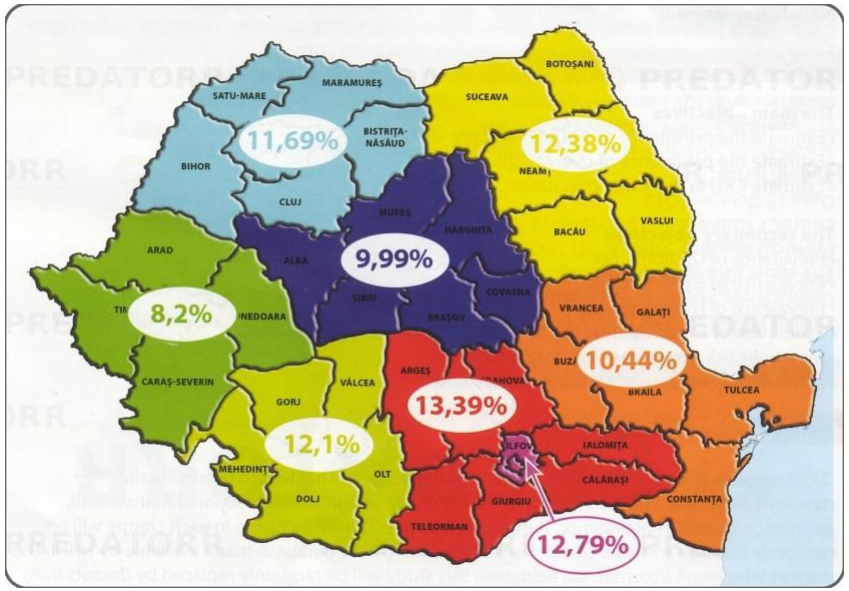
Cum realizăm studiul?

măsurăm glicemia după minim 8 ore de pauză alimentară pentru:

- 1000 - 2000 de persoane - **da, se poate**
  - avem personal
    - o persoană e suficient,
  - erori mici
    - un singur aparat de calibrat
    - personal ușor de instruit
- toată populația țării, 10.000 – 20.000 de persoane, 100.000 – 200.000 de persoane - **nu e fezabil**
  - costuri mari,
  - timp îndelungat,
  - lipsă de personal,
  - erori mari etc.

Scenariu – studiul Predatorr

- Scop: primul studiu de măsurare a prevalenței diabetului în România
- Metoda de eșantionare:
  - stratificat,
  - clusterizat
  - selecţie aleatorie.
- 2728 de participanți
- Prevalența diabetului a fost de 11,6%



Studiul Naţional privind Prevalenţa Diabetului, Prediabetului, Supraponderii, Obezităţii, Dislipidemiei, Hiperuricemiei şi Bolii Cronice de Rinichi (PREDATORR)
Mota M, Popa SG, Mota E, Mitrea A, Catrinoiu D, Cheta DM et al. Prevalence of diabetes mellitus and prediabetes in the adult Romanian population: PREDATORR study. Journal of Diabetes 2015; doi: 10.1111/1753-0407.12297

Obiectiv:
Cum influențează iluminatul nocturn calitatea somnului?

De unde recrutăm subiecții?

Numai din urban

Numai din rural

Un oraș, un sat

} Erori de eșantionare

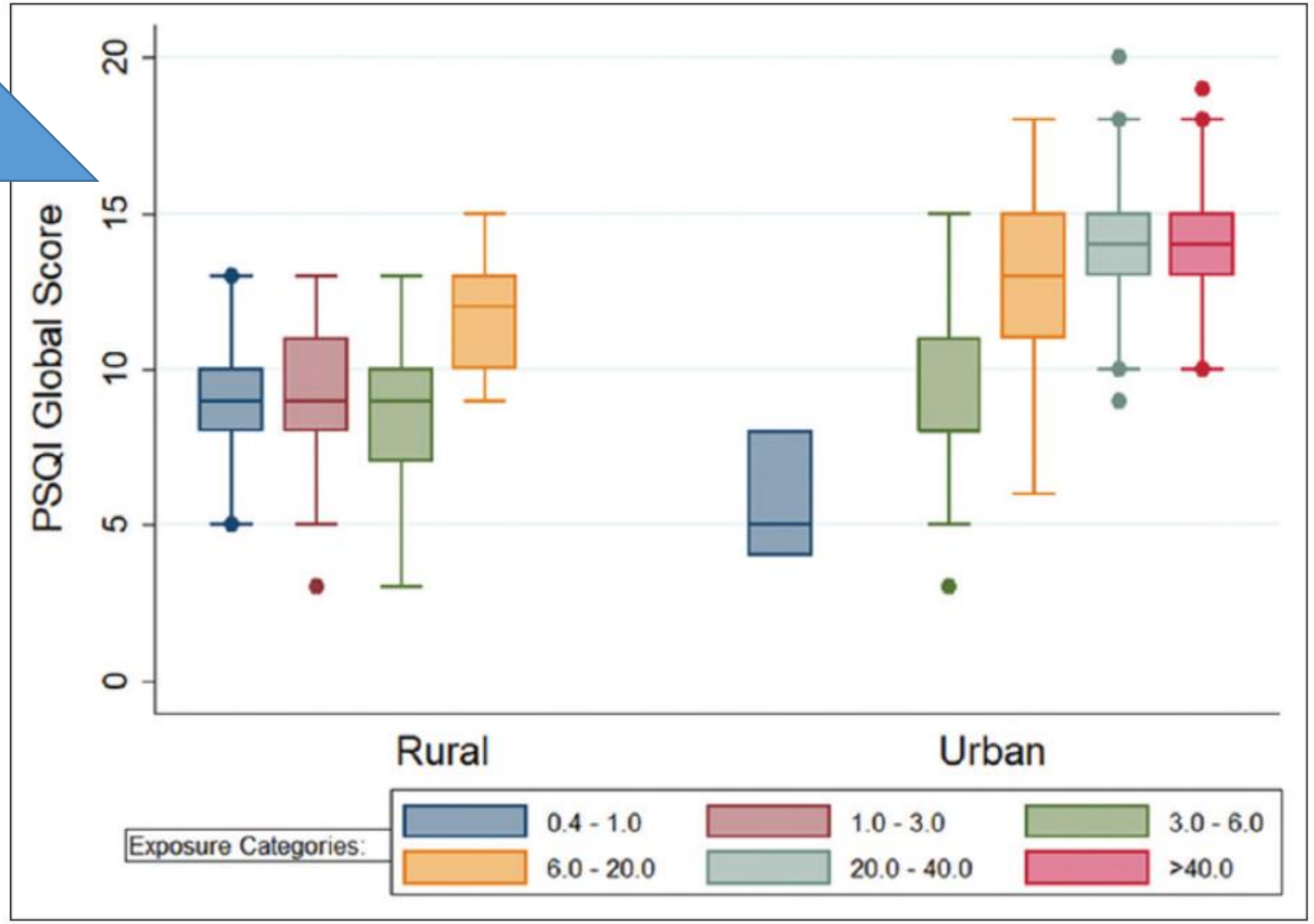
- mai sunt și alți factori care influențează calitatea somnului
  - nivelul zgomotului
- iluminat stradal diferit
  - led/tub fluorescent etc.

Din rural și din urban, din multiple orașe și din multiple comune

Ceva mai corect

Figure 1: Box and whisker plot showing comparison of Pittsburgh Sleep Quality Index global scores' distribution as per different levels of exposure among the rural and urban respondents.

Scade calitatea somnului



2 grupuri: urban / rural

Impartite in 4 categorii de expunere la lumina artificiala in timpul somnului

PSQI scor – chestionar de evaluare a calității somnului

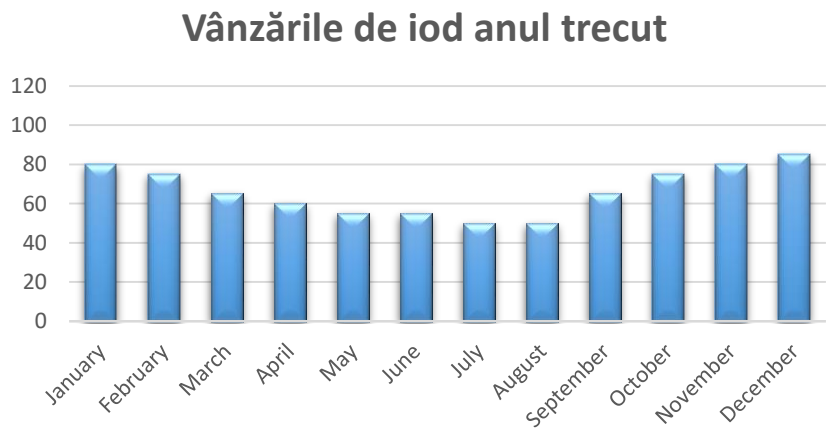
Lahiri A, Chakraborty A, Roy AK, Dasgupta U, Bhattacharyya K. Effect of light pollution on self-reported sleep quality and its components: Comparative assessment among healthy adult populations in a rural and an Urban area of West Bengal, India. Indian J Public Health 2020;64:229-35

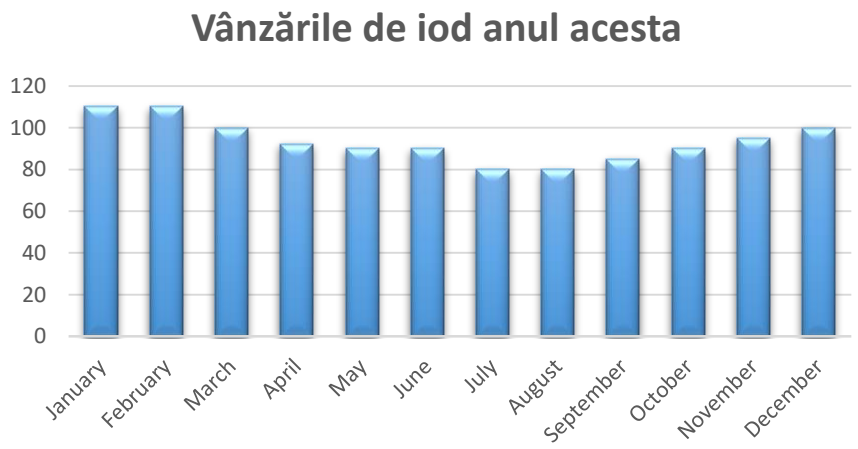
Eșantion

- o submulțime a populației

Eșantion

- ne dă o idee asupra întregii populații
  - ne dă o idee despre cum arată dispersia datelor
  - de ce e importantă dispersia?
    - compararea a două grupuri

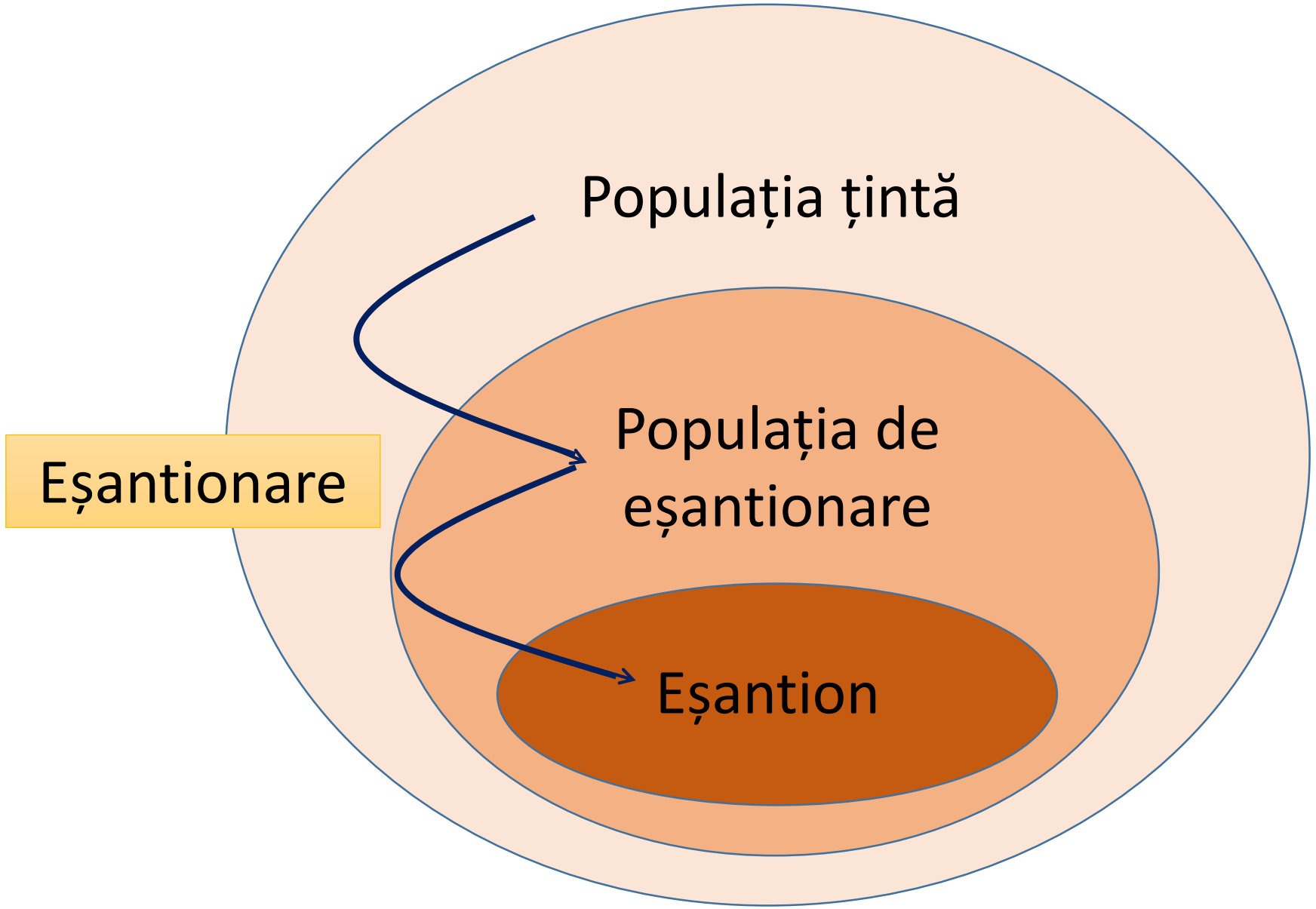




De ce să studiem eșantioane în locul întregii populații?

Se studiază eșantioane în locul întregii populații

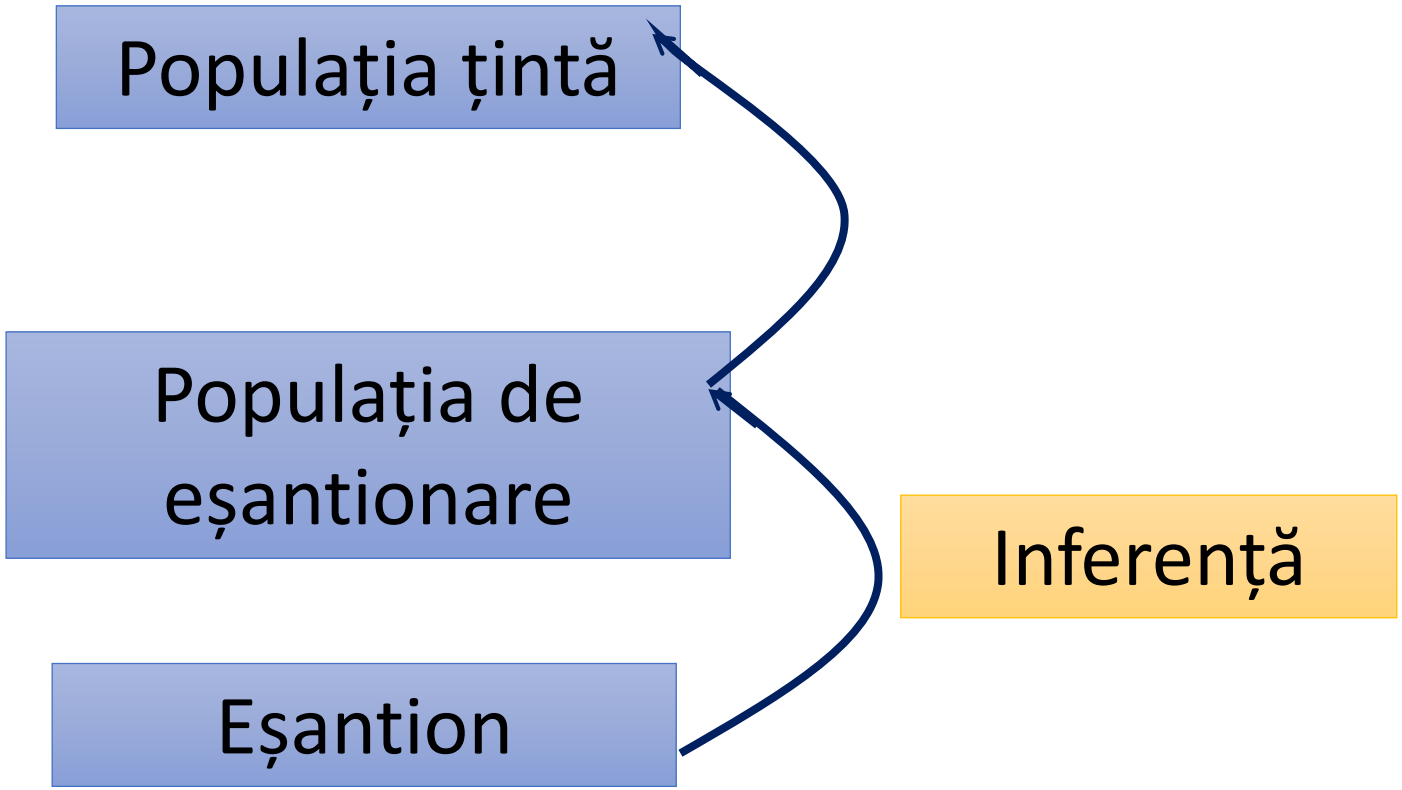
- Mai rapid
  - epidemie COVID-19, e nevoie rapidă de soluții
- Mai puțin costisitor
- Mai puțin periculos
  - întreaga populație primește un tratament nou netestat înainte
- Concluzii mai precise
  - mulți investigatori, multe măsurători – probabilitate mare de eroare



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

- Prin eșantionare obținem eșantionul

- Realizăm studiul, obținem rezultatele pe eșantion



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

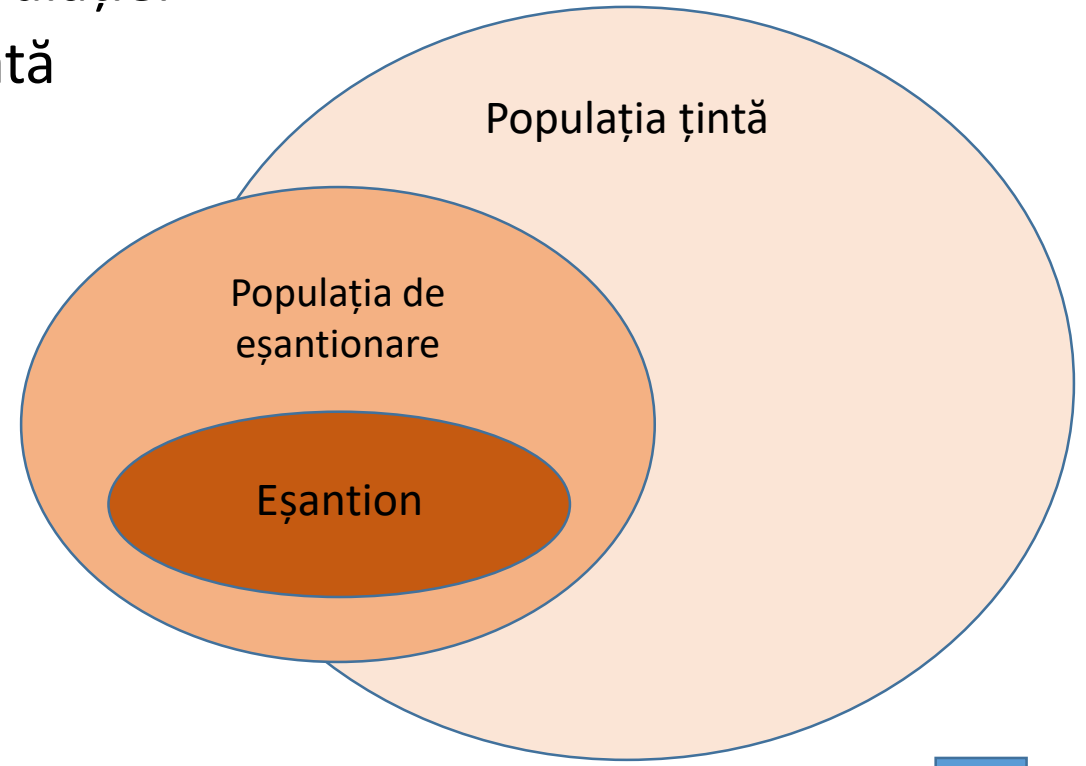
- După obținerea rezultatelor pe eșantion
  - generalizăm la întreaga populație
    - (se numește inferență statistică)

Eroare (bias) de selecție

- atunci când am greșit selecția
  - populația țintă nu este o submulțime a populației
  - o parte din populația țintă nu a fost selectată

  în mod sistematic

- Dacă selecția este biasată
  - inferența nu se poate realiza



Eroare (bias) de selecție

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

Avem nevoie de un rezultat valid

Cum facem selecţia?

Metode de eșantionare

Eşantion **reprezentativ** pentru populație

Eșantion reprezentativ

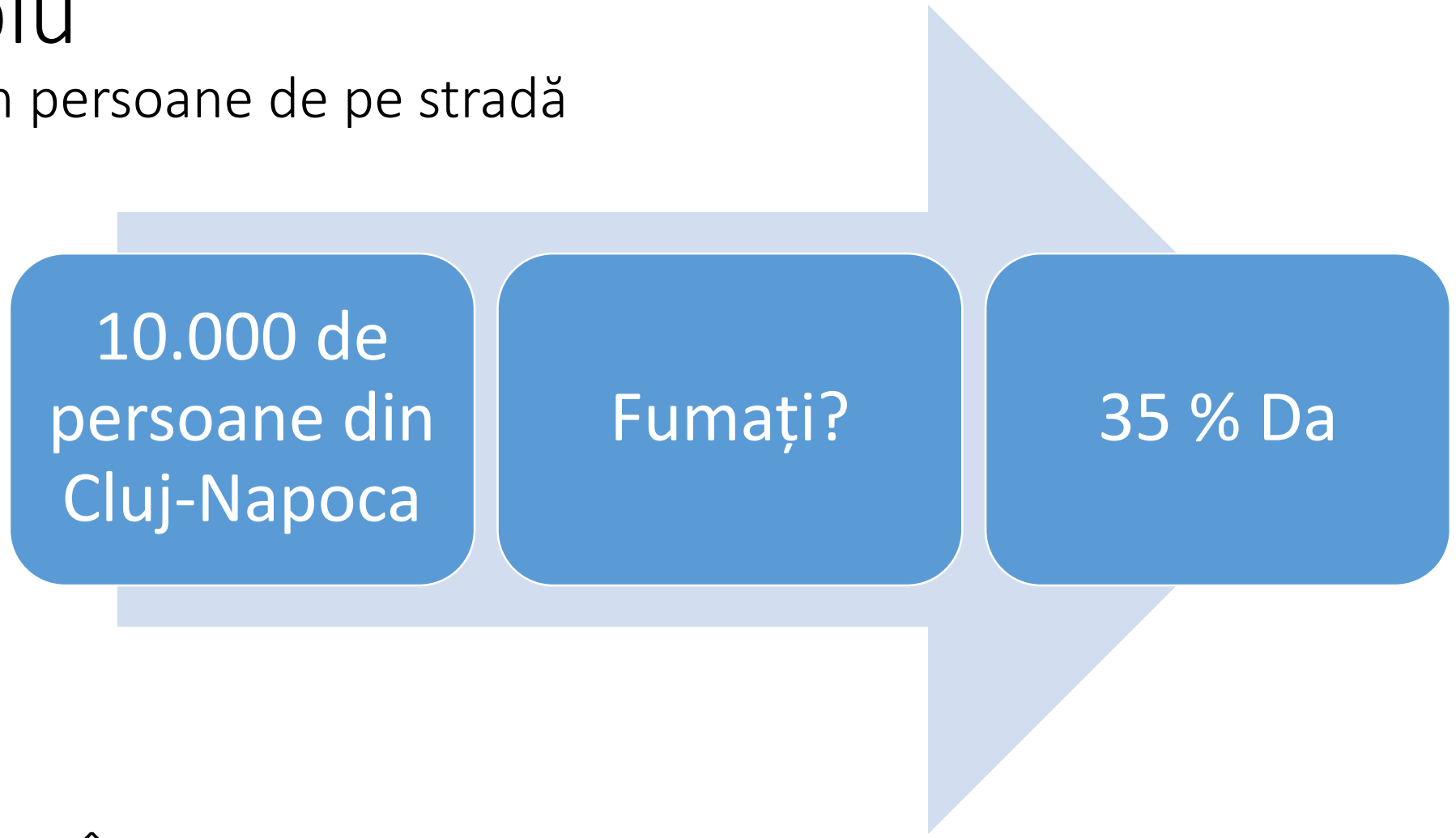
- în care sunt prezente toate subcategoriile populației cu aceeași frecvență ca și în populație
  - categoriile care reprezintă un interes pentru rezultatul urmărit

  - reprezintă interes
    - rezultat diferit al studiului pe respectivele categorii

    - ex. în ziua alegerilor dorim să realizăm un studiu care să ne arate rezultatul alegerilor
      - avem diferențe de vot între mediul rural și urban
      - prezența la vot a fost 40% din mediul urban/60% din mediul rural
      - dacă în eșantionul nostru selectat la ieșirea de la urne nu avem o prezență asemănătoare 40%/60% vom avea diferențe de rezultat.
      - nu ne interesează distribuția după gen, dacă nu există diferențe de vot între genuri

Exemplu

– selectăm persoane de pe stradă



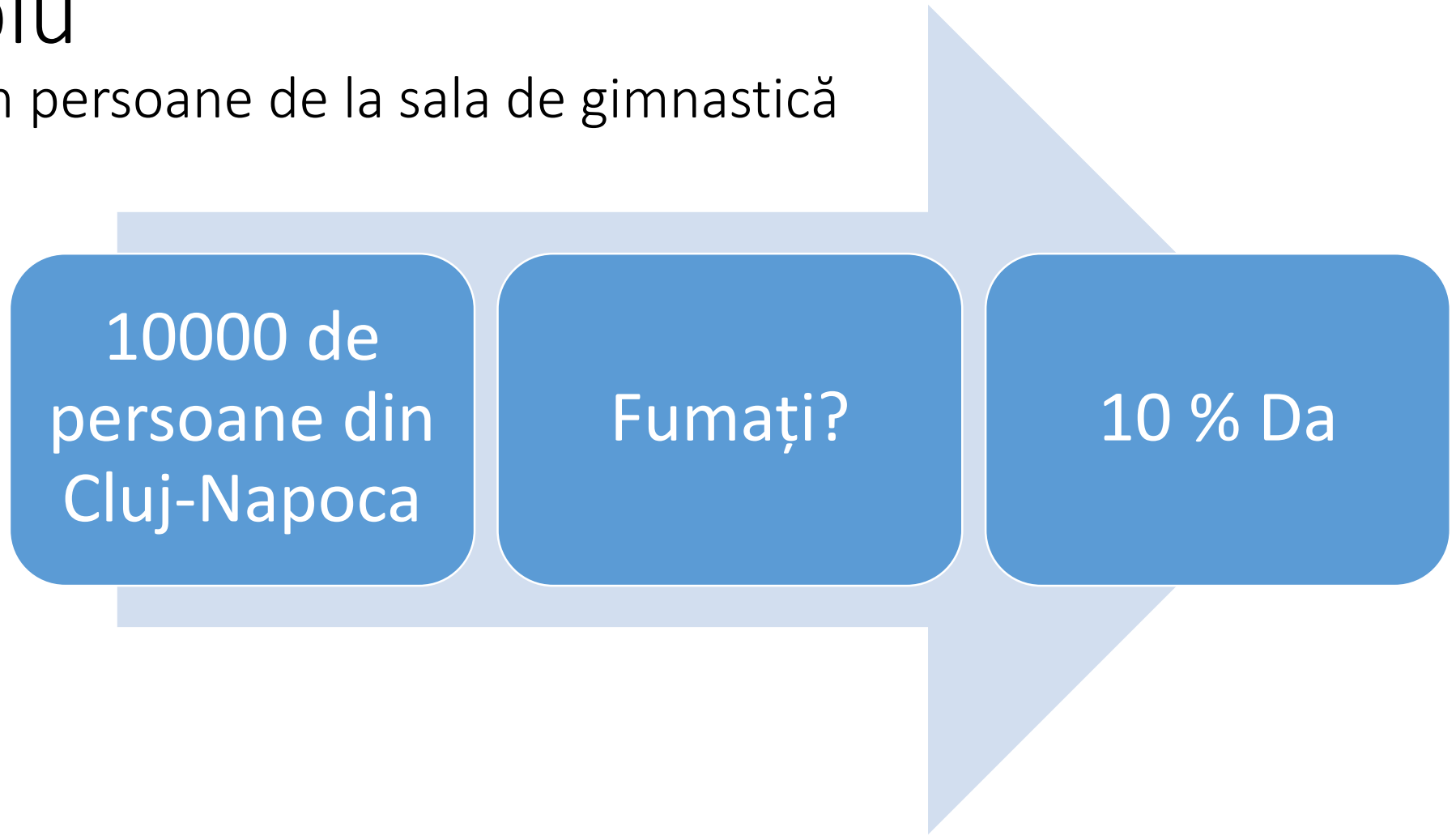
Generalizarea: În Cluj-Napoca probabilitatea ca o persoană să fumeze este 0,35

Avem 35% fumători?

Când se poate realiza generalizarea?

Exemplu

– selectăm persoane de la sala de gimnastică



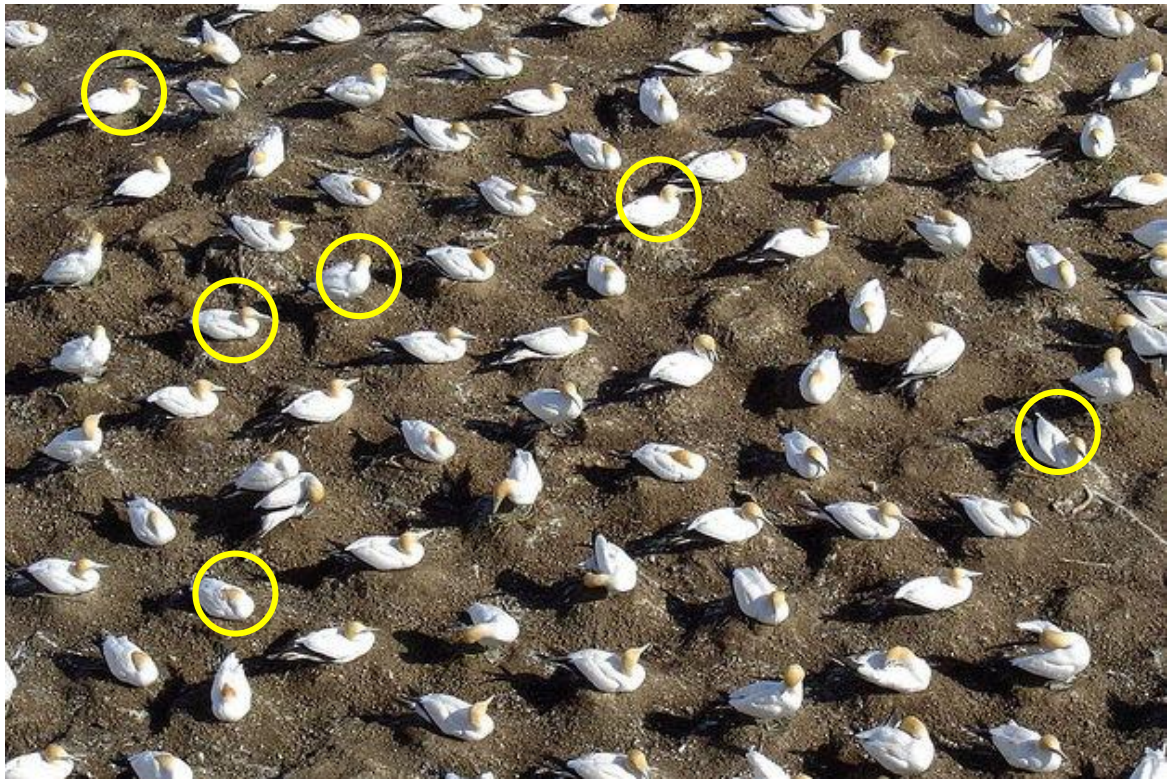
Selecția influențează rezultatul!

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

Cum putem selecta un eșantion reprezentativ?

???

prin selecție aleatoare



- o metodă de eșantionare

Selecție aleatoare

- Fiecare individ din populație are aceeași pobabilitate de a fi selectat în eșantion

- Ex. Mergeți la primărie. Luați toate CNP-urile – extrageți aleator

De ce selecție aleatoare?

- Micșorarea/eliminarea erorilor experimentale = micșorarea/eliminarea **biasului de selecție**

- la selecția aleatoare

  P (eșantion reprezentativ) > P (eșantion nereprezentativ)

probabilitatea de a selecta un eșantion reprezentativ e mai mare decât probabilitatea de a selecta un eșantion nereprezentativ

Chiar dacă selecția este aleatoare

- există o probabilitate mică să avem la selecție un eșantion nereprezentativ

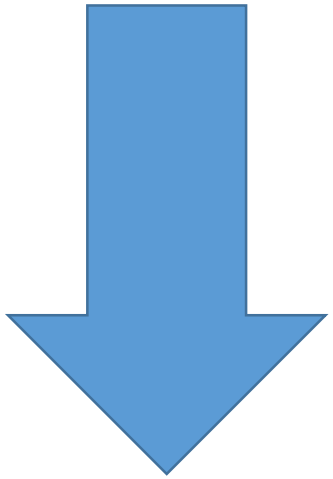
- ex. prevalența fumatului – sunt șanse mici, dar există să selectăm numai persoane care fac sport de performanță (care posibil să fumeze într-o proporție mai mică decît restul populației)

- Ce putem face?
  - studiile sunt replicate
  - dacă rezultatele sun consecvente în mai multe studii → evidențe medicale

Concluzie

Condiţia inferenței eşantion → populaţie

Selecţia aleatoare



Probabilitatea mare de a selecta un eșantion reprezentativ

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

- Preferabil să citiți
  - rezultate ale unor studii realizate pe eșantioane selectate aleator

Metode de eșantionare

Metode de eșantionare

Probabilistice

probabilitatea unui individ de a fi selectat este cunoscută

Nonprobabilistice

probabilitatea unui individ de a fi selectat este necunoscută

Metode de eșantionare

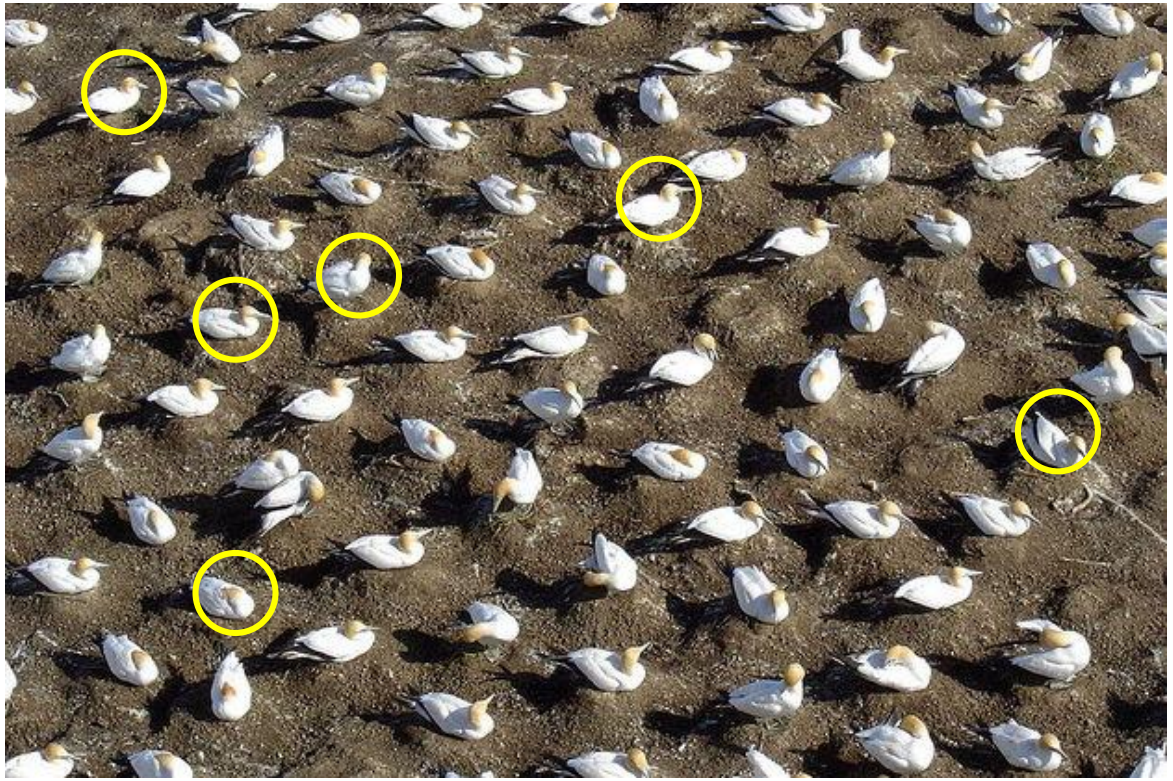
Probabilistice: fiecare subiect din populaţie are o probabilitate cunoscută de a fi selectat

- 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

Eșantionare aleatorie simplă

- fiecare subiect are o probabilitate egală de a fi selectat

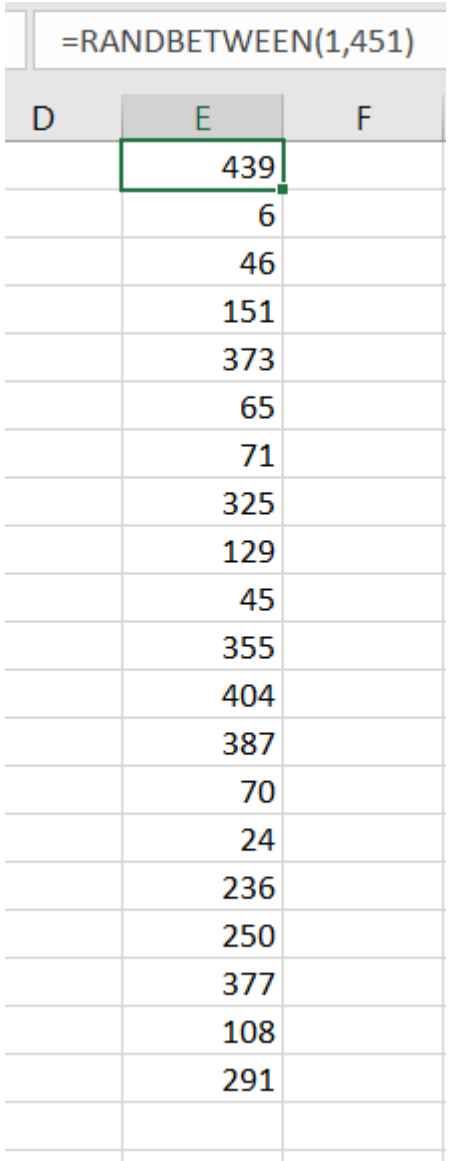
- Cum se realizează?
- Subiecților li se atribuie un număr
- Se extrag numere aleatorii din listă



Eșantionare aleatorie simplă

- Dorim sa luăm la întâmplare 20 de studenți din anul I medicină, total 450 de studenți
- Numerotăm studenții cu numere de la 1 la 451
- Folosim în Excel funcția RANDBETWEEN

- ce facem cu numerele duble?
  - eșantionare cu/fără înlocuire
- este eșantionarea oferită de calculator aleatorie?



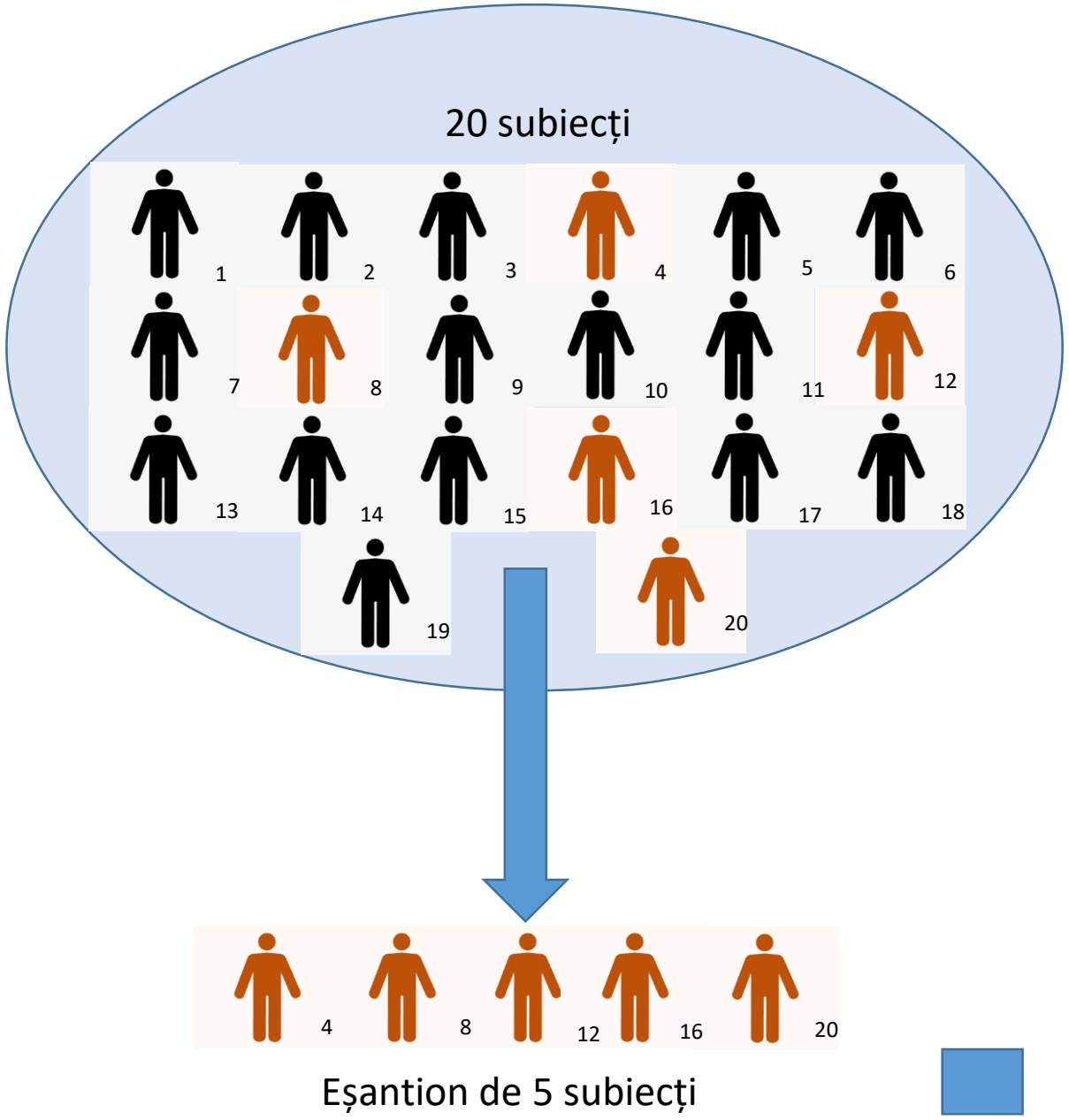
Eșantionare sistematică

- tot al k-lea individ se alege pentru a fi inclus în eșantion.
- Pentru N = 20, eșantion de 5 indivizi,

k= 20/5 = 4

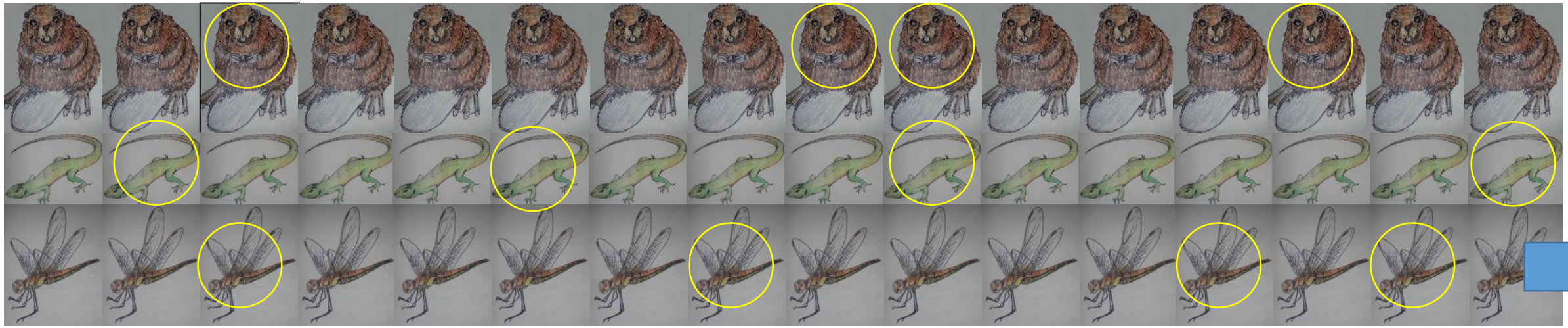
1. Se extrage aleator punctul de start
2. Tot al 4-lea individ va fi selectat

- Nu se recomandă
  - în cazul datelor ciclice
- ex. cazuri din clinica de ginecologie
  - sarcinile ectopice apar în general primăvara



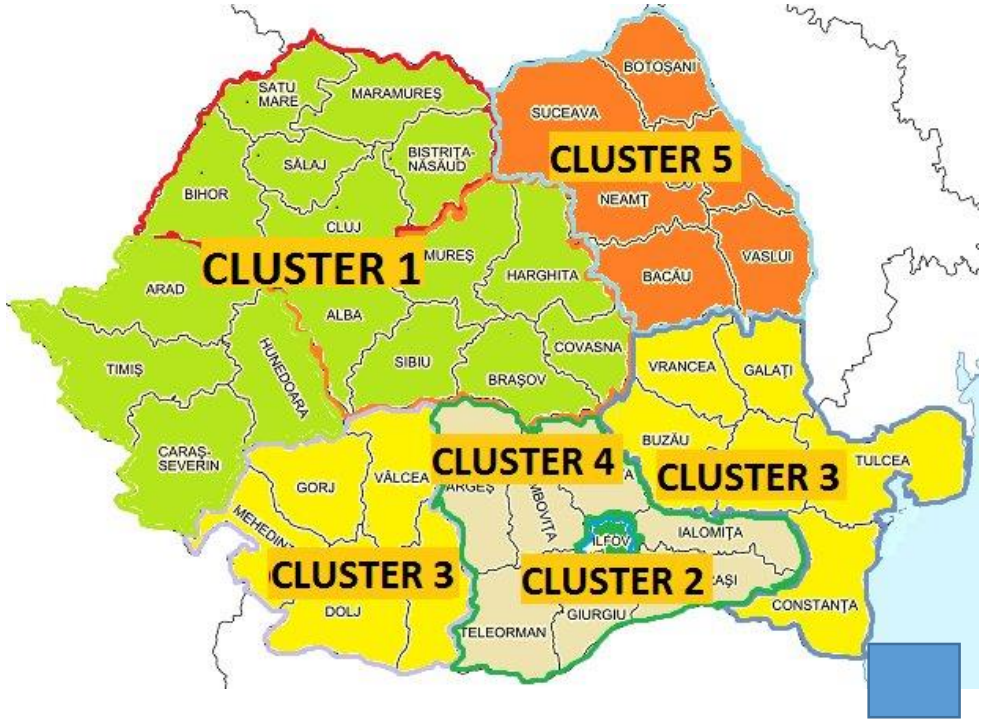
Eșantionare stratificată

- Populația este împărțită în straturi după însușiri care nu sunt echiprobabile, dar care pot influența obiectivul studiului
- ex.
  - categorii de vârstă
  - gen
- Se extrage aleator din fiecare grup câte un eșantion în funcție de cât de reprezentativ este stratul respectiv



Eșantionare de tip cluster

- Cluster= arie delimitată geografic
- Pas 1. Selectarea clusterelor
  - Pas 2. Selectare aleatorie a subiecților din fiecare cluster
- Ex. în studii multicentrice



Ulman S Particularitati regionale ale saraciei rurale in Romania

Metode de eșantionare

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

- Convenient:
    - Participanţii sunt selectaţi deoarece sunt accessibili
- Bulgărele de zăpadă:
    - Subiecţii incluşi în studiu vor aduce alţi potenţiali participanţi
    - ex.
        - membrii ai aceluiaşi grup
        - activităţi comune
- Deliberat
    - Grup de tehnici de eşantionare care au la bază gândirea cercetătorului
    - ex.
        - eșantionarea cu variație maximă,
        - eșantionarea cazurilor extreme,
        - eșantionarea realizată de experți

Eșantionarea non-probabilistică

- de multe ori - Reflectă erorile de gândire ale cercetătorului
- poate duce la rezultate
  - subiective
  - eronate
- au un posibil bias de selecție
- nu putem estima erorile de eșantionare

Recensământ

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

Variabila aleatoare, distribuția de probabilitate

Eșantioane selectate aleator

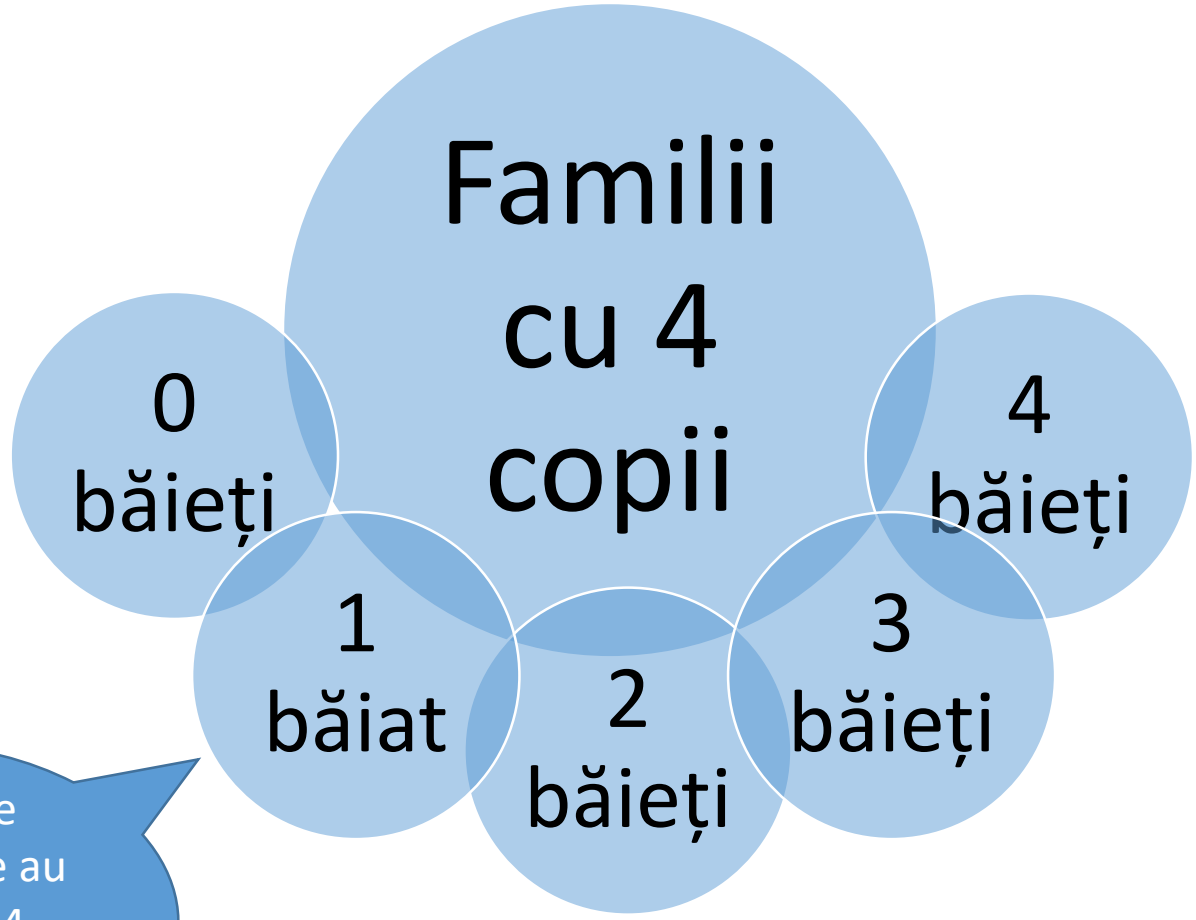
Măsurători

Rezultatul imprevizibil

Rezultatul = variabilă aleatoare

Experiment

La naștere
P(sex masculin)=0,5
(50%)



Nu sunt alte posibilități: fie au 0 băieți din 4 copii, fie au 1 băiat, etc.

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 băieți	1 băiat	2 băieți	3 băieți	4 băieți	**Total**
Nr. de familii	4	29	40	24	9	100

- Numarul de băieți într-o familie cu 4 băieți – **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

Distribuția de probabilitate obținută empiric

Numim **distribuție de probabilitate a variabilei aleatoare X** numărul de apariții a valorilor posibile ale 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 cu 4 copii – **Variabila aleatoare**

Distribuție de probabilitate

Frecvențele de apariție a valorilor unei variabile aleatoare

→sumarizate într-o **distribuție de frecvențe** =

= distribuție de probabilitate

Cum calculăm distribuția de probabilitate?

- Empiric
  - eșantion, apoi inferență
- Teoretic
  - Formulă
  - Regulă
  - Aproximare cu o distribuție de probabilitate teoretică cunoscută

Ex. **Distribuția de probabilitate teoretică** a numărului de baieți 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 (probabilitatea teoretică),
  - un comportament “normal” = neinfluențat de diverși factori

La naștere
P(sex masculin)=0,5
(50%)

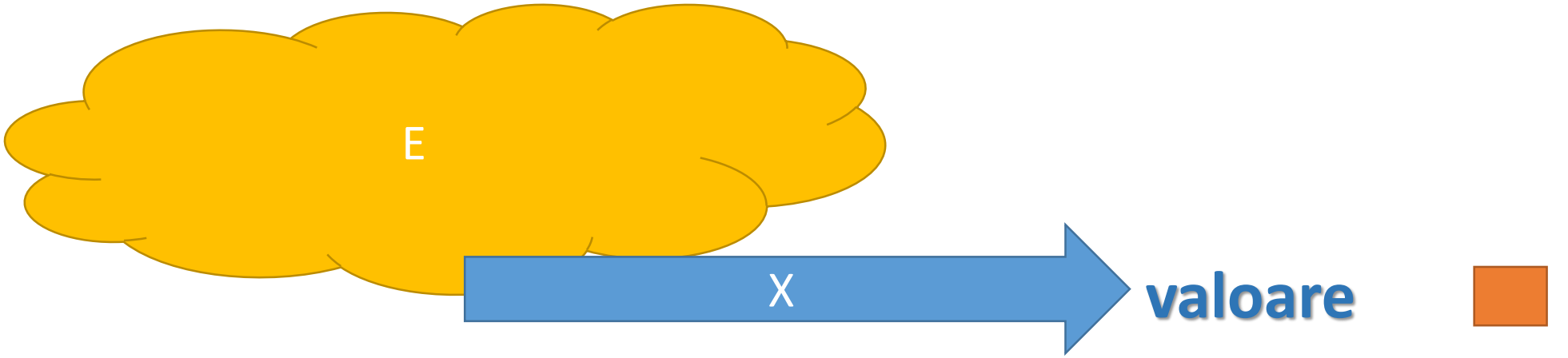
Dacă ne interesează același obiectiv la clinica de infertilitate

- Familii cu gemeni:

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

Definiţie

- **Se numeşte <u>variabilă aleatoare</u> pe un spaţiu fundamental E şi se notează prin X, o funcţie definită pe E cu valori în mulţimea numerelor reale.**
- Unei variabile aleatoare X i se pot asocia diferite probabilităţi pentru fiecare valoare posibilă a sa, ca de exemplu:
  - Pr(X = a) - probabilitatea ca “X să ia valoarea a”;
  - Pr($a \leq X \leq b$) - probabilitatea ca “X să ia o valoare în intervalul [a,b]“.



Definiţie

- O variabilă aleatoare se numește discretă dacă ea poate lua un număr finit sau cel mult numărabil de valori
- O variabilă aleatoare este *continuă* atunci când variază în mod continuu într-un interval şi poate lua o mulţime nenumărabilă de valori.

Exemple

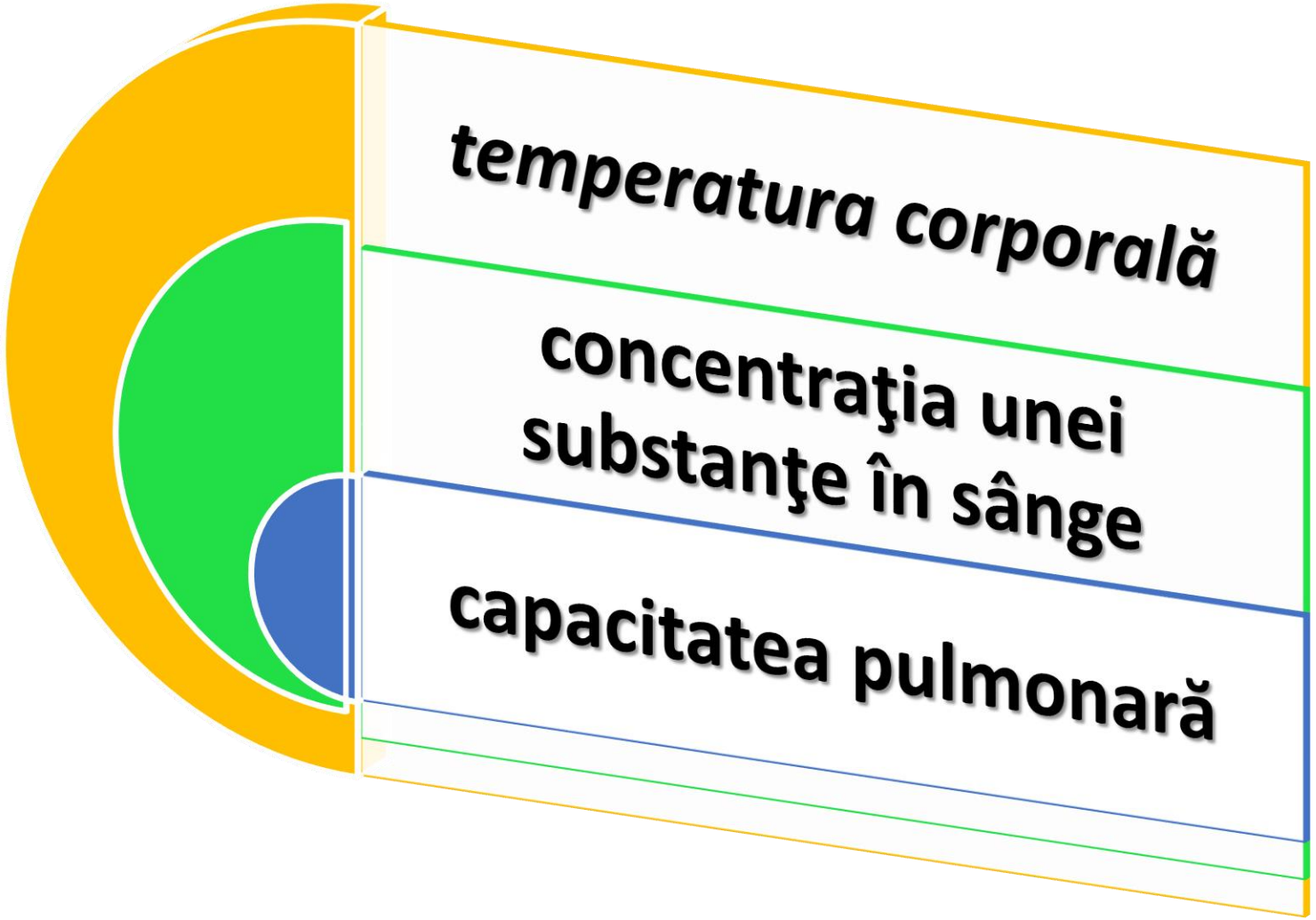
variabilă aleatoare discretă infinită

- Numărul de internări într-un spital într-un interval de timp dat $X=\{0,1,2,...,n,...\}$.
- Numărul de bacterii într-un mililitru de apă $X=\{0,1,2,...,n,...\}$.
- Numărul de prezentari la medic pentru otita in primii doi ani de viata $X=\{0,1,2,...,n,...\}$.

variabilă aleatoare discretă finită

- Numărul de indivizi cu RH-negativ dintr-un grup de n persoane luate la întâmplare $X=\{0,1,2,...,n\}$.

Exemple: variabile aleatoare continue



LEGEA DE PROBABILITATE A UNEI VARIABILE ALEATOARE FINITE

- Fie X o variabilă aleatoare pe un spaţiu fundamental E finit, adică

$$X = \{x_1, x_2, \ldots, x_i, \ldots, x_n\}.$$

- Mulţimea de probabilităţi

$P(x_1), P(x_2), \ldots, P(x_i), \ldots, P(x_n)$ asociate valorilor: $x_1, x_2, \ldots, x_i, \ldots, x_n$

- se numeşte distribuţia sau legea de probabilitate a variabilei aleatoare X.

- Distribuţia de probabilitate in cazul unei variabile aleatoare finite X:

$$X: \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ P(x_1) & P(x_2) & \dots & P(x_n) \end{pmatrix}$$

- Proprietate

$P(x_1) + P(x_2) + \dots + P(x_n) = 1$

Conditie obligatorie

Exemplu

- Apariţia uneia dintre feţele unui zar:

$$X_1 : \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix}.$$

- probabilitatea P(x) - constantă - distribuţia lui X1 este **uniformă**.

$$Pr(x_1) + Pr(x_2) + \ldots + Pr(x_n) = 1/6 + 1/6 + 1/6 + 1/6 + 1/6 + 1/6 = 6/6 = 1$$

Valoarea aşteptată sau **speranţa matematică**

$$\bar{X} = \sum_{i=1}^{n} x_i * p(x_i)$$

- **Observaţii:**
  - Dacă legea de probabilitate a lui X este uniformă, adică $P(x_i) = 1/n$, pentru orice $i = 1,2,...,n$, atunci $\bar{X}$ este media aritmetică a numerelor x_1 , x_2, ..., x_i, ..., x_n.

Exemplu

- Numarul de episoade de otită în primii doi ani de viață:

r	0	1	2	3	4	5	6
$Pr(X = r)$	.129	.264	.271	.185	.095	.039	.017

- $\bar{X}$= 0 x 0,129 + 1 x 0,264 + 2 x 0,271 + 3 x 0,185 + 4 x 0,095 + 5 x 0,039 + 6 x 0,017
- $\bar{X}$=2,038

Ne vom aștepta la aproximativ 2 episoade de otită pentru un copil în primii doi ani de viață

Variaţia şi abaterea standard

- Variaţia variabilei aleatoare X se defineşte prin

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

- Prin definiţie abaterea standard este:

$$S = \sqrt{S^2}$$

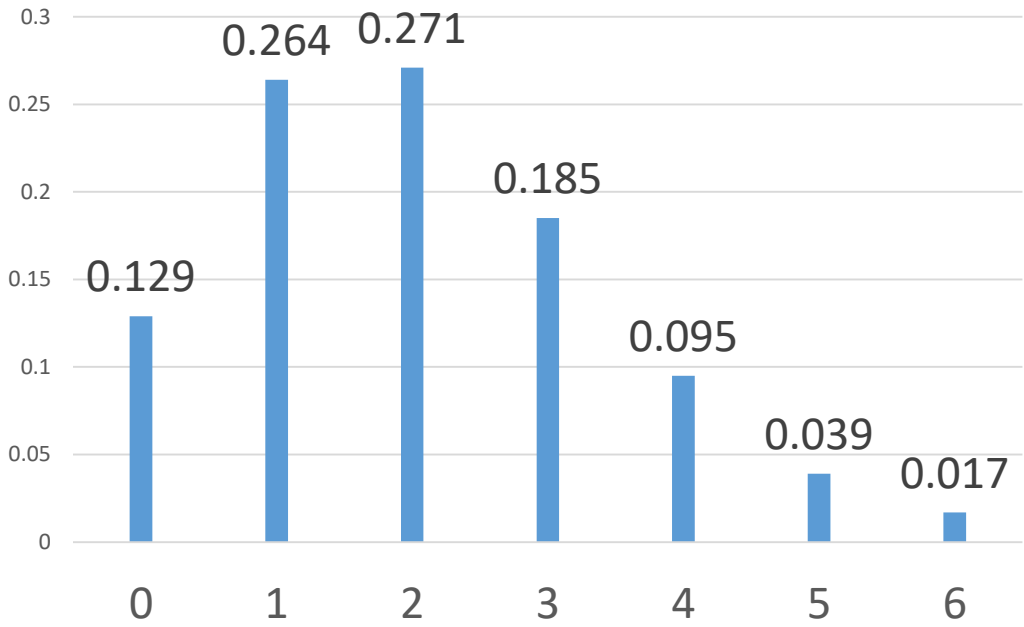
Exemplu

- Numărul de episoade de otită în primii doi ani de viață:

r	0	1	2	3	4	5	6
$Pr(X = r)$	.129	.264	.271	.185	.095	.039	.017

- $S^2 = (0-2,138)^2 \times 0,129 + (1-2,138)^2 \times 0,264 + (2-2,138)^2 \times 0,271 + (3-2,138)^2 \times 0,185 + (4-2,138)^2 \times 0,095 + (5-2,138)^2 \times 0,039 + (6-2,138)^2 \times 0,017$
- $S^2 = 1,967$
- $S = 1,402$

In acest caz



- Numarul de episoade de otita in primii doi ani de viata:

- $\bar{X}$=2,038
- S =1,402

r	0	1	2	3	4	5	6
$Pr(X = r)$	.129	.264	.271	.185	.095	.039	.017

În interval medie +/- 2 S = 2,038 ± 2,8 = [-0,766; 4,842] sunt cuprinse

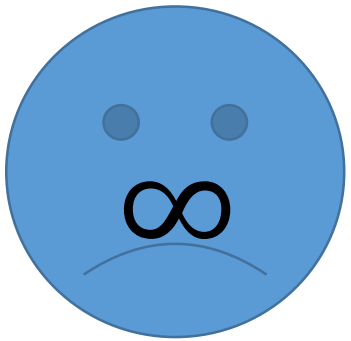
0,129 + 0,264 + 0,271 + 0.185 + 0,095 = 0,944 => 94,4%

Cazul discret finit – Cazul continuu

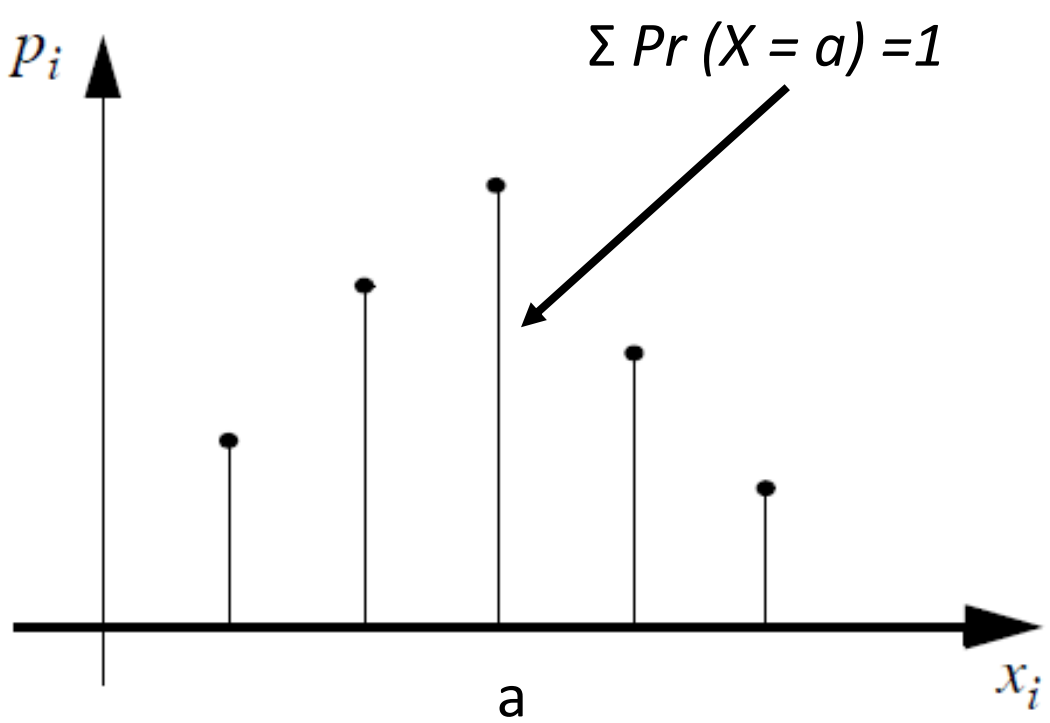
- Analog, dar ∞

$$\sum_{i=1}^{n}$$

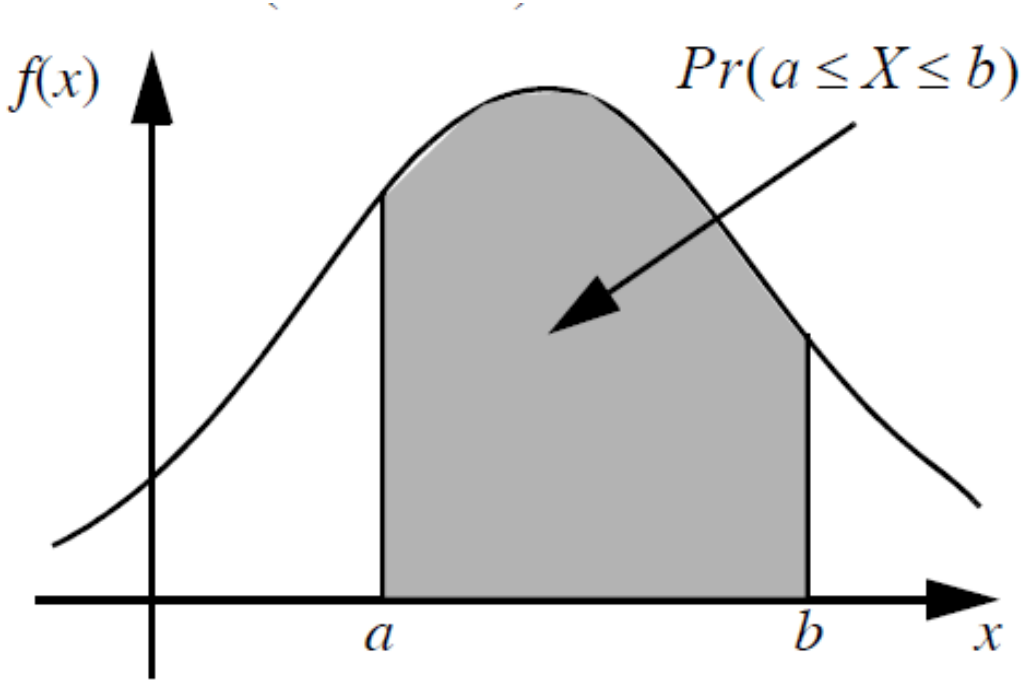
$$\sum_{i=1}^{\infty}$$



Cazul discret – Cazul continuu



Aria de sub curbă =1



! Avem noroc – în secolul nostru calculează calculatorul

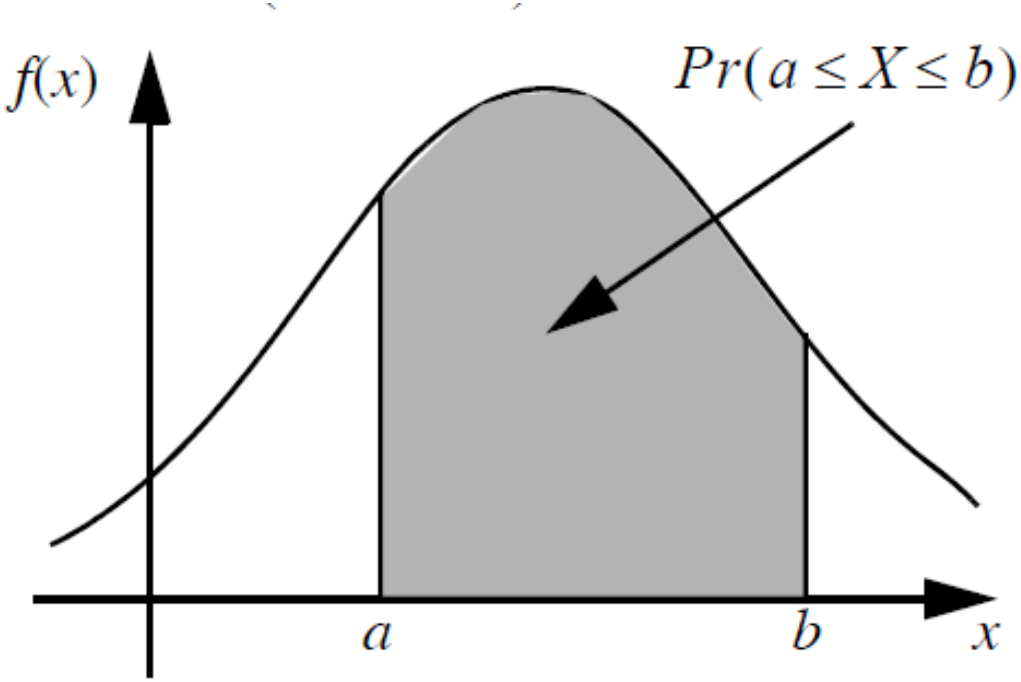
Cazul continuu

- In cazul unei variabile aleatoare continue X, se consideră o funcţie f:R→R numită *densitate de probabilitate*, care are proprietăţile:

$$\Pr(a \le X \le b) = \int_a^b f(x)dx$$

$f(x) \ge 0, \forall x \in R$

$$\int_{-\infty}^{\infty} f(x)dx = 1$$



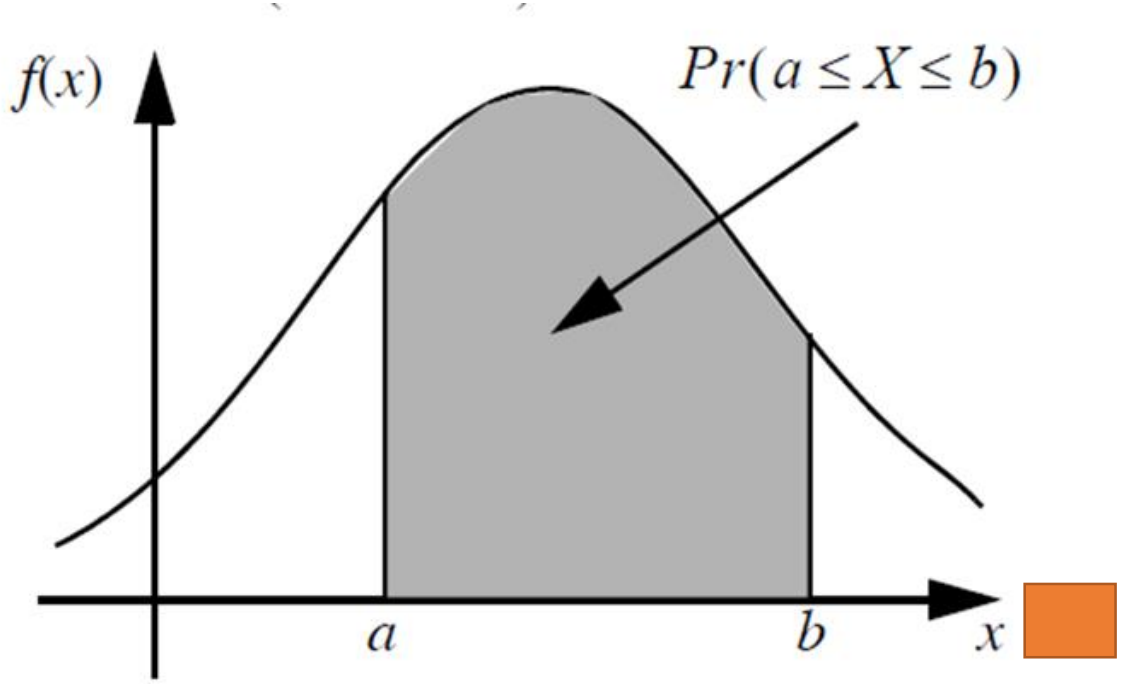
Cazul continuu

- In acest caz funcţia de repartiţie F asociată variabilei aleatoare X este definită prin:

$$F(x) = Pr(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

- De asemenea, media lui X este definită prin :

$$M(X) = \int_{-\infty}^{\infty} xf(x)dx$$



- iar variaţia lui X

$$V(X) = \int_{-\infty}^{\infty} [x\text{-}M(X)]^2 f(x)dx$$

Cum aflăm distribuția de probabilitate?

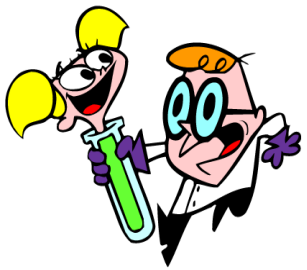
Dacă variabila e finită

- Empiric – experiment – distribuție de frecvențe

Dacă e ∞?

! suntem norocoși - găsim

- Formulă
- Regulă



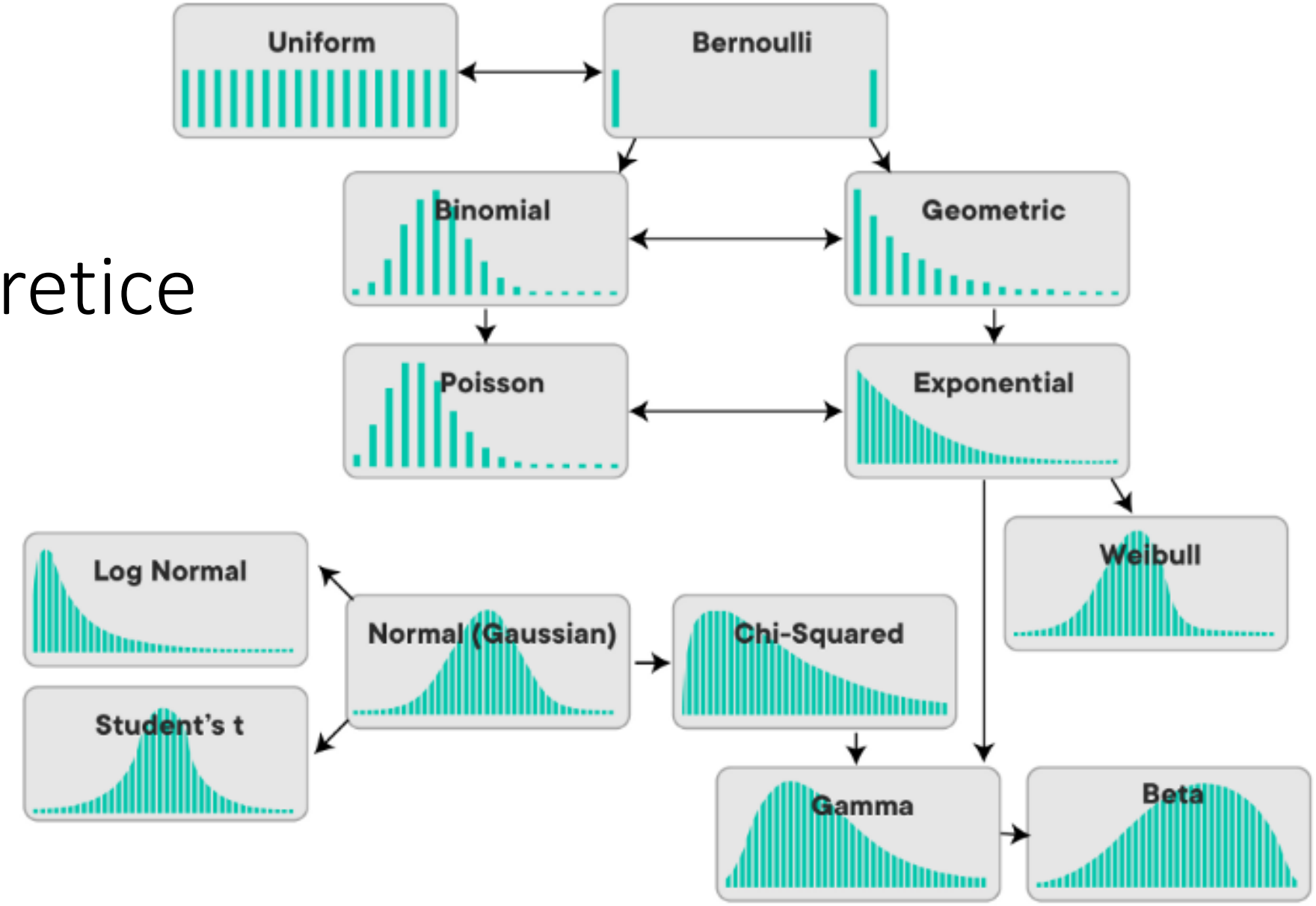
Dacă nu suntem norocoși:

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

Legi de distribuţie

(distribuţii de probabilitate)

Distribuţii teoretice



- Au o funcție cunoscută, medie și deviaţie standard deductibilă

Akshay S. Understanding Different Types of Distributions You Will Encounter As A Data Scientist

Principalele legi de distribuţie

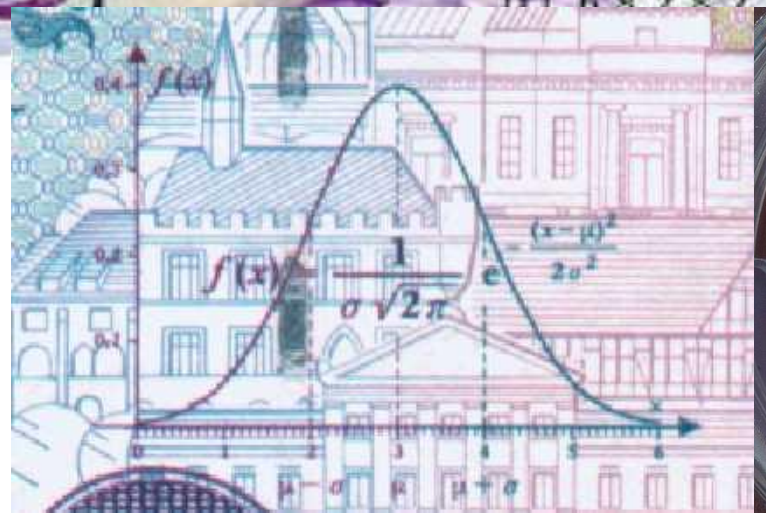
legea BINOMIALĂ (BERNOULLI)	• probabilitatea unei variabile de tip succes/eșec. De câte ori apare evenimentul intr-un număr dat de trialuri
legea POISSON	• probabilitatea evenimentelor rare
legea normală sau legea LAPLACE-GAUSS	• probabilitatea unui eveniment în cazul variabilelor continue
legea STUDENT (t)	• probabilitatea unui eveniment în cazul variabilelor continue
legea $\chi 2$ a lui PEARSON	• sume de pătrate a unor variabile independente normal distribuite
legea F a lui FISHER.	• comportarea câtului a două variabile cu distribuţie Hi-pătrat

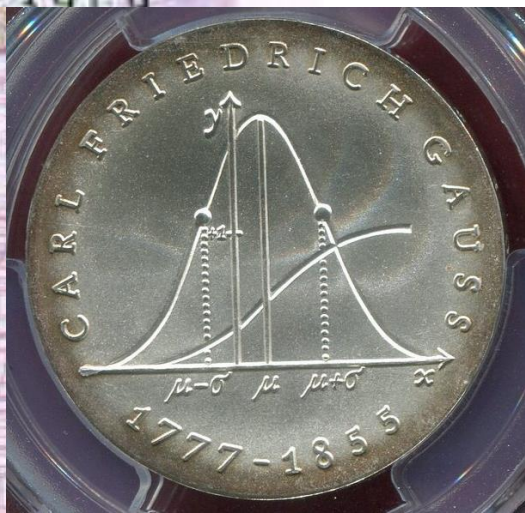
Simboluri utilizate frecvent în inferența statistică

	Simbolul pentru parametru calculat pe populaţie	Simbolul pentru statistica calculată pe eşantion
Media	μ	$\bar{X}$
Deviația standard	σ	S
Proporția	π	p

LEGEA NORMALĂ - Karl Friedrich Gauss

1777–1855





LEGEA NORMALĂ

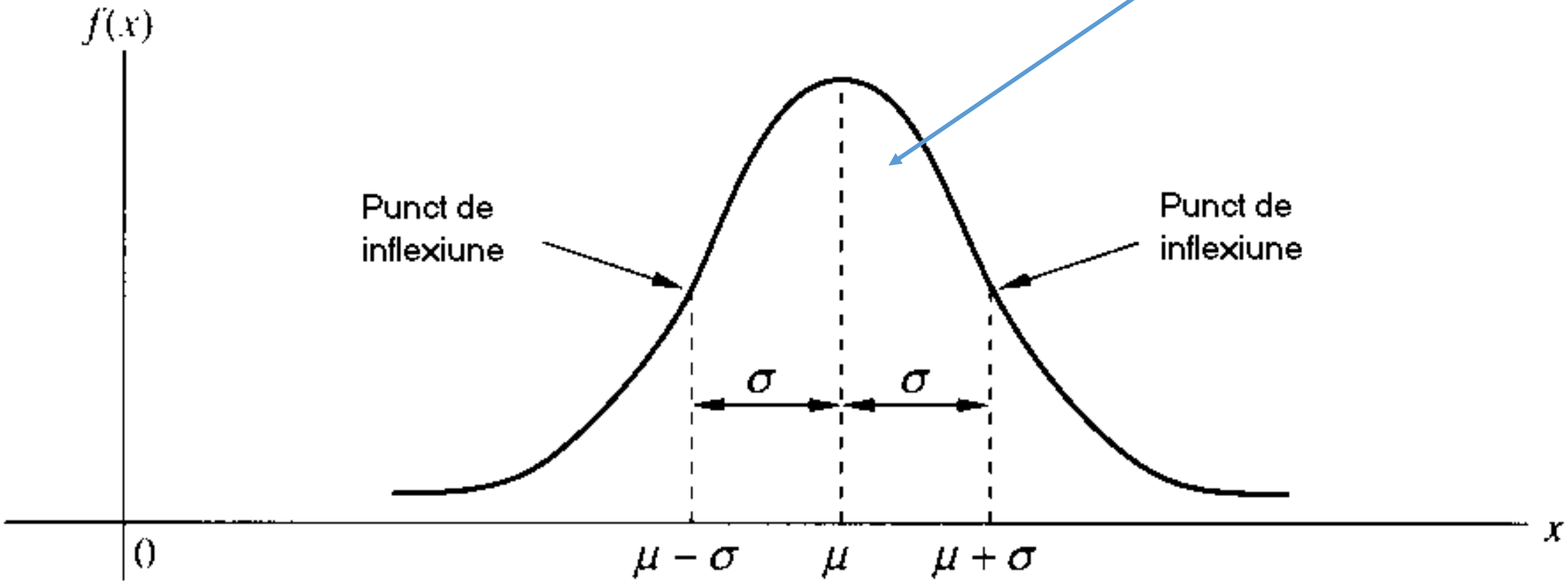
1777–1855

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

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

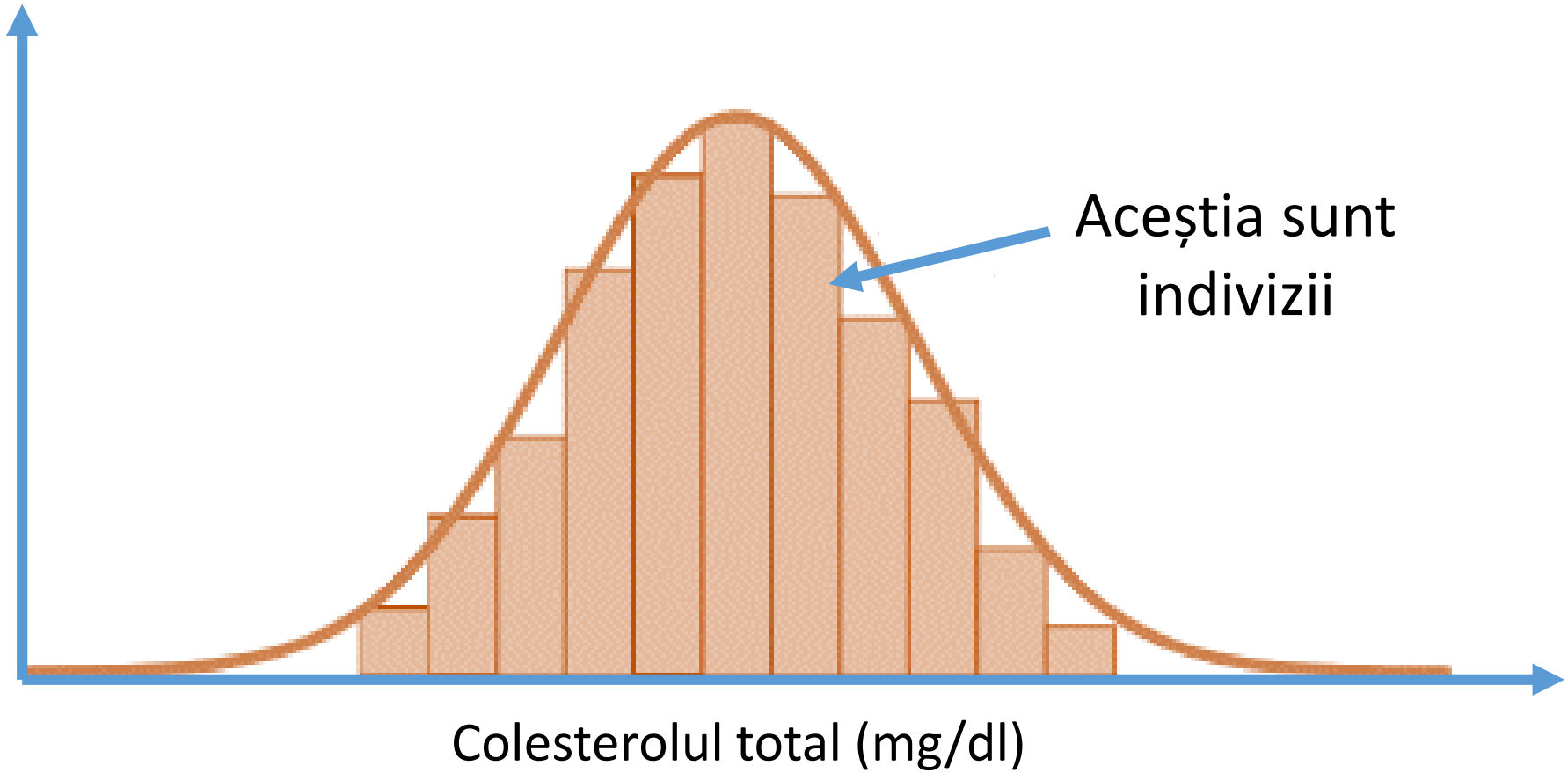
LEGEA NORMALĂ

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)

n=100



Transformarea Z

- când media aritmetică a unei distribuții ≠ 0 și varianța ≠ 1

Pasul 1. Mutăm distribuția în sus sau în jos pe linia numerică, astfel încât media să fie 0, adică $X - \mu$

Pasul 2. Ajustăm distribuția fie mai largă, fie mai îngustă /σ

$$Z = \frac{X - \mu}{\sigma}$$

numit și scor z, o abatere normală

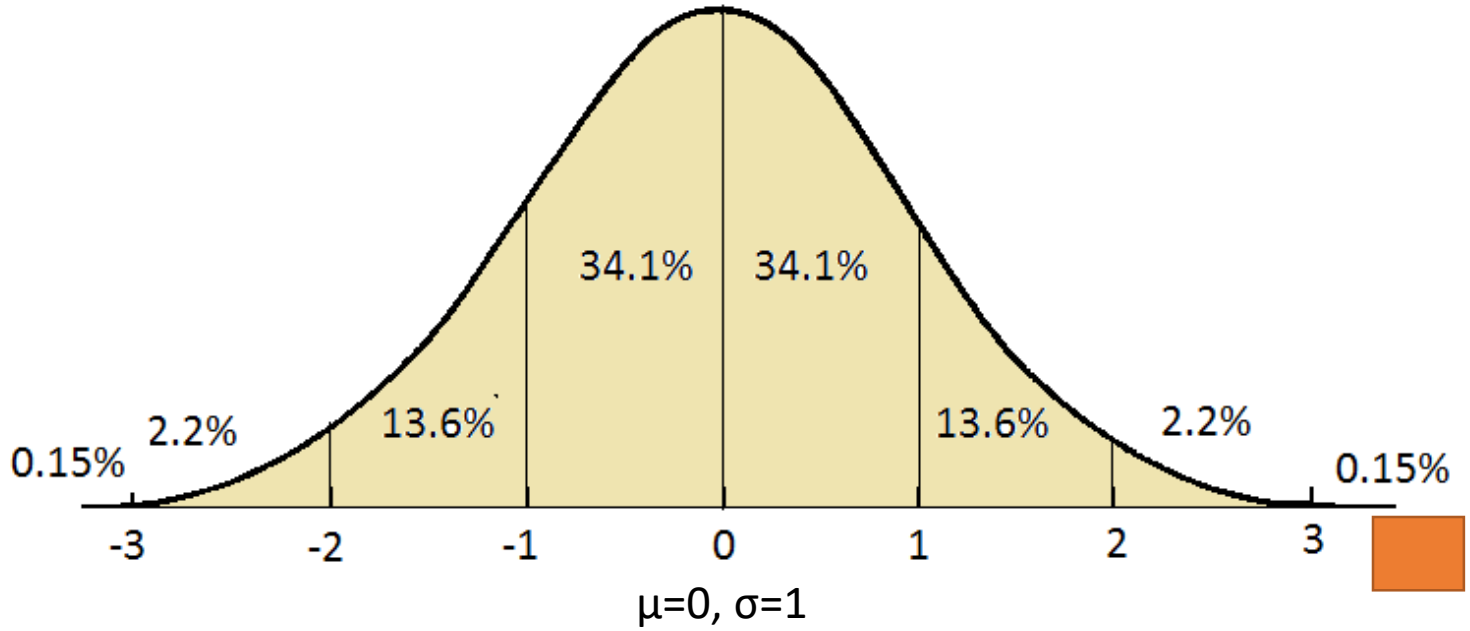
Distribuția normală standardizată

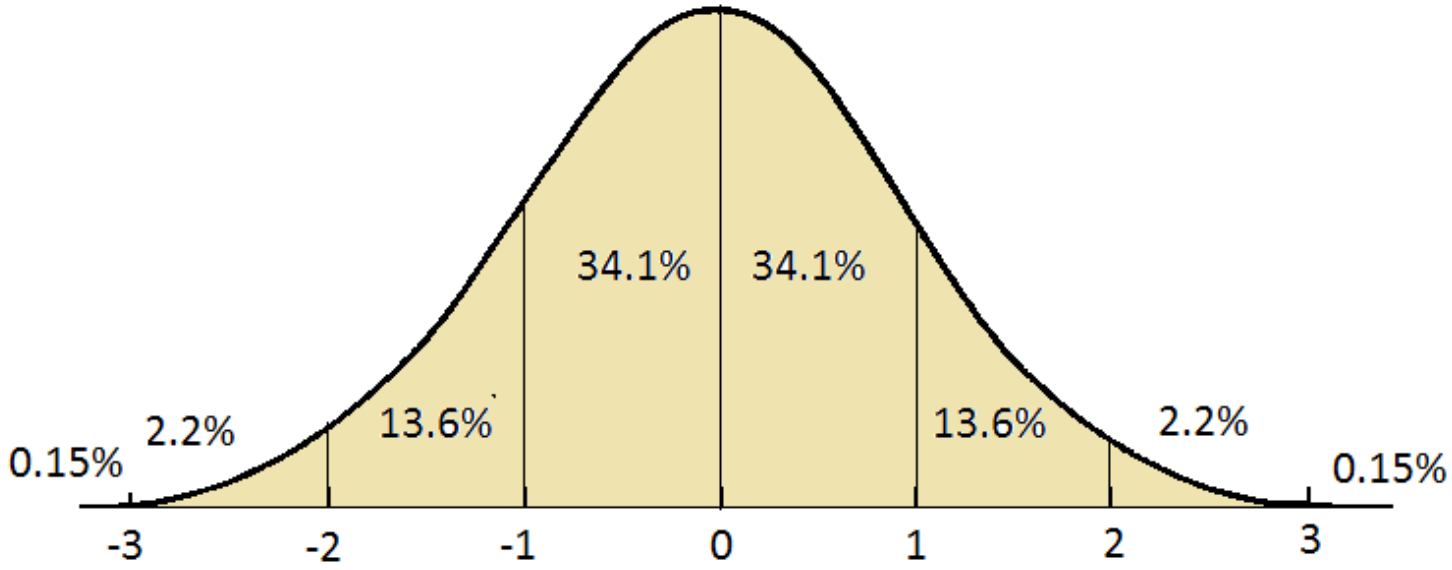
O distribuție normală cu μ=0 si σ=1.

$$Z = \frac{X - \mu}{\sigma}$$

Formula distribuției z:

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





- Câți subiecți sunt peste 0?
  - 50%
- Câți subiecți sunt între 0 și 1?
  - 34,1%
- Câți subiecți sunt peste 2 (două deviații standard)?
  - 2,35%

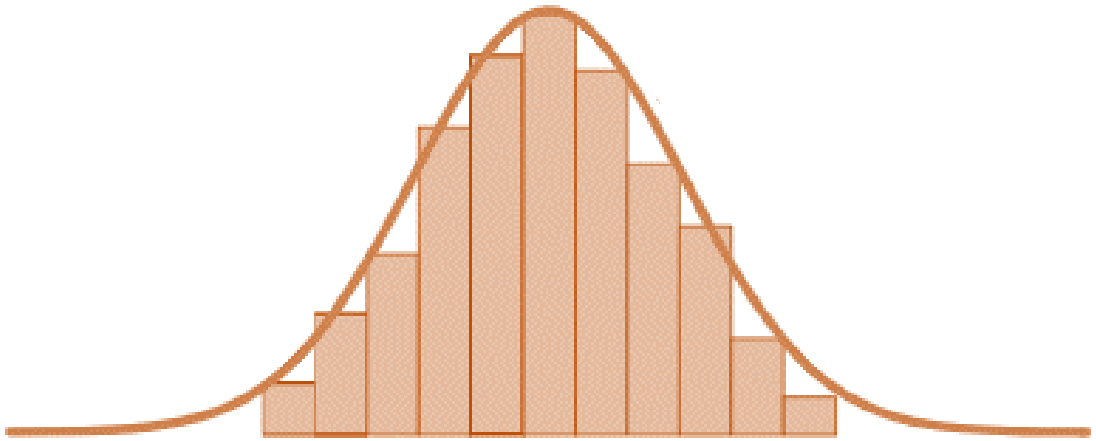
Proprietăți:

- punct de maxim este media aritmetică

  media aritmetică = modul

- simetrică față de media aritmetică

  media aritmetică = mediana

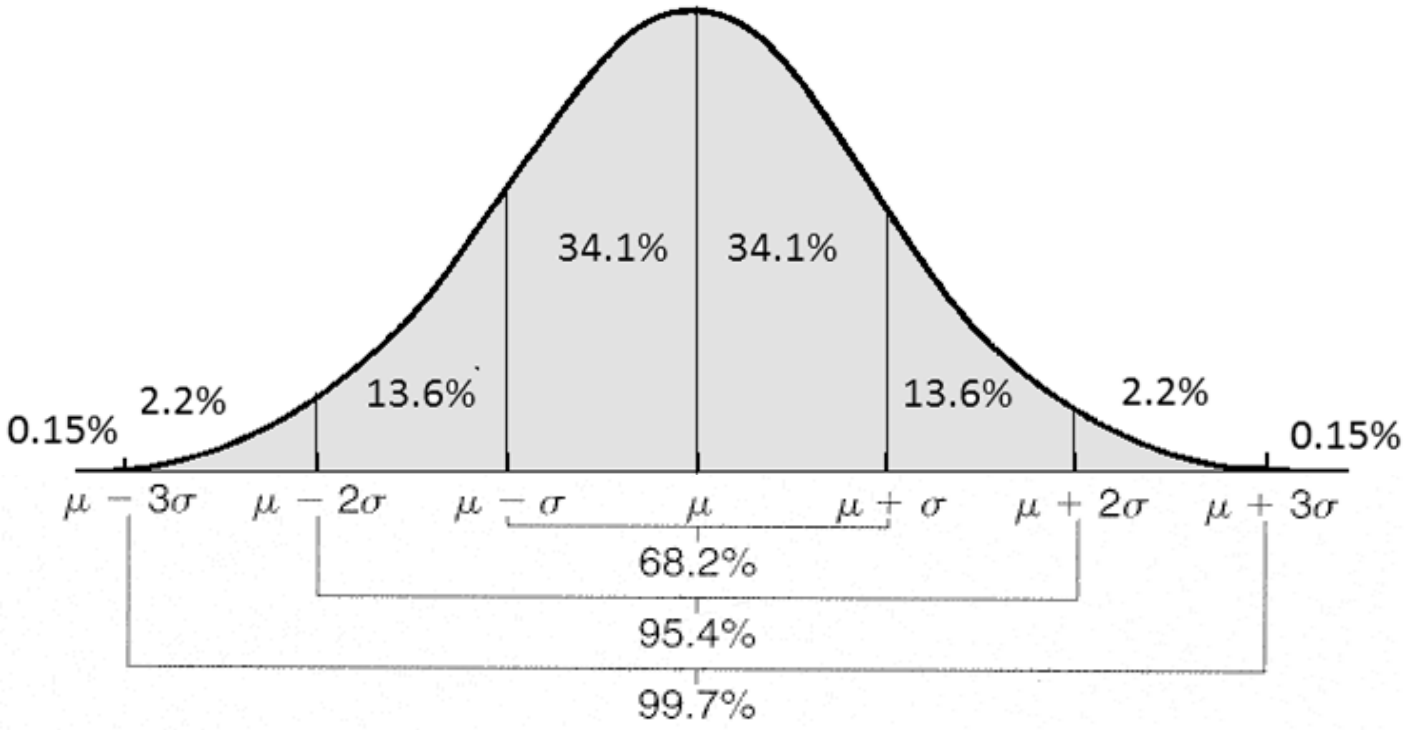


Proprietăți:

În intervalul medie ± abatere standard - minim 68,2% din observaţii;

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

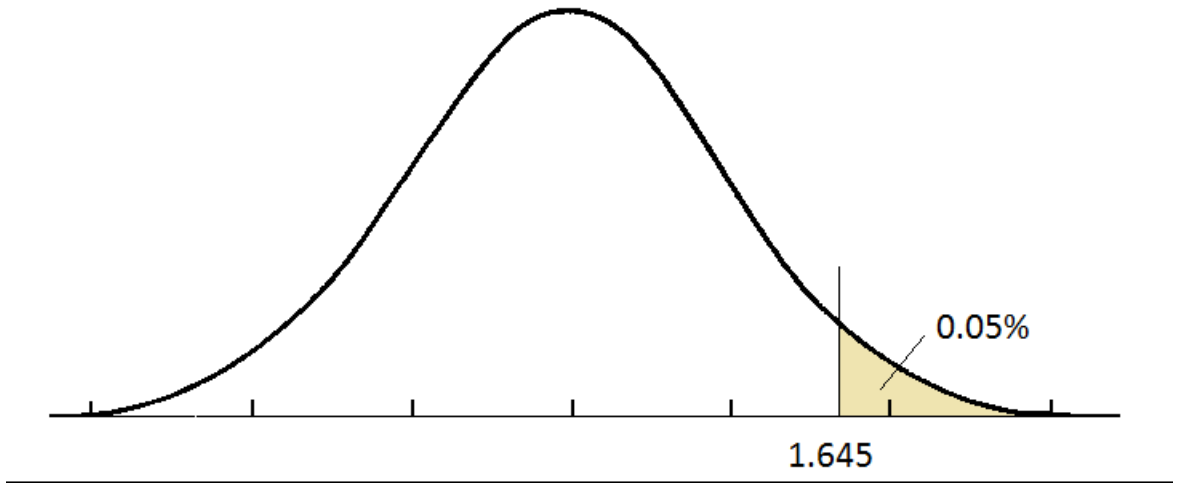
În intervalul medie ± 3 * abatere standard - minim 99,7% din observaţii.

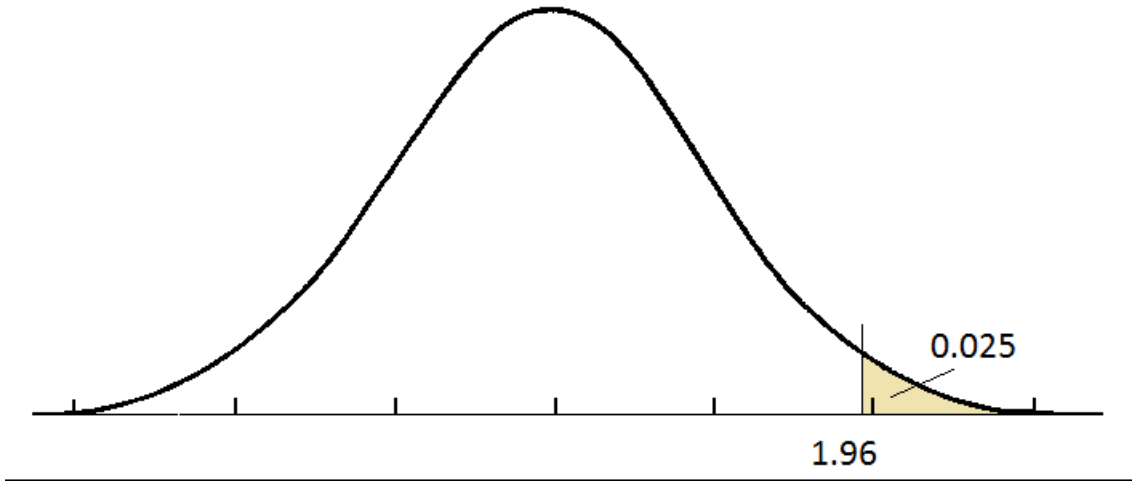


Intrebări

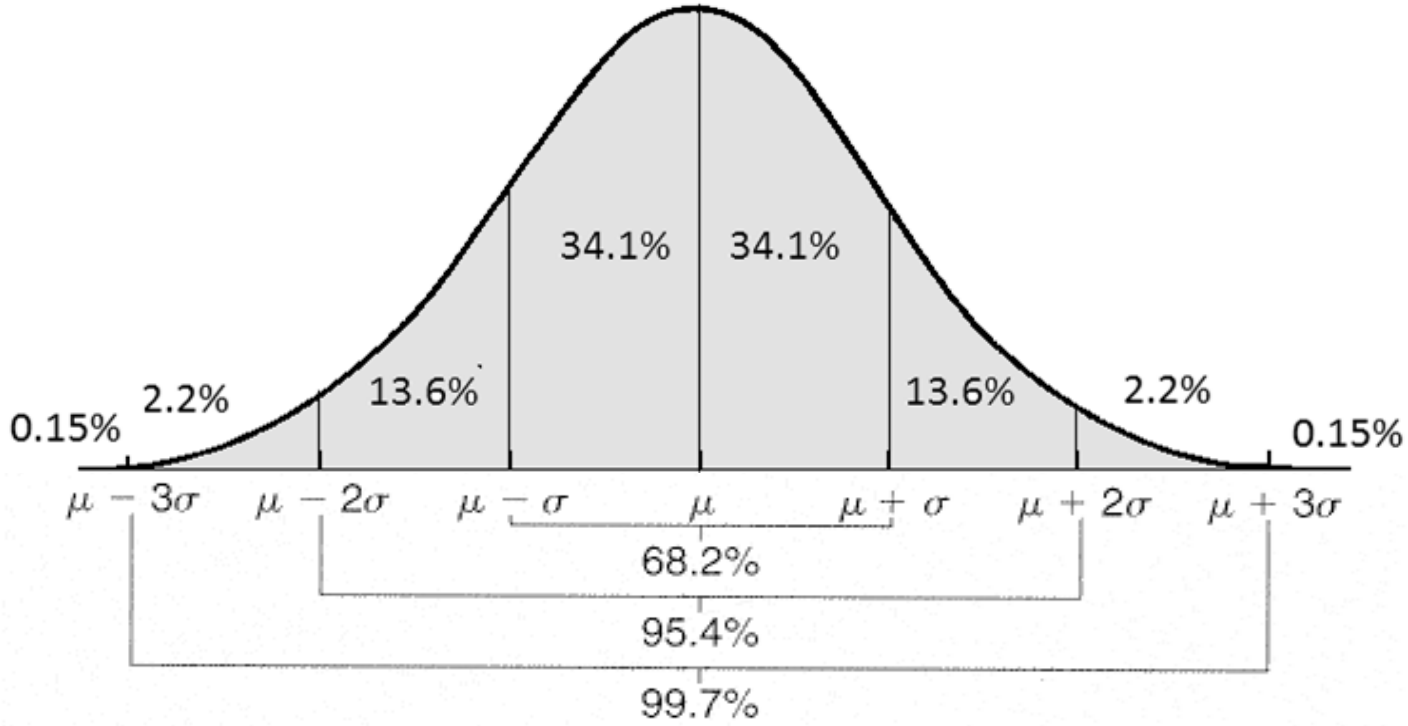
Există tabele cu aceste valori, există funcție în Excel care întoarce aceste valori, nu este necesar să știm să le calculăm, însă aceste două valori sunt de reținut, vom face estimări cu 5% eroare (vezi cursurile viitoare)

- Care valoare a lui Z divide aria in 95% și 5%? $Z_\alpha = 1{,}645$
- Care valoare a lui Z divide aria in 97,5% și 2,5%? $Z_\alpha = 1{,}96$





In mod normal în populație

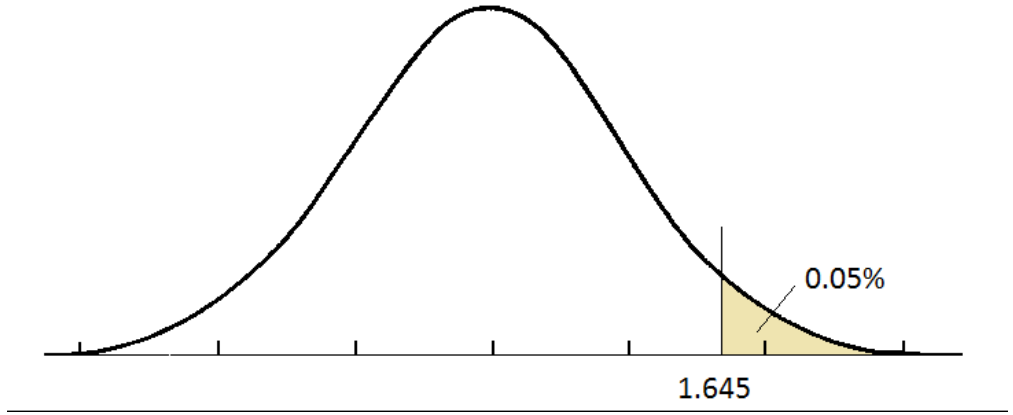


- Aproximativ 95,4% din distribuția de probabilitate - cuprinsă în intervalul medie +/- 2 σ (abaterea standard-populație)
- 95% din distribuția de probabilitate - cuprinsă în intervalul medie +/- 1,96 σ

Exerciții

Colesterolul - distribuit normal cu μ=160 si σ=15 dL/mg.

1. Care valoare a Colesterolului divide aria de sub curbă în 95% si 5%?



$$Z = \frac{X - \mu}{\sigma}$$

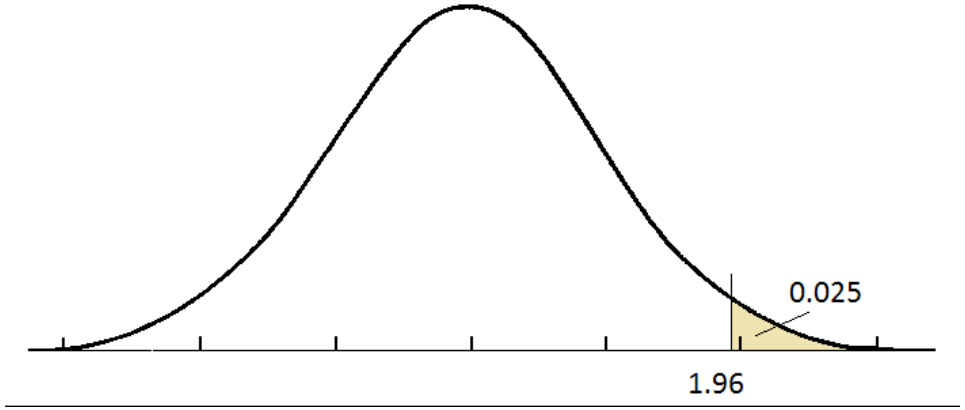
$$1{,}645 = \frac{X - 160}{15}$$

1,645*15=X-160

24,675=X-160

X=184,675

2. Care valoare a Colesterolului divide aria de sub curbă în 97.5% si 2.5%?



$$1{,}96 = \frac{X - 160}{15}$$

1,96*15=X-160

29,4=X-160

X=189,4

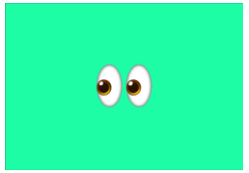
Aplicații: Cum este distribuția datelor?

Dacă aceste condiții sunt îndeplinite

- media ≈ mediana ≈ modulul
- simetria ≈ 0 (între -1 și 1)
- boltirea ≈ 0 (între -1 și 1)
- cuartilele 1 și 3 simetrice față de media aritmetică
- În intervalul medie±abatere standard ∋ minim 68,2% din observaţii;
- În intervalul medie±2*abatere standard ∋ minim 95,4% din observaţii;
- În intervalul medie±3*abatere standard ∋ minim 99,7% din observaţii,

- atunci distribuția datelor obținute empiric se apropie de distribuția normală

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

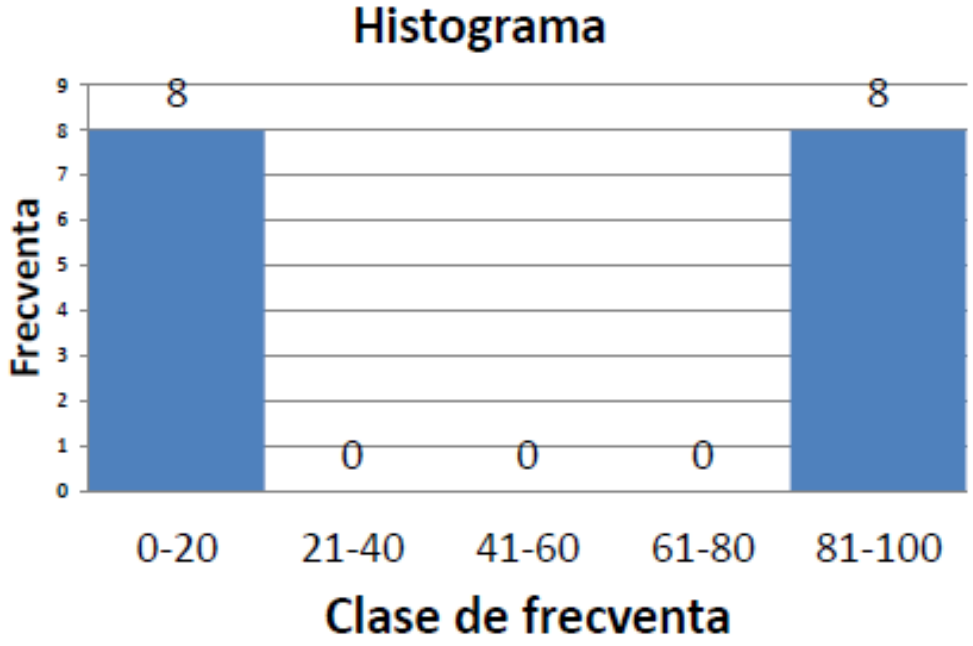


https://app.wooclap.com/CURS7MGRO?from=instruction-slide

Exemplu – Seria 1

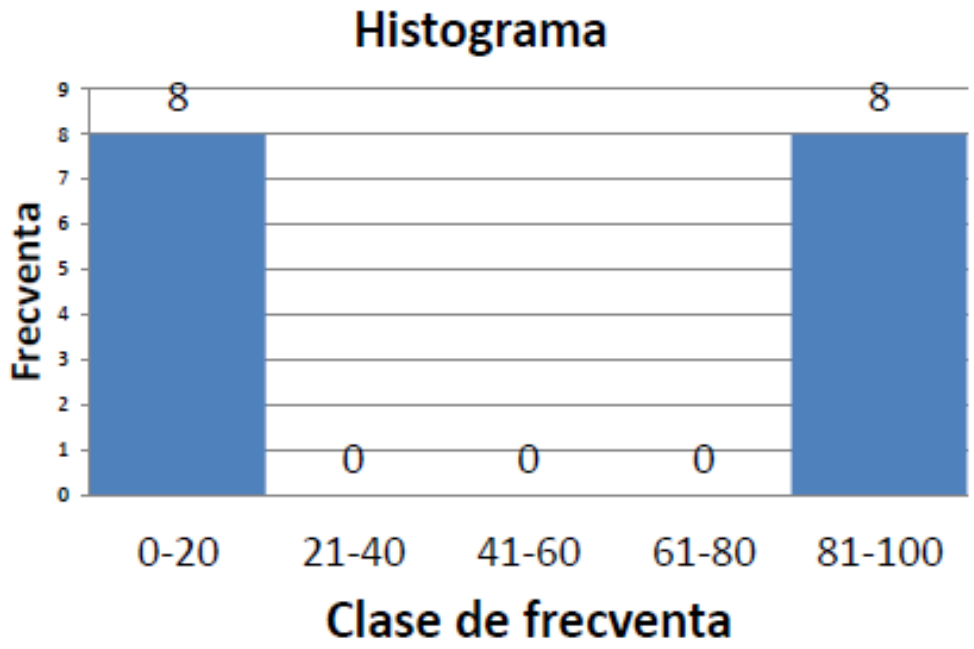
Seria 1
1
1
2
3
5
6
6
7
93
94
94
95
97
98
98
100

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



Ne arată diferențe mari față de distribuția normală

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



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

Deviaţia 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
- Deviaţia standard = 47,70

- Media - deviaţia standard = 50 - 47,7 = 2,3
- Media + deviaţia standard = 50 + 47,7 = 97,7

- intervalul media ± deviaţia 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

16

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

Media ± deviaţia standard = [50 - 47,7; 50 + 47,7] = [2,3; 97,7]

- 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 ± deviaţia standard sunt minim 68,3% din date

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

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

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

- Media ± deviaţia 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 ± deviaţia standard sunt minim 68,3% din date

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

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

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

- Media ± deviaţia 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 ± 2*deviația standard = [50-2*47,7; 50+2*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 media ± deviaţia standard sunt minim 68,3% din date

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

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

16

Medie ± 2*deviația standard = [50-2*47,7; 50+2*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

16

Ca să fie distribuție normală:

in intervalul media ± deviaţia standard sunt minim 68,3% din date

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

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

Media aritmetică = 50

Deviaţia standard = 47,70

Media ± deviaţia 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 ± deviaţia standard sunt minim 68,3% din date

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

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

16 Media aritmetică = 50

Deviaţia standard = 47,70

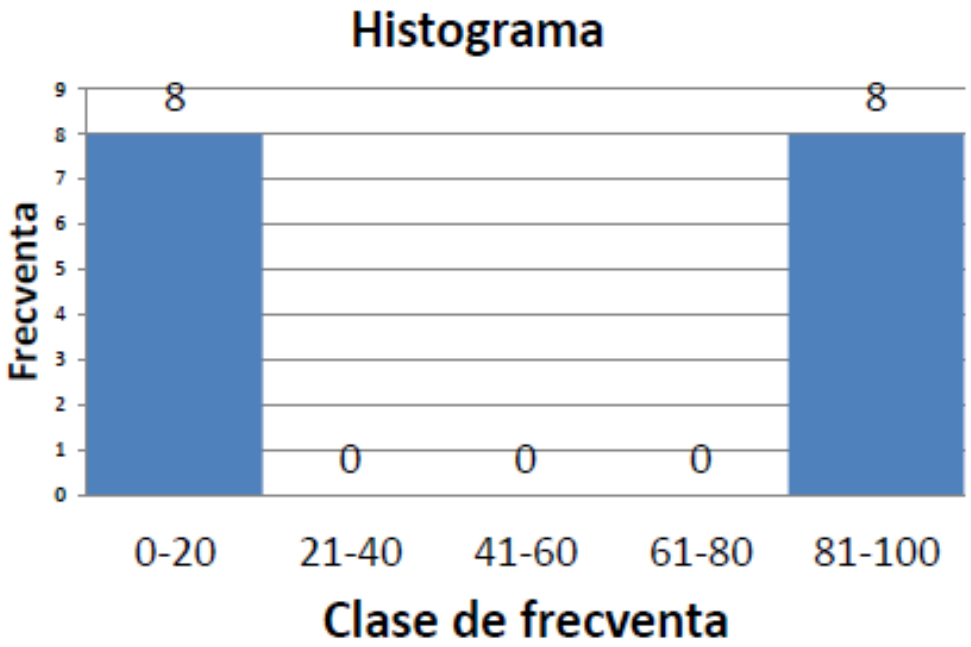
Media ± deviaţia standard = [2,3; 97,7] cu 10 valori, adică 62,5% dintre date

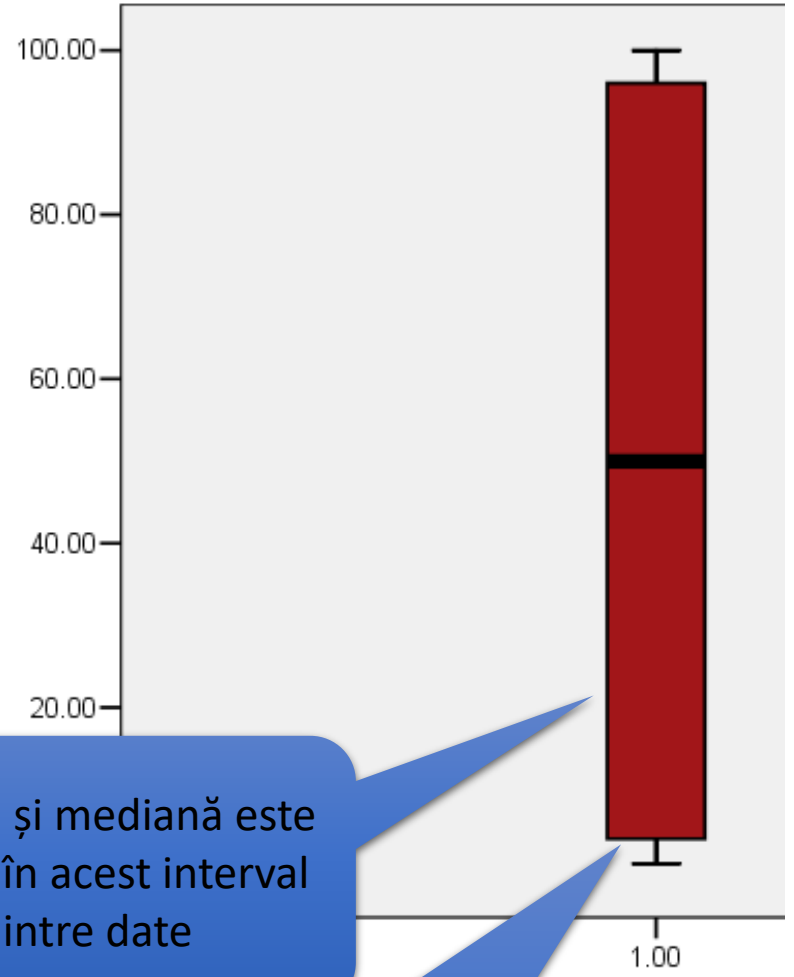
Mean ± 2*st.dev = [-45,39; 145,39] sunt 16 valori, adica 100% dintre date

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





Între percentila 25 și mediană este o distanţă mare = în acest interval avem 25% dintre date

Între minim şi percentila 25 este o distanţă mică = în acest interval avem 25% dintre date

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 - nu are

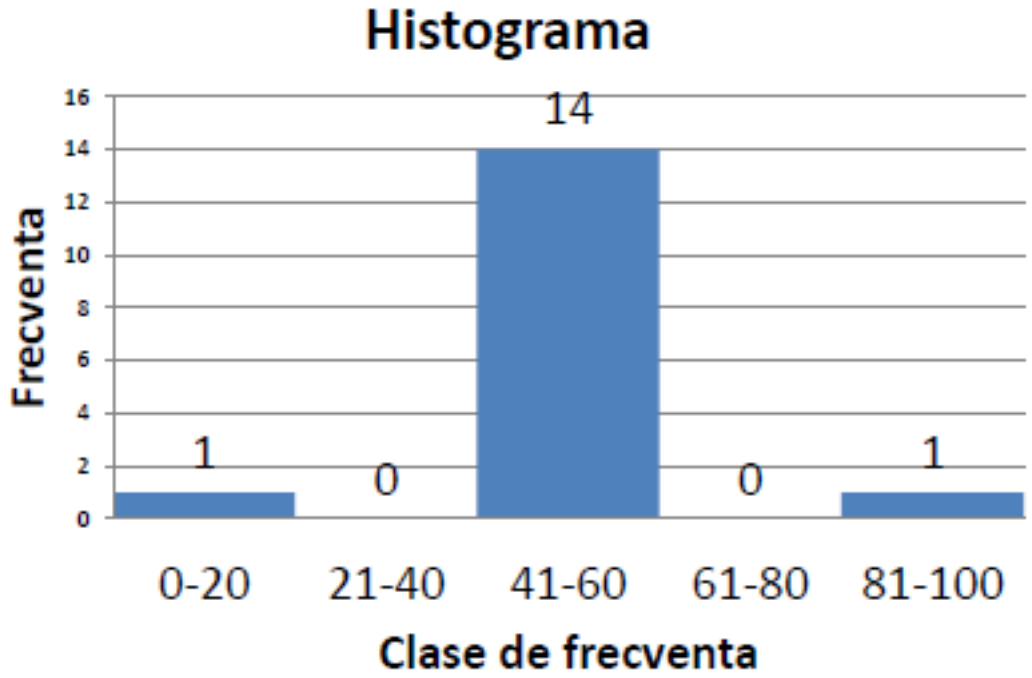
Deviaţia standard = 18,37

Cuartila 1 = 47,5

Cuartila 3 = 52,5

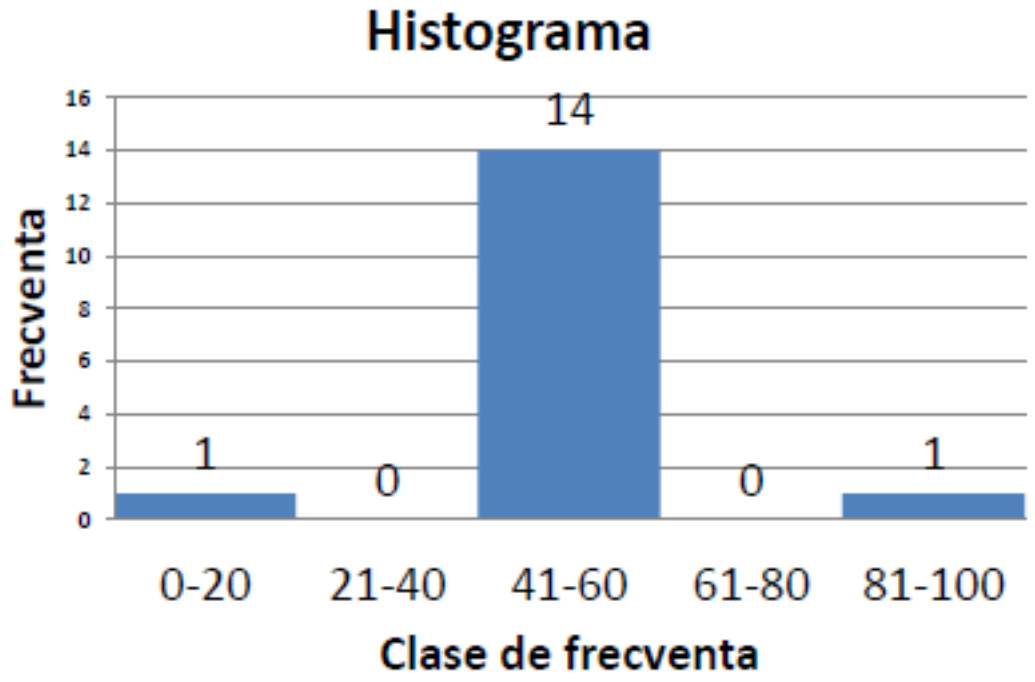
Simetria = 0,09

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
56
100



Ca să fie distrib. normală:

Minim 68,3% din date

Minim 95,4% din date

Minim 99,7% din date

Deviaţia standard este mică, concluzie: cazurile sunt apropiate de medie

Media aritmetică = 50

Deviaţia standard = 18,37

Seria 2
1
44
45
46
48
48
49
50
50
51
52
52
54
55
56
100

16

Media aritmetică = 50
Deviaţia standard = 18,37

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

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

Seria 2
1
44
45
46
48
48
49
50
50
51
52
52
54
55
56
100

16

Ca să fie distrib. normală:

Minim 68,3% din date

Minim 95,4% din date

Minim 99,7% din date

Media aritmetică = 50
Deviaţia standard = 18,37

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

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

Seria 2
1
44
45
46
48
48
49
50
50
51
52
52
54
55
56
100

16

Ca să fie distrib. normală:

Minim 68,3% din date

Minim 95,4% din date

Minim 99,7% din date

Media aritmetică = 50

Deviaţia standard = 18,37

Media ± 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

Seria 2
1
44
45
46
48
48
49
50
50
51
52
52
54
55
56
100

16

Ca să fie distrib. normală:

Minim 68,3% din date

Minim 95,4% din date

Minim 99,7% din date

Media aritmetică = 50

Deviaţia standard = 18,37

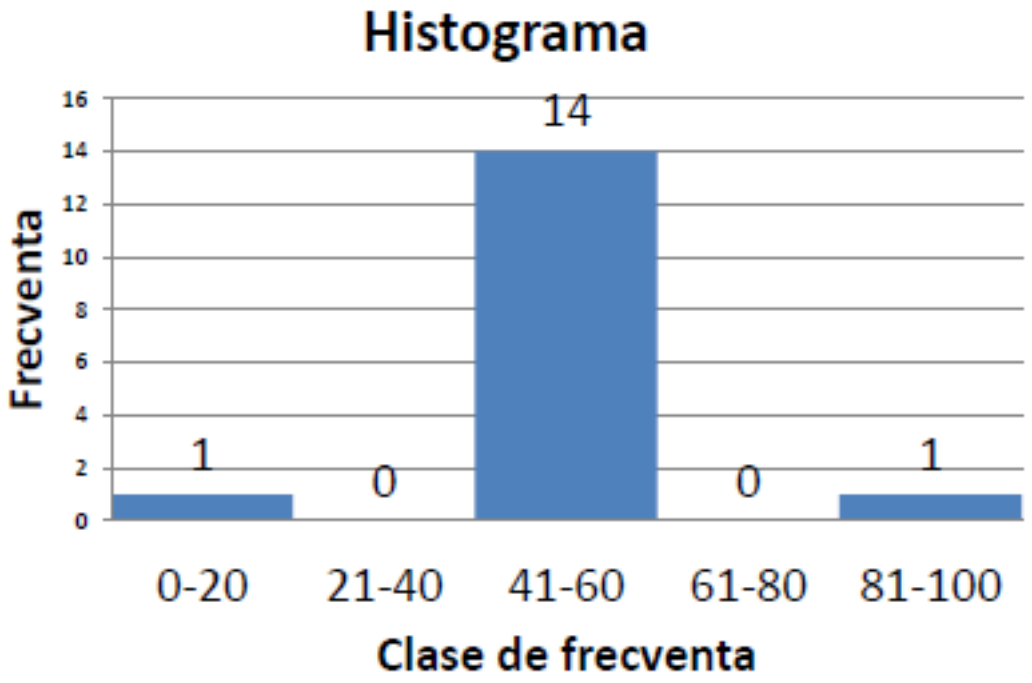
Media ± dev.st = [31,63; 68,37] sunt 14 valori - 87,5% din date

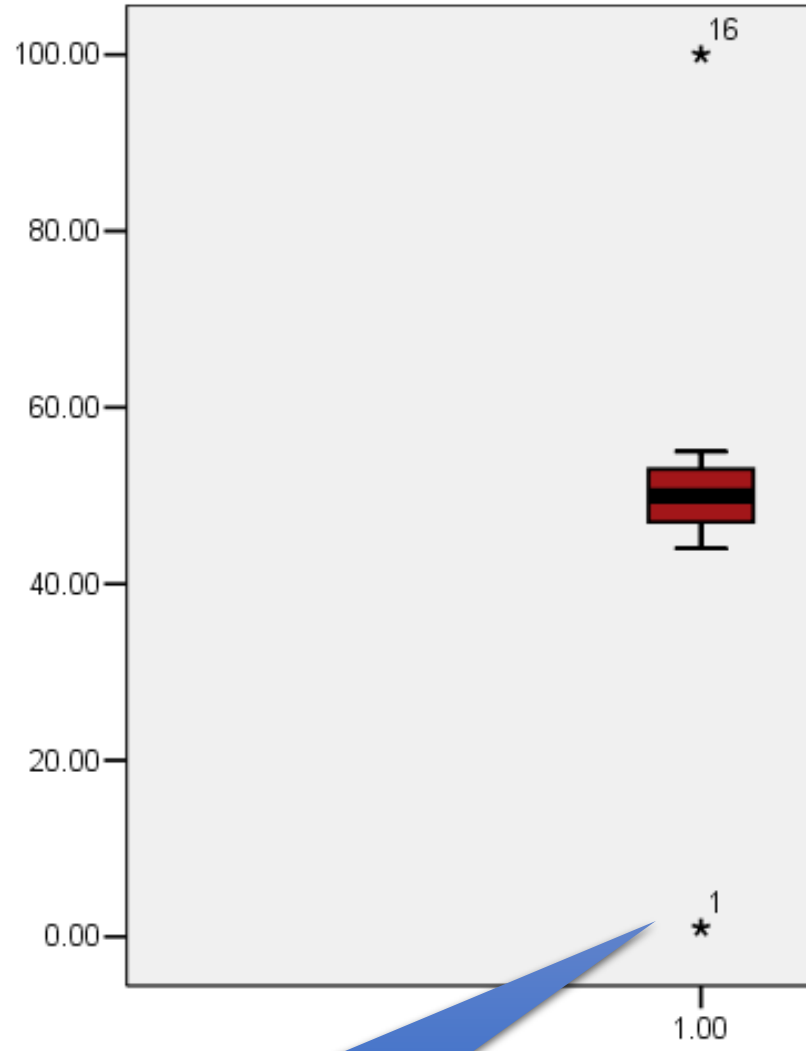
Media ± 2*dev.st. = [13,26; 86,74] sunt 14 valori - 87,5% din date

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

Distribuția nu este apropiată de cea normală

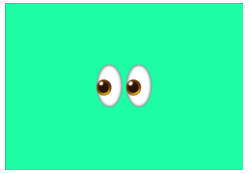
Seria 2
1
44
45
46
48
48
49
50
50
51
52
52
54
55
56
100





Caz extrem

Seria 3
1
11
24
29
36
41
45
50
50
55
59
64
71
76
88
100



https://app.wooclap.com/CURS7MGRO?from=instruction-slide

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

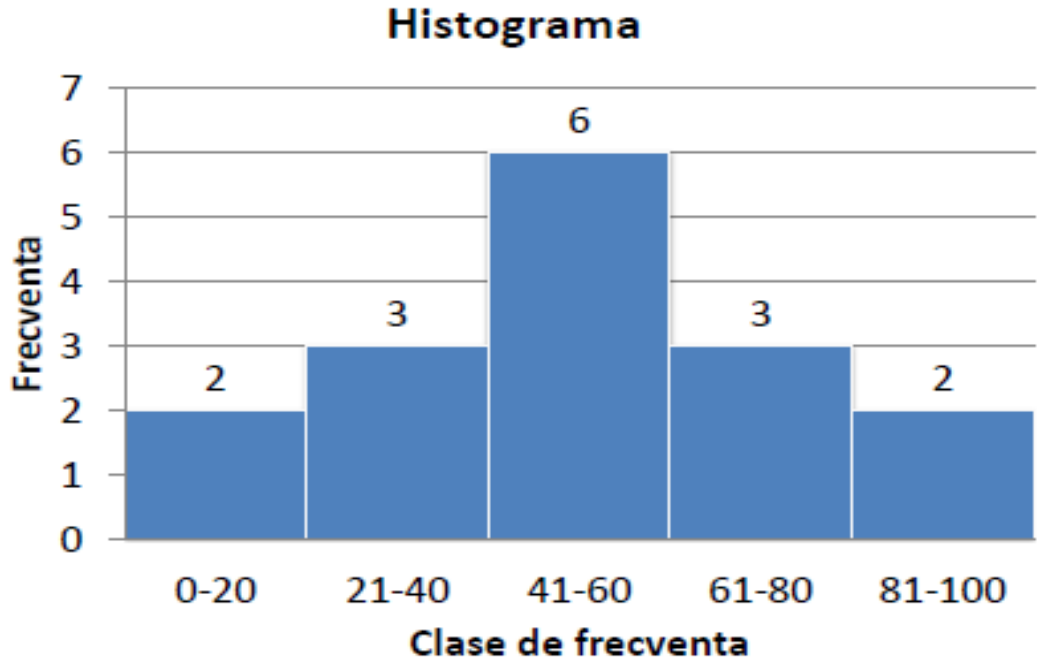
Deviaţia standard = 26,71

Cuartila 1 = 34,25

Cuartila 3 = 65,75

Simetria = 0,01

Boltirea = -0,22



Distribuția este apropiată de cea normală

Seria 3
1
11
24
29
36
41
45
49
51
55
59
64
71
76
88
100

Ca să fie distrib. normală:

Minim 68,3% din date

Minim 95,4% din date

Minim 99,7% din date

Media aritmetică = 50

Deviaţia standard = 26,71

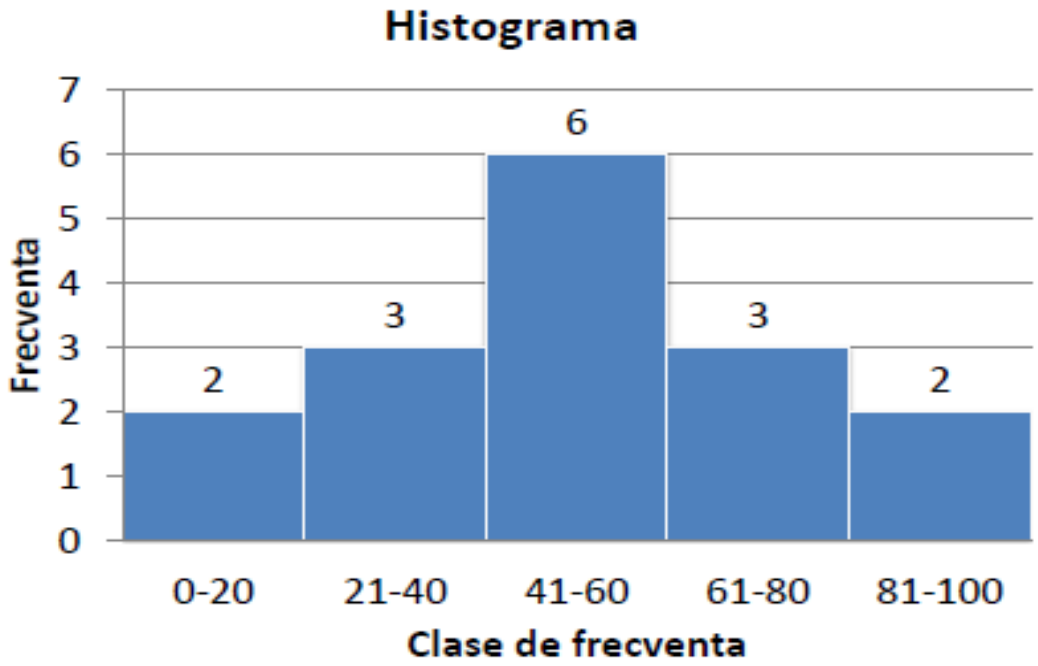
Media ± dev.st. = [23,28; 76,72] sunt 87,5% din date

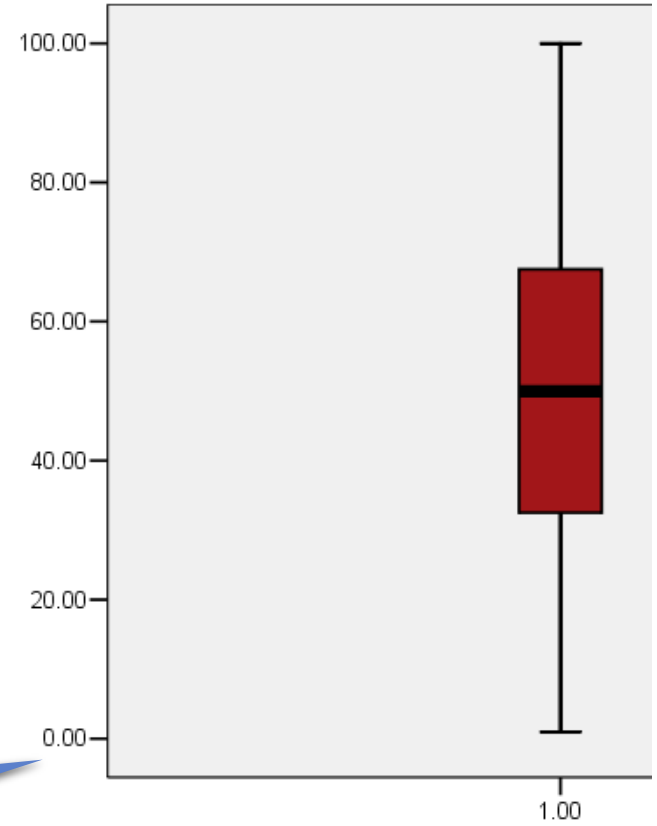
Media ± 2*dev.st. = [-3,43; 103,43] sunt 100% din date

Media ± 3*dev. st. = [-30,15; 130,15] sunt 100% din date

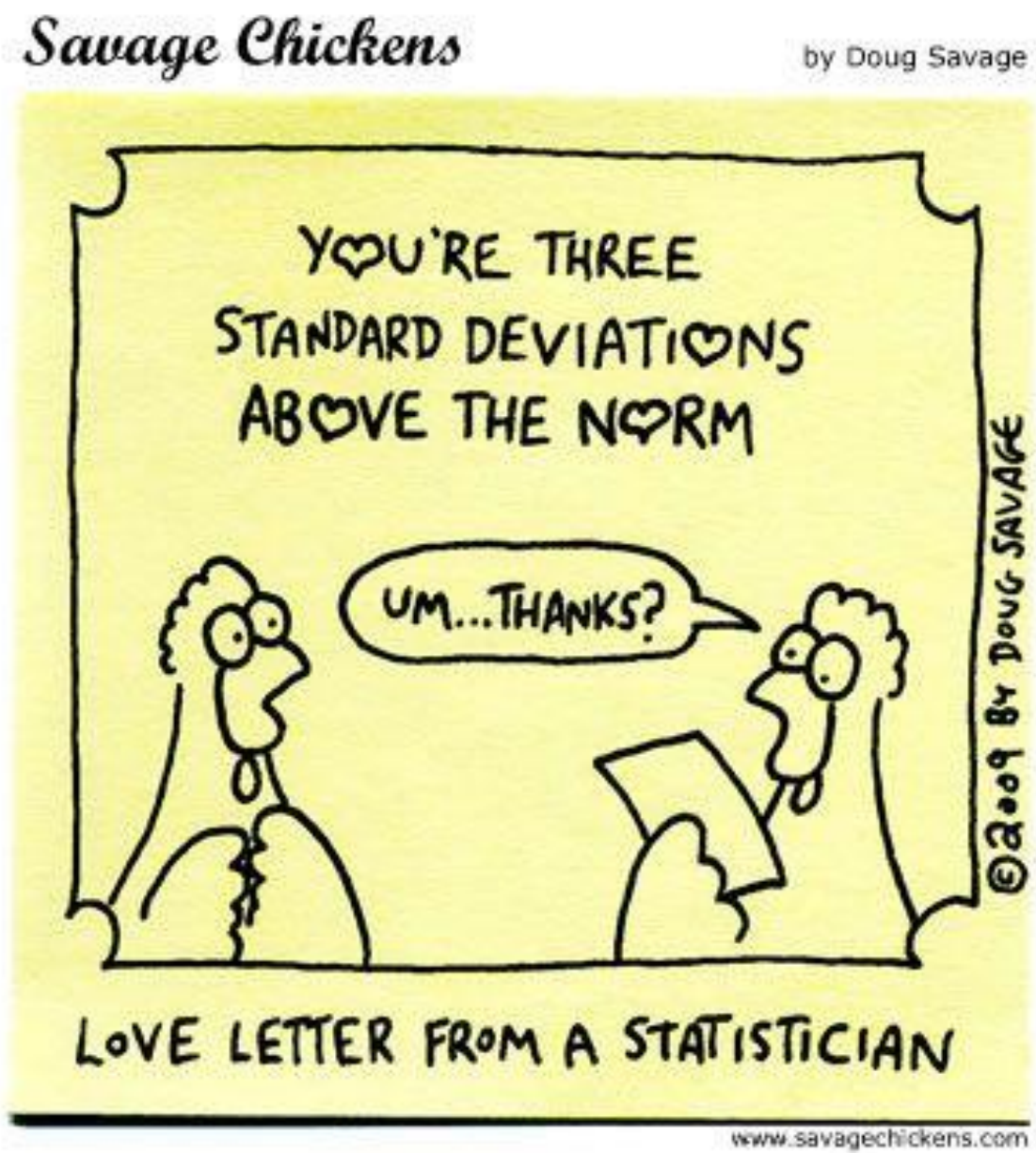
Distribuția este apropiată de cea normală

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





Distribuția este apropiată de cea normală


Savage Chickens
by Doug Savage
YOU'RE THREE STANDARD DEVIATIONS ABOVE THE NORM
UM...THANKS?
©2009 BY DOUG SAVAGE
LOVE LETTER FROM A STATISTICIAN
www.savagechickens.com

LEGEA BINOMIALĂ
DISTRIBUŢIA LUI BERNOULLI

1700-1782

- eveniment cu 2 rezultate posibile: *succes* şi *eşec*
- probabilitatea p de succes şi q = 1 - p de eşec – aceleași la fiecare încercare
- n încercări repetate - independente una de cealaltă
- Care este probabilitatea ca din n încercări să avem succes de k ori?

Estimăm cu ajutorul legii binomiale

Ex.

- moartea
  - 10 pacienți Covid-19 la secția ATI, moartea are o probabilitate constantă pentru fiecare dintre aceștia 0,66
    - care este probabilitatea ca să moară 2 din cei 10 pacienți?
- ameliorarea în cazul tratamentului cancerului
  - 5 pacienți cu leucemie, ameliorarea are o probabilitate constantă pentru fiecare dintre aceștia 0,5
    - care este probabilitatea ca să se amelioreze 4 din cei 5 pacienți? dar 5 din 5?
- trăsături genetice, apariția bolii la cei expuși la un factor de risc etc.

- 2 pacienți Covid-19 secția ATI,
  - supraviețuirea are o probabilitate p = 0,7; moartea q=0,3

- Care este probabilitatea ca cei doi pacienți să supraviețuiască?
- P(ambii pacienți supraviețuiesc) = 0,7 * 0,7 = 0,49

	Primul pacient	Al doilea pacient	Probabilitate
P(ambii pacienți supraviețuiesc)	S	S	= 0,7 * 0,7 = 0,49
P(un pacient supraviețuiește și un pacient moare)	S	M	= 2 * 0,7 * 0,3 = = 0,42
	M	S	
P(ambii pacienți mor)	M	M	=0,3 *0,3 = 0,09

Auld SC, Harrington KRV, Adelman MW, Robichaux CJ, Overton EC, Caridi-Scheible M, Coopersmith CM, Murphy DJ; Emory COVID-19 Quality and Clinical Research Collaborative. Trends in ICU Mortality From Coronavirus Disease 2019: A Tale of Three Surges. Crit Care Med. 2022 Feb 1;50(2):245-255.

LEGEA BINOMIALĂ SAU DISTRIBUŢIA LUI BERNOULLI

- n încercări repetate

p – probabilitatea de a avea reuşită

q – probabilitatea de a avea ne-reuşită

- Probabilitatea pentru k reușite:

$$\Pr(X = k) = C_n^k p^k q^{n-k}$$

1700-1782

Combinări de n luate câte k

$$C_n^k = \frac{n!}{k!*(n-k)!}$$

k factorial

$k! = k \times (k-1) \times (k-2) \times \ldots \times 2 \times 1$

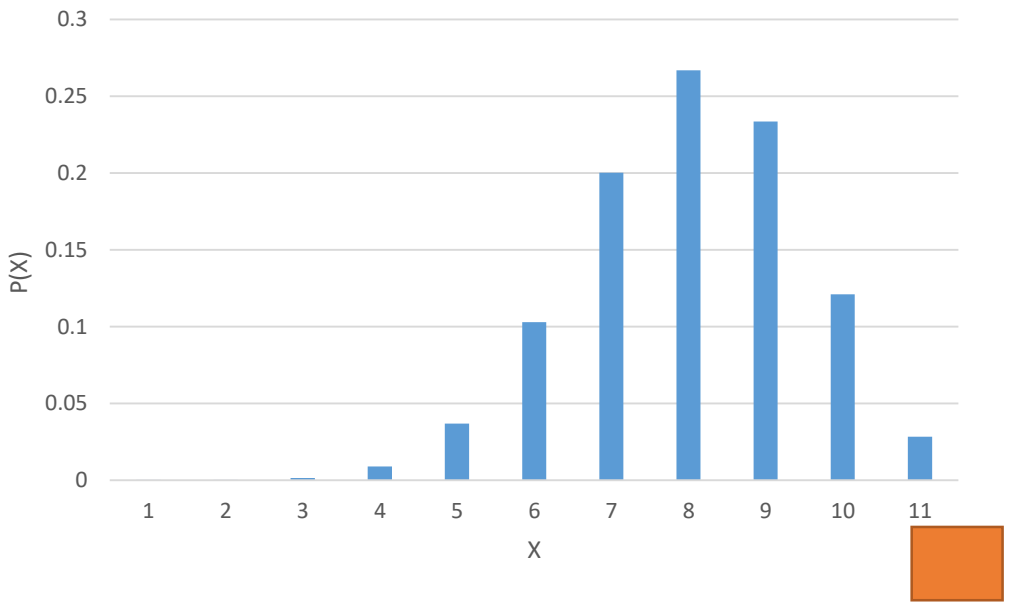
- 10 pacienți Covid-19 secția ATI,
  - supraviețuirea are o probabilitate p = 0,7; moartea q=0,3

- Care este probabilitatea ca exact unul să supraviețuiască?

- P(1 supraviețuiește din 10)=$\frac{10!}{1!(10-1)!}0{,}7^{1}0{,}3^{9} = \frac{3628800}{362880} * 0{,}7 * 0{,}000019683 = 0{,}000138$

- 10 pacienți Covid-19 secția ATI,
  - supraviețuirea are o probabilitate p = 0,7; moartea q=0,3

Nr. pacienți care supraviețuiesc	C_n^k	p^k	$q^{(n-k)}$	P(X)
0	1	1	0.00000590	0.000005
1	10	0,7	0.00001968	0.000137
2	45	0.49	0.00006561	0.001446
3	120	0.343	0.0002187	0.009001
4	210	0.2401	0.000729	0.036756
5	252	0.16807	0.00243	0.102919
6	210	0.117649	0.0081	0.200120
7	120	0.082354	0.027	0.266826
8	45	0.057648	0.09	0.233474
9	10	0.040354	0.3	0.121062
10	1	0.028248	1	0.028248



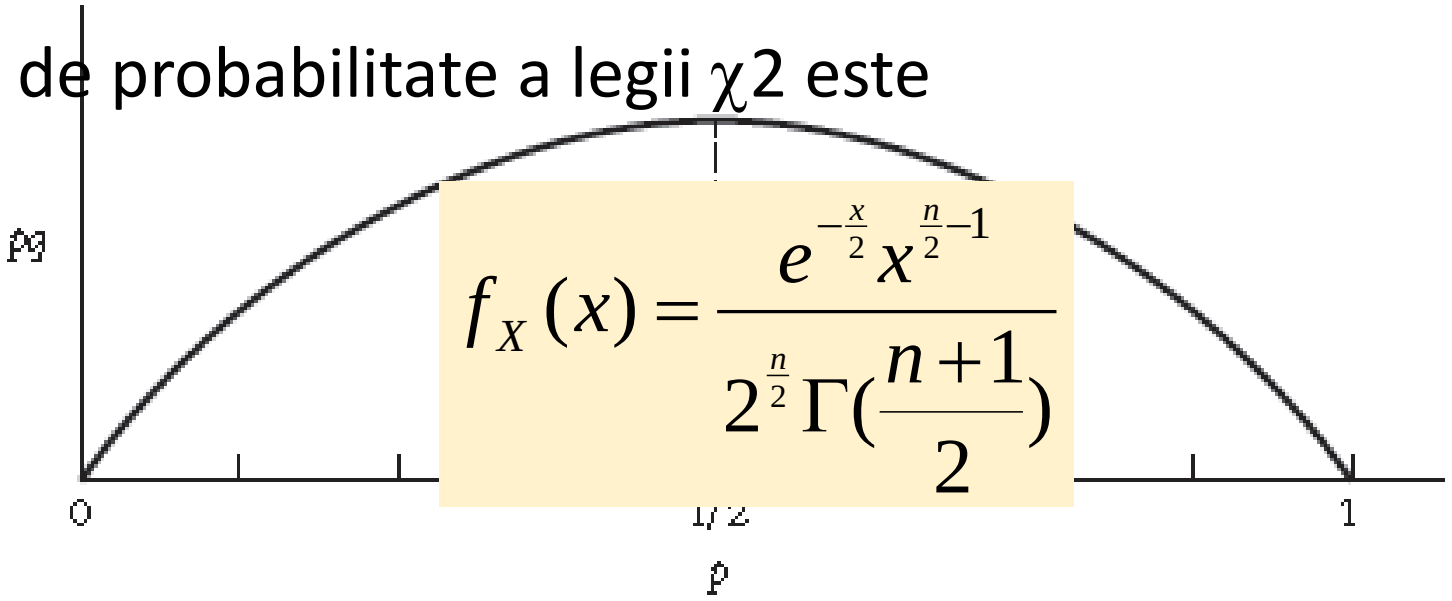
- Cercetarea medicală
  - pe >10 pacienți → calcule elaborioase
  - se va folosi aproximarea distribuției binomiale, adică distribuția normală

LEGEA BINOMIALĂ SAU DISTRIBUŢIA LUI BERNOULLI

- speranţa matematică a legii binomiale este:
  - $\bar{X}$ = np,
- iar variaţia:
  - S^2 =npq,
- şi deci abaterea standard: S=$\sqrt{npq}$

$$X : \begin{pmatrix} k \\ C_n^k p^k q^{n-k} \end{pmatrix}$$

- .
- Densitatea de probabilitate a legii $\chi 2$ este



$$f_X(x) = \frac{e^{-\frac{x}{2}} x^{\frac{n}{2}-1}}{2^{\frac{n}{2}} \Gamma(\frac{n+1}{2})}$$

$$X : \begin{pmatrix} k \\ e^{-\theta} \frac{\theta^k}{k!} \end{pmatrix}$$

Exemplu: utilizarea de anticoncepționale de generația IV

- Literatura de specialitate: efecte secundare la 3% dintre utilizatoare.
- Care este probabilitatea ca **din 10 utilizatoare 3** să aibă un astfel de efect secundar?
- Răspuns: In acest caz avem n = 10, k = 3, p =0.03, q = 1-p = 0.97
- Atunci: $\boldsymbol{Pr(X = 3) = C_{10}^{3}(0,03)^{3}(0,97)^{7} = 0,0026}$

LEGEA LUI POISSON

1781–1840

- variabilă discretă infinită
- Modelul legii POISSON:
  - Numărul de realizări ale unui eveniment într-un interval dat de timp sau spaţiu

- exemplu
  - numărul de internări pe an într-un spital,
  - numărul de ambulanțe necesare
  - numărul de bacterii într-un mililitru de apă,
  - numărul de dezintegrări ale unei substanţe radioactive într-un interval de timp T dat

LEGEA LUI POISSON

- caracterizată de un parametru θ - numărul mediu teoretic (aşteptat) de realizări ale evenimentului în intervalul considerat
- Probabilitatea ca să se realizeze evenimentul de k ori în intervalul considerat:

$$\Pr(X = k) = \frac{e^{-\theta}\theta^{k}}{k!}$$

Unde e≈ 2,71828

LEGEA LUI POISSON

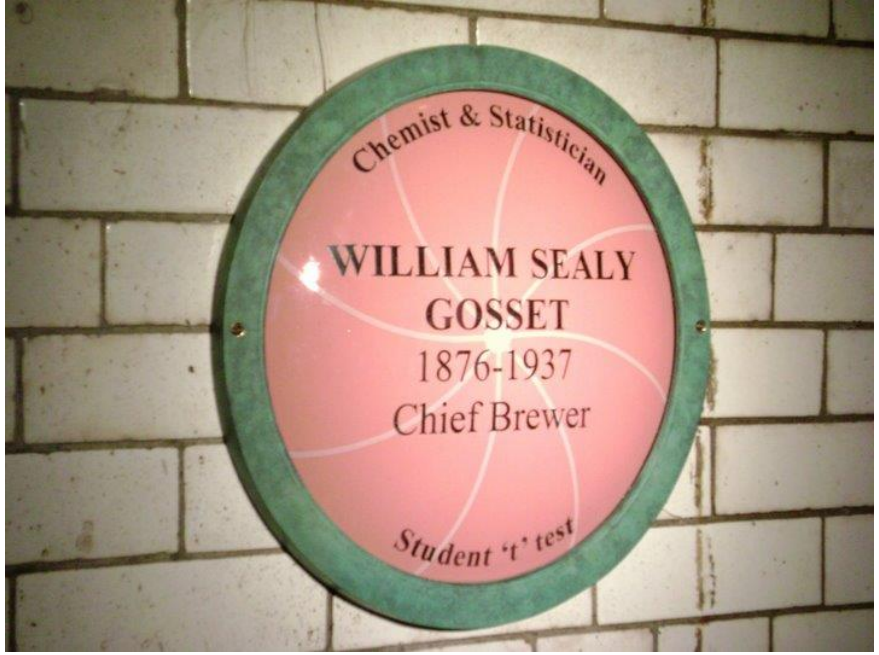
- Speranţa matematică şi variaţia în cazul legii lui Poisson sunt egale ambele cu θ, adică :
  - $M(X) = Var(X) = \theta$.

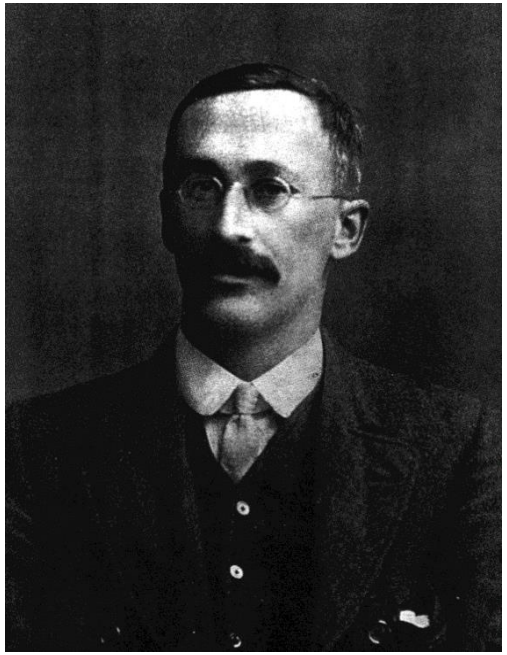
Exemplu: boala Crohn într-un interval de 1 an

- Frecvența bolii Crohn (inflamație intestinală) = 3 la 100 = 0,03.
- Care este probabilitatea ca sa avem **6 persoane cu boala Crohn într-un an**?

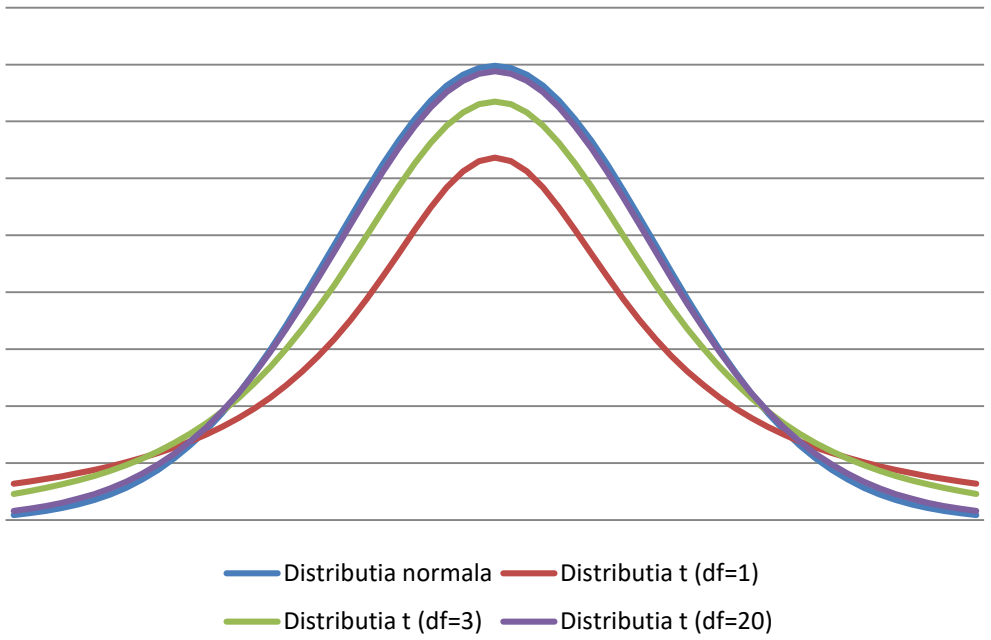
- Răspuns:
- θ =0,03,
- Pr(k=6/an)=$\frac{e^{-0,03}\cdot 0,03^{6}}{6!}$=0,83*10^{-13}

LEGEA STUDENT (T)





1876–1937



LEGEA STUDENT (T)

- variabile aleatoare continue
- Variabila aleatoare Student t
  - depinde de un singur parametru - numărul de grade de libertate
- X_0, X_1, …, X_n variabile aleatoare independente care urmează legea normală centrată redusă. Atunci variabila aleatoare

$$T_n = \frac{X_0 \sqrt{n}}{\sqrt{\sum_{i-1}^{n} X_i^2}}$$

- urmează o lege de probabilitate Student cu n grade de libertate.

LEGEA STUDENT (T)

- Densitatea de probabilitate a variabilei aleatoare Student T_n este:

$$f_{T_n}(x) = \frac{1}{\sqrt{2\pi}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \frac{1}{(1+\frac{x^2}{n})^{\frac{n+1}{2}}}$$

- unde Γ este funcţia Gamma definită astfel:

$$\Gamma(\frac{n+1}{2}) = \int_0^{+\infty} e^{-t} t^{n-1} dt$$

LEGEA STUDENT (T)

- Distribuţia acestei variabile aleatoare este simetrică în raport cu originea şi are o formă de clopot:
- $$Pr[T_k < -x] = Pr[T_k > x].$$
- Atunci când k tinde la infinit, distribuţia Student tinde către o distribuţie normală redusă.
- Dacă n > 30 legea lui Student şi legea normală sunt foarte apropiate.
- Această variabilă aleatoare este utilizată, în anumite condiţii de normalitate, în testul de comparaţie a mediilor numit şi testul Student sau testul t.

LEGEA HI-PĂTRAT (PEARSON)

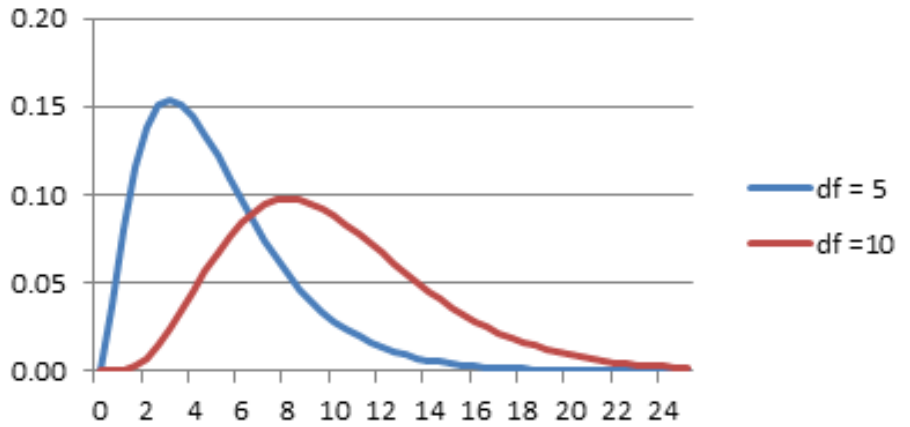
1857–1936

X_0, X_1, …, X_n variabile aleatoare independente care urmează legea normală centrată redusă,

- fiecare având o medie egală cu zero şi abatere standard egală cu 1. Variabila X

$$X = X_1^2 + X_2^2 + \ldots + X_n^2$$

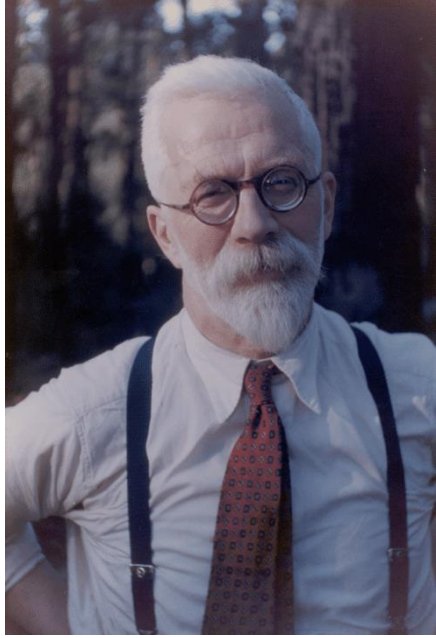
este $\chi 2$ distribuită cu n grade de libertate.



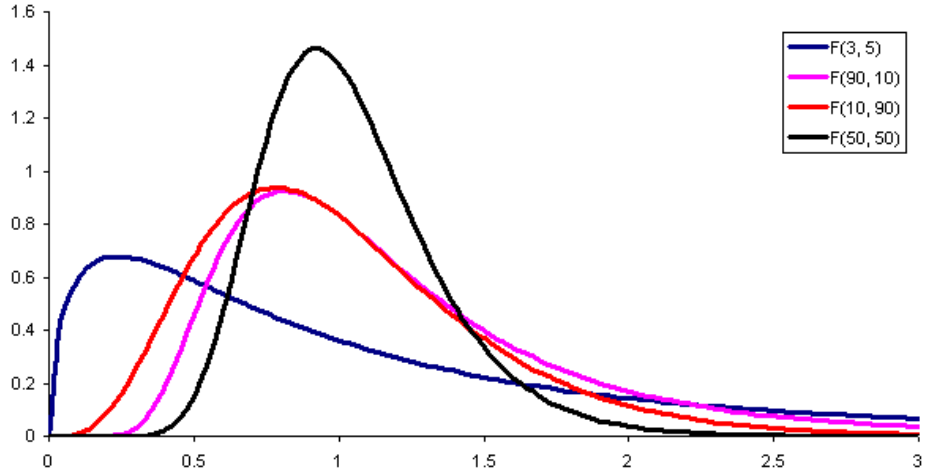
LEGEA F (FISHER)

- Distribuţia F introdusă de R. A. Fisher,
- X_1, X_2 variabile cu distribuţie Hi-pătrat, d_1, d_2 numărul gradelor lor de libertate

$X = \frac{\frac{X_1}{d_1}}{\frac{X_2}{d_2}}$ este o variabilă ce urmează legea Fisher



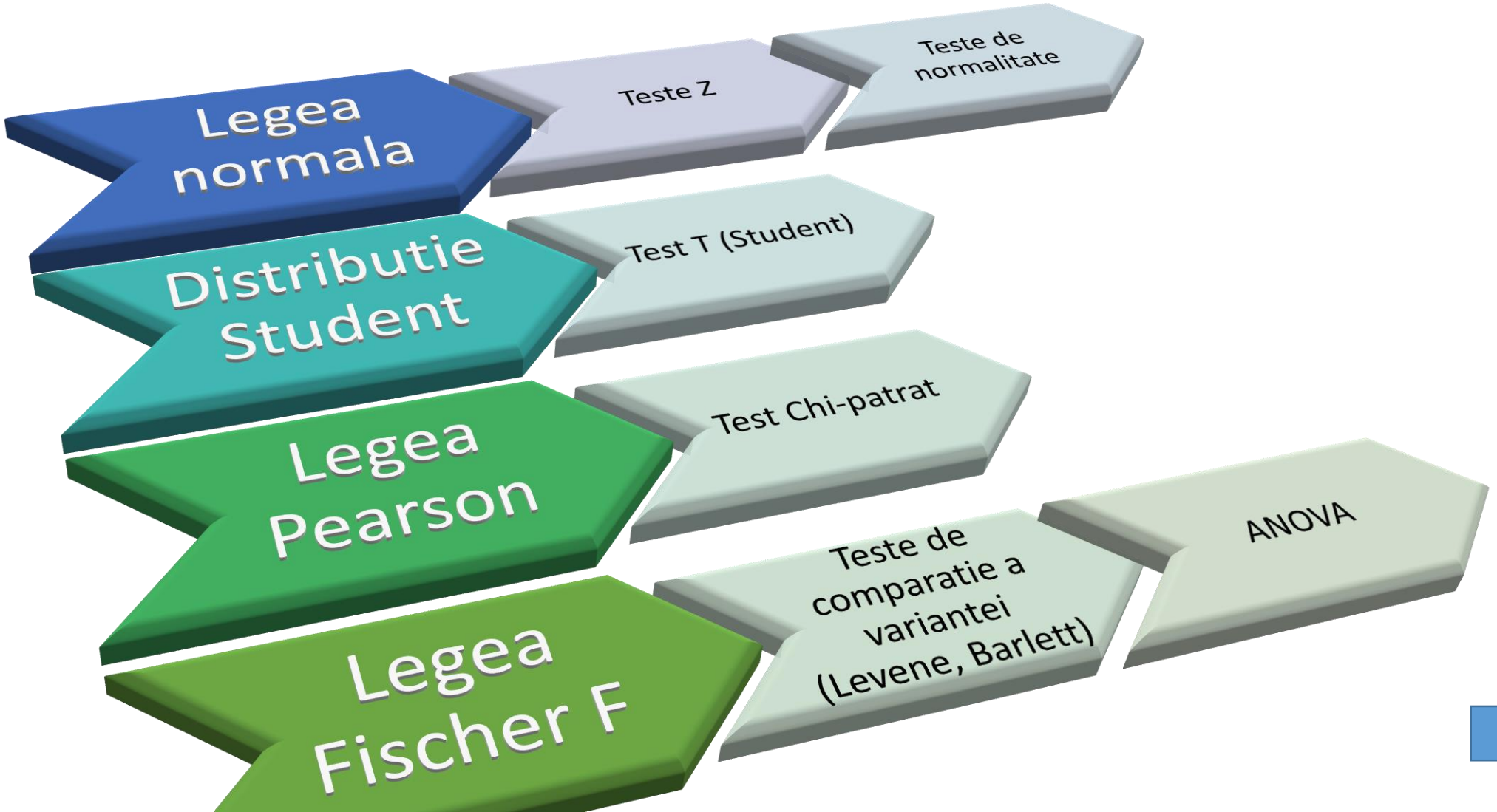
1890–1962



LEGEA F (FISHER)

- In general, pentru dn şi dm > 2 distribuţia F este unimodală şi pozitiv asimetrică. Atunci când numărul gradelor de libertate creşte distribuţia F se apropie pe domeniul său de definiţie de o distribuţie normală.
- Această distribuţie este utilizată în testele de comparaţie a variaţiilor şi ca aplicaţie a acestora în testele ANOVA.

Distributie – Test statistic



Concluzie

- Selectăm eșantioane aleator
- Putem modela fenomenele
  - cu legi de probabilitate cunoscute

Distribuția paranormală

- Mulțumesc!!!

Legendă de ținut minte pentru pasionați ! important pentru înțelegerea noțiunilor ce urmează a fi prezentate