UMF
UNIVERSITATEA DE
MEDICINĂ ŞI FARMACIE
IULIU HAŢIEGANU
CLUJ-NAPOCA

Noţiuni de statistică inferenţială

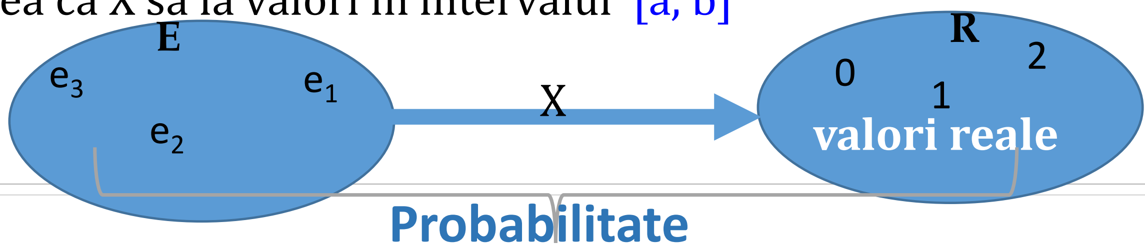
Intervalul de încredere (IC)

Variabile aleatoare

- Variabilă **aleatoare =** variabilă a cărei valoare este un număr determinat de un eveniment asociat unei experiențe aleatoare (**valorile unei variabile aleatoare sunt mutual exclusive având o probabilitate dată**)
- O variabilă aleatoare X este o variabilă numerică care poate lua diferite valori cu anumite probabilități
- se notează cu majuscule: X, M, N,…în timp ce valorile lor sunt notate cu litere mici (x, m, n…)
- **Probabilitatea ca variabila X să ia valoarea x**: $Pr(X=x)=p(x)$
- **Probabilitatea ca variabile X să ia o valoare mai mică sau egală cu x**:

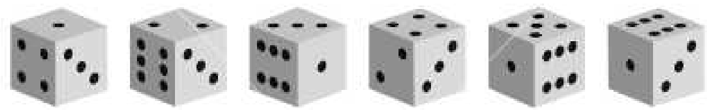
  $Pr(X \le x)=F(x)$ (funcție de repartiție)

  $Pr(a \le X \le b)$ → probabilitatea ca X să ia valori în intervalul [a, b]



Exemple: variabile aleatoare

Rezultatele unei experiențe aleatoare



Spațiul fundamental

→ Asociem numere

1 2 3 4 5 6

Variabila aleatoare

Distribuția de probabilitate

→ Asociem probabilități

1/6 1/6 1/6 1/6 1/6 1/6

Distribuția de probabilitate a unei variabile

- **Distribuția de probabilitate**: mulțimea (valorilor posibile ale variabilei cu probabilitățile acestora

ou

- **Legea de probabilitate:** relația (matematică) între valorile variabilei și probabilitățile lor
- Distribuția de probabilitate poate fi discretă sau continuă (în funcție de tipul variabilei)

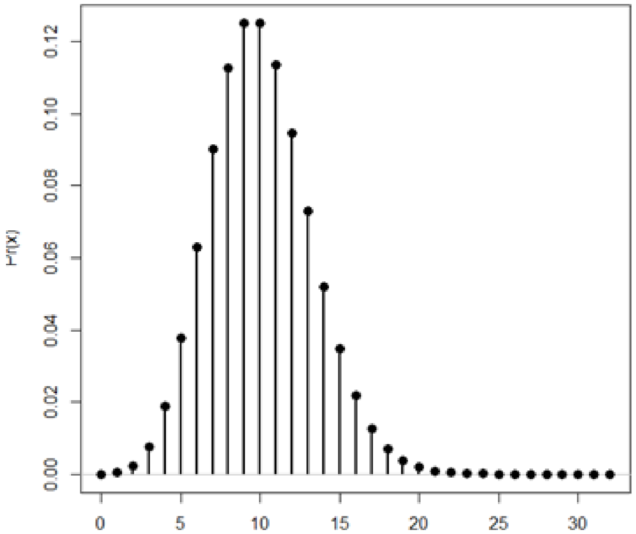
• Distribuția de probabilitate discretă: tabel de forma

$$\boldsymbol{X}:\begin{pmatrix} \boldsymbol{x_1} & \boldsymbol{x_2} & \dots & \boldsymbol{x_n} \\ \boldsymbol{Pr(x_1)} & \boldsymbol{Pr(x_2)} & \dots & \boldsymbol{Pr(x_n)} \end{pmatrix} \text{ unde } \mathbf{Pr}(\boldsymbol{x_1}) + \boldsymbol{Pr}(\boldsymbol{x_2}) + \cdots + \boldsymbol{Pr}(\boldsymbol{x_n}) = \mathbf{1}$$

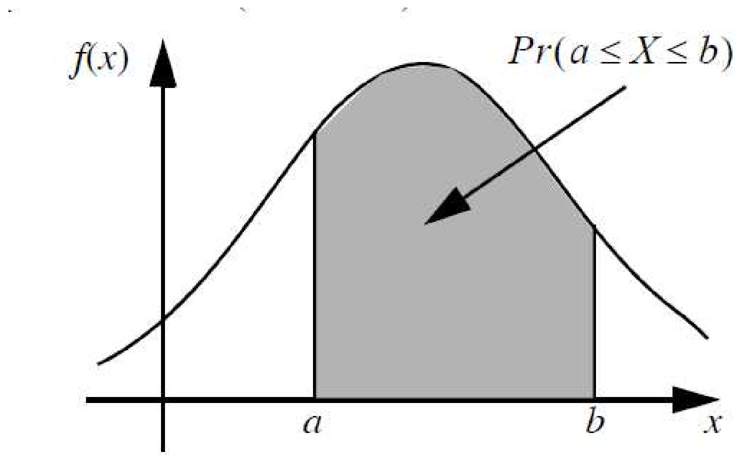
Distribuţia de probabilitate discretă vs. continuă

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

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



Valori posibile ale variabilei X

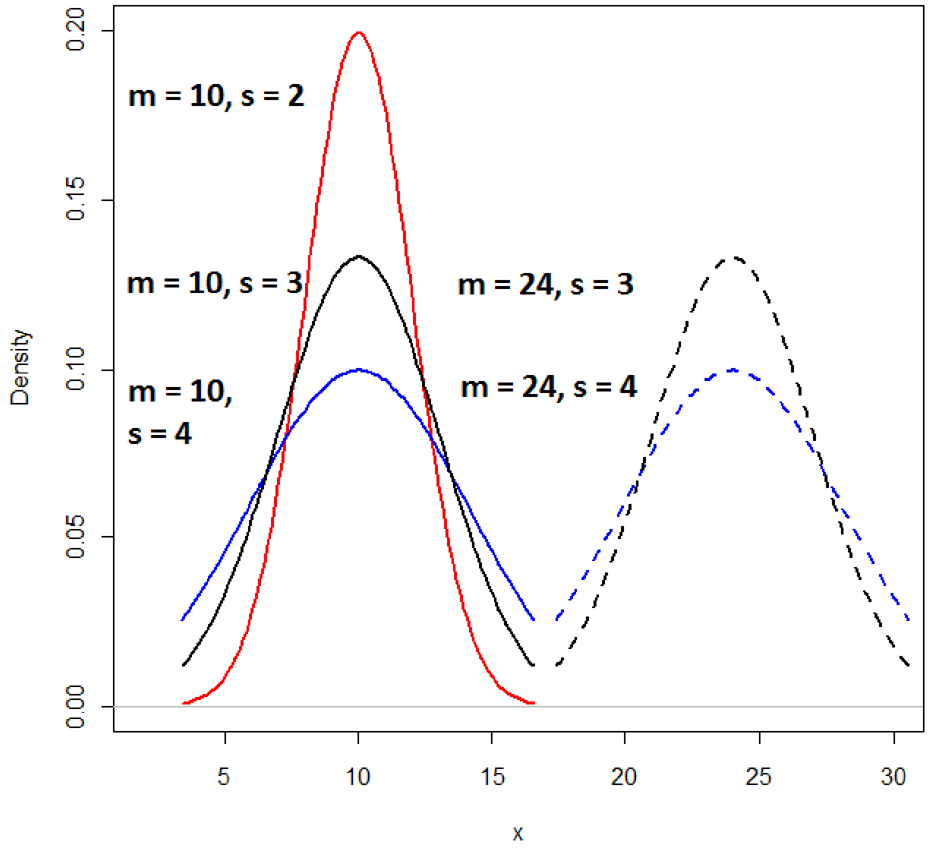


Distribuții continue: Legea Normală (lui GAUSS)

Model: **N(**m, σ [2]**)**, depinde de medie și variație/deviație standard

- glicemie, trigliceride…, etc.

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

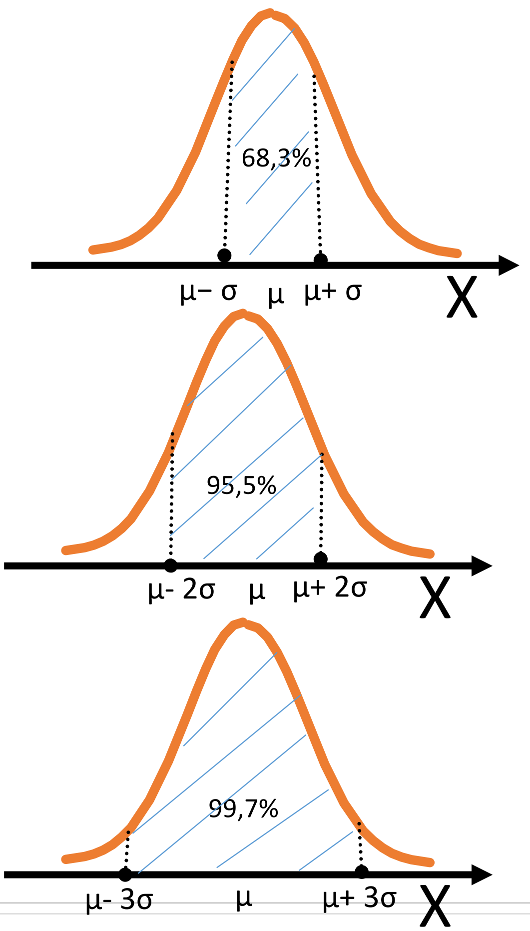


$$E(X) = \mu$$
$$V(X) = \sigma^2$$

Aplicații: intervalele de încredere pentru medii/frecvenț

Curba Normală (curba lui Gauss)

- ✓ Distribuție simetrică în jurul mediei
- ✓ Intervalul de valori **[μ –σ; μ +σ]** conține aproximativ **~ 68**,3 **%** din valorile distribuției.
- ✓ Intervalul de valori **[μ –2σ; μ +2σ]** conține aproximativ **~95**,5 **%** din valorile distribuției.
- ✓ Intervalul de valori **[μ –1,96σ; μ +1,96σ]** conține **95 %** din valorile distribuției.
- ✓ Intervalul de valori **[μ –3σ; μ +3σ]** conține aproximativ **~ 99**,7 **%** din valorile distribuției.



Statistica inferențială

Estimarea parametrilor statistici (necunoscuţi) dintr-o populație

Testarea ipotezelor statistice

Principii generale

Pentru studiul într-o populație $\mathcal{P}$ a parametrilor unei caracteristici oarecare (cantitative sau calitative) adesea este necesar să se urmeze procedeul:

1. Se extrage un **<u>eșantion reprezentativ</u>** al acestei populații.
2. Prin mijloacele statisticii descriptive se descrie distribuția caracteristicii pe eșantionul extras:
   - Calitativă: frecvența observată,
   - Cantitativă: media și variația.
3. Prin mijloacele statisticii inferențiale se extind la întreaga populație rezultatele observate pe eșantion.

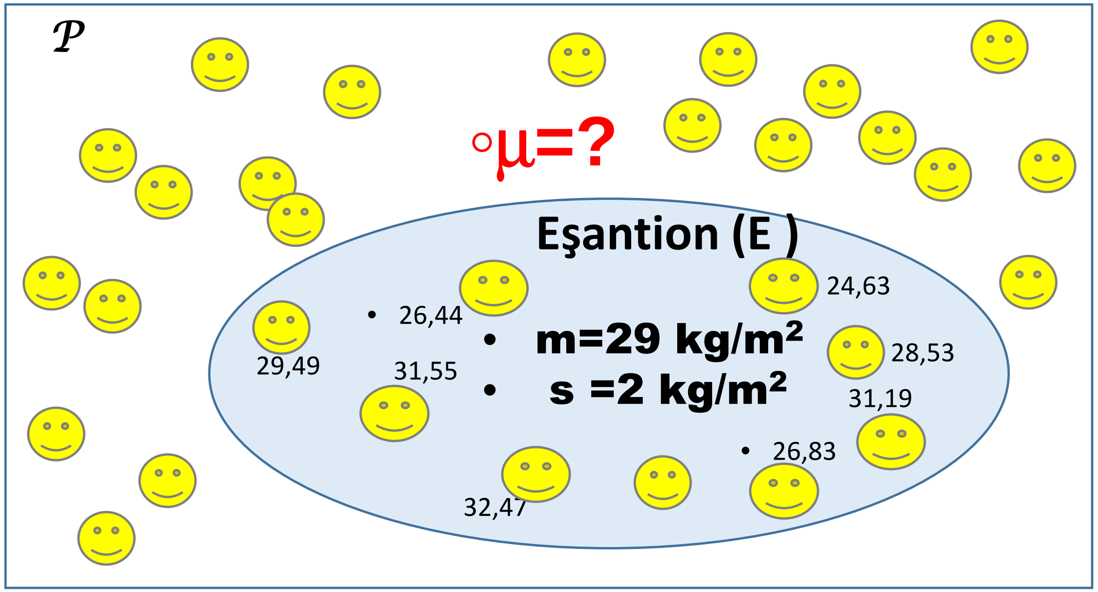
Cazul unei variabile cantitative

Media teoretică μ a variabilei X și variația sa teoretică σ^2 în populația $\mathcal{P}$ sunt necunoscute.

- Din populația P se extrage la întâmplare **eșantionul E reprezentativ.**
- În eșantionul reprezentativ E pentru variabila X se observă o medie **m** și o variație s^2.
- Se încearcă să se estimeze valorile necunoscute ale lui μ și σ^2 cu ajutorul lui m și s^2 observate.

Populația (P): media μ, σ^2 asociată unei variabile X

Variabila cantitativă: Indice de masă corporală (IMC)



Care este media Indicelui de masă corporală (IMC) la bolnavii de diabet?

Într-un eșantion de 10 bolnavi, s-au obtinut rezultatele :

28.40; 26.44; 29.49; 31.55; 31.40; 32.47; 24.63; 28.53; 31.19; 26.83

Statistica inferențială: ESTIMĂRI

estimarea unui parametru necunoscut pe o populaţie cu ajutorul statisticilor determinate pe un eșantion reprezentativ

Simbolul ^ este frecvent utilizat pentru a descrie parametrul ce se doreste a fi estimat

Statisticile determinate pe eșantion:

$$\hat{\mu} = m = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\widehat{\sigma^2} = S^2 = \frac{\sum_{i=1}^{n}(x_i - m)^2}{n-1}$$

Parametrii pe populație:

Eșantion (E) de talie n

E

X: variabila de studiu (cantitativa)

x_i-valorile lui X masurate de esantion

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

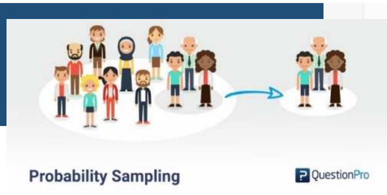
Extrapolarea (generalizarea) rezultatului la nivelul populației

Cum se obține eșantionul reprezentativ? Metode de eșantionare

- **Probabilistică**: fiecare unitate din populaţie are o şansă mai mare de zero de a fi selectată şi inclusă în eşantion
  1. Eșantionare aleatorie simplă
  2. Eșantionare sistematică
  3. Eșantionare stratificată
- **Non-probabilistică**: unele unităţi ale populaţie nu au nici o şansă de a fi selectate şi incluse în eşantion
  1. Eșantionare convenientă
  2. Eșantionare de tip bulgăre de zăpadă
  3. Eșantionare deliberată

Toți indivizii din populație au aceeași șansă de a fi selectați în eșantion.

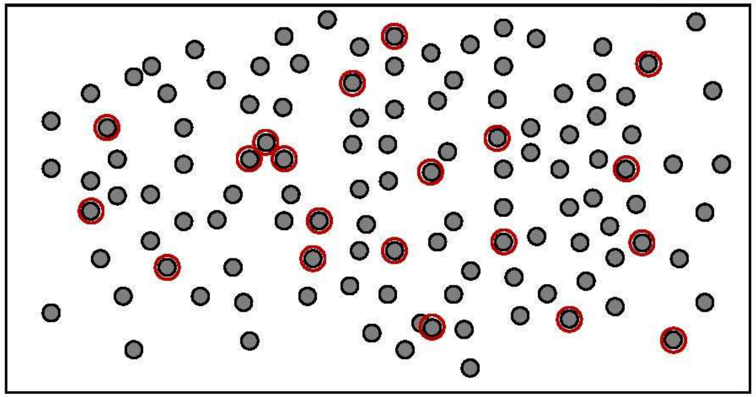
Eșantionarea probabilistică: esantionare aleatorie



Fiecărui individ i se asignează un număr și se utilizează un proces automat de selectare a numerelor pentru a fi incluse în eșantion

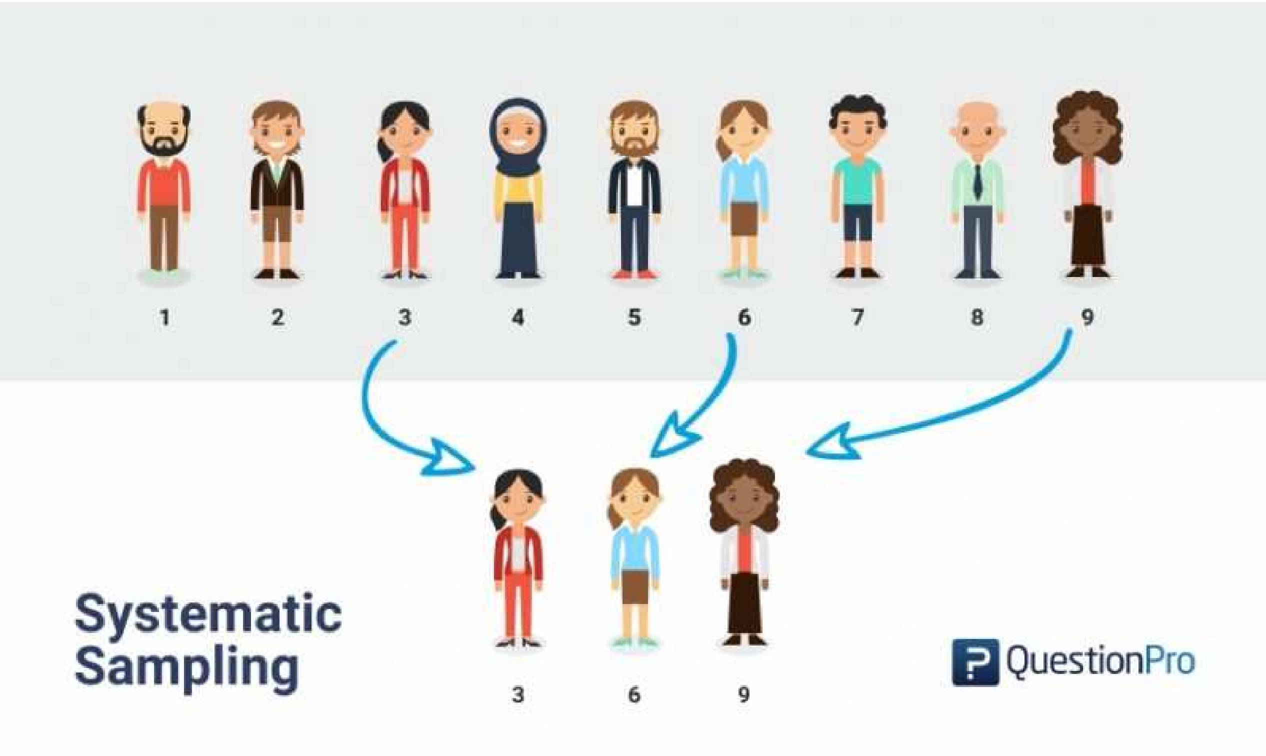
Eşantionare ALEATORIE simplă

- Subiecţi extraşi la întâmplare din populaţia statistică
- Fiecare subiect are aceiaşi şansă de a fi inclus în eşantion
  - Generarea aleatorie a unui set de numere discrete
  - Folosirea funcţiilor Excel pentru generarea numerelor randomizate (RANDBETWEEN)



Toți indivizii din populație au aceeași șansă de a fi selectați în eșantion.

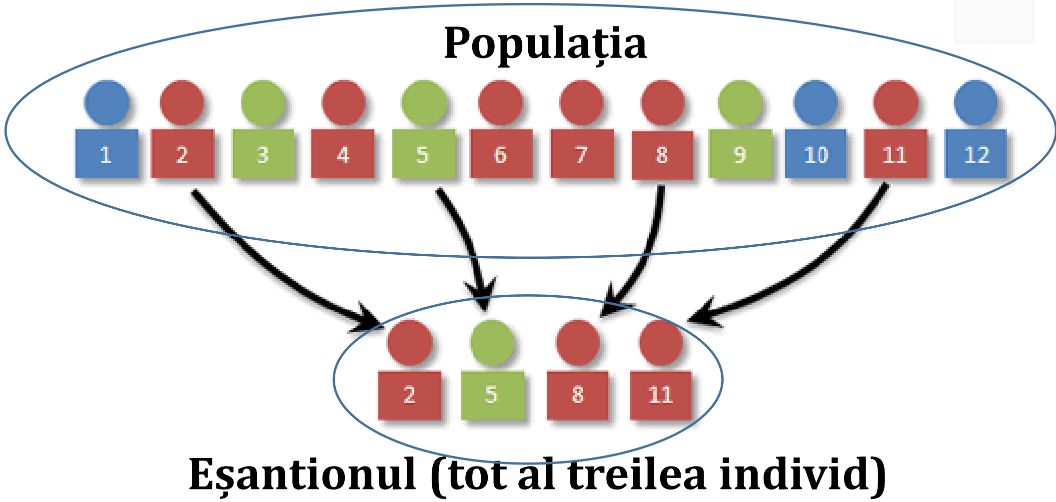
Eșantionarea probabilistică: Eșantionare sistematică



Se alege pentru a fi inclus în eșantion tot al k-lea individ.
Pentru N = 2000
Eșantionul trebuie să aibă
5%×2000 = 100 →
k= 2000/100 = 20
Tot al 20lea individ va fi inclus în eșantion

Eşantionare sistematică

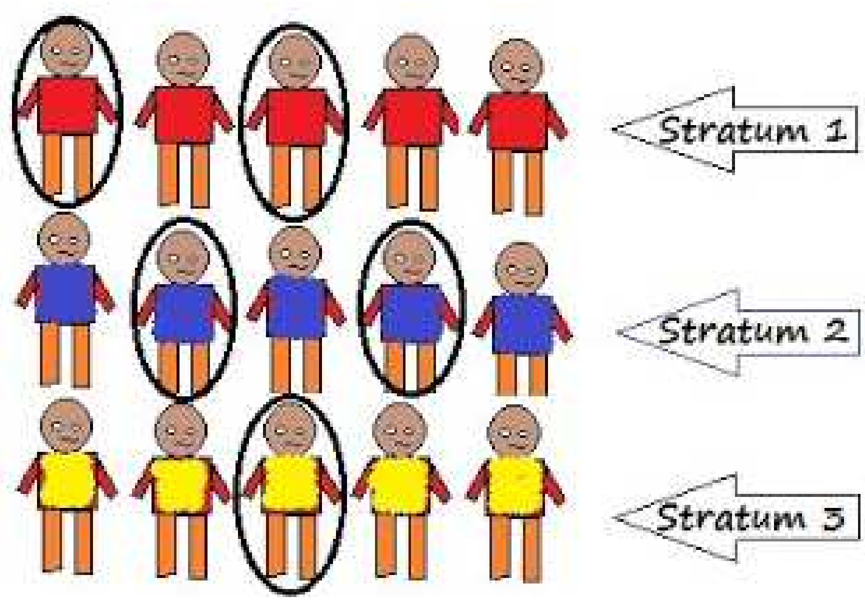
- Este selectat pentru a fi inclus în eşantion fiecare al k-lea element din structura de eşantionare
- Numărul k se obţine împărţind talia populaţiei la talia dorită a eşantionului
    - Se selectează prin randomizare punctul de start (0 - 9)



Toți indivizii din populație au aceeași șansă de a fi selectați în eșantion.

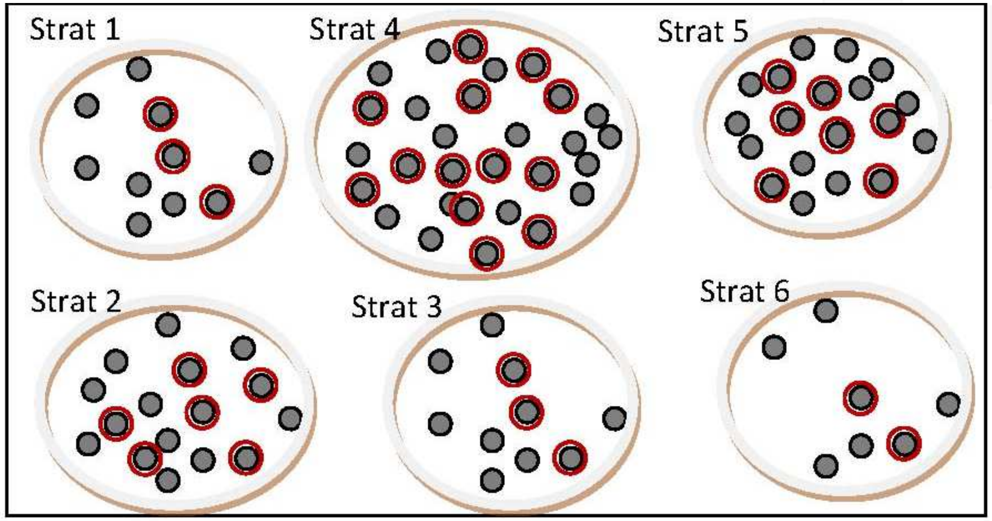
Eșantionarea probabilistică: esantionare stratificată

Populația este împărțită în grupuri sau straturi mici (ex., sex, clase de vârstă, etnie, etc.) exclusive (care nu se suprapun) → Se extrage aleatoriu din fiecare strat câte un eșantion.



Eșantionare stratificată

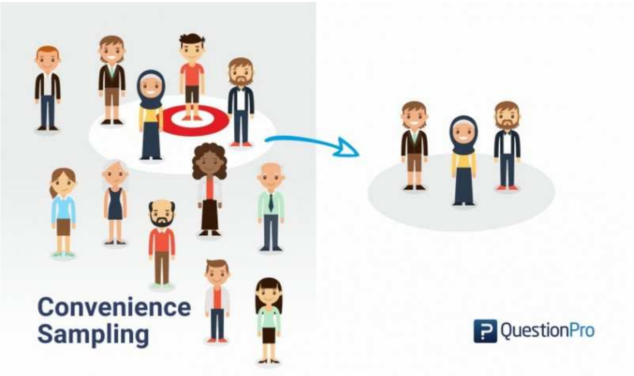
- Se împarte populaţia în mai multe subgrupe relevante numite **straturi**
- Se constituie eşantionul prin extrageri aleatoare din straturi
- Fiecare strat trebuie să fi reprezentat în eşantion în funcţie de importanţa sa în populaţie



Eșantionarea non-probabilistică

- De convenienţă:
  - Participanţii sunt selectaţi doarece sunt accessibili
    - Relativ uşor
- Bulgărele de zăpadă:
  - Subiecţii incluşi în studiu vor aduce alţi potenţiali participanţi (ex. membrii ai aceluiaşi grup, activităţi comune, etc.)
- Deliberată
  - Grup de tehnici de eşantionare care au la bază gândirea cercetătorului (ex. eșantionarea cu variație maximă, eșantionarea omogenă, eșantionarea cazurilor comune, eșantionarea cazurilor extreme, eșantionarea realizată de experți)

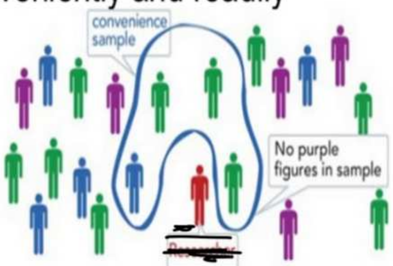
Eșantionarea non-probabilistică



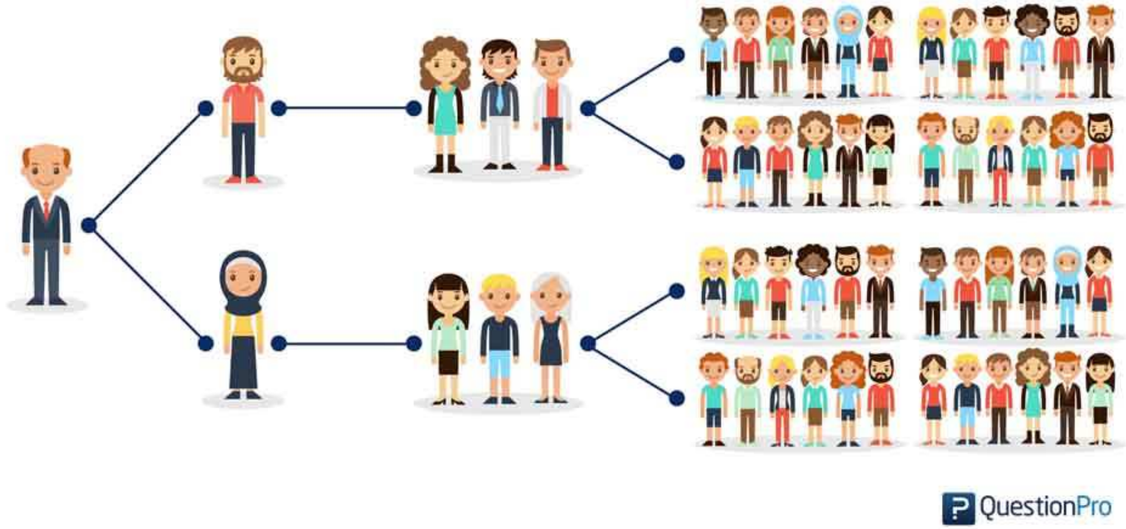
Convenientă

How?

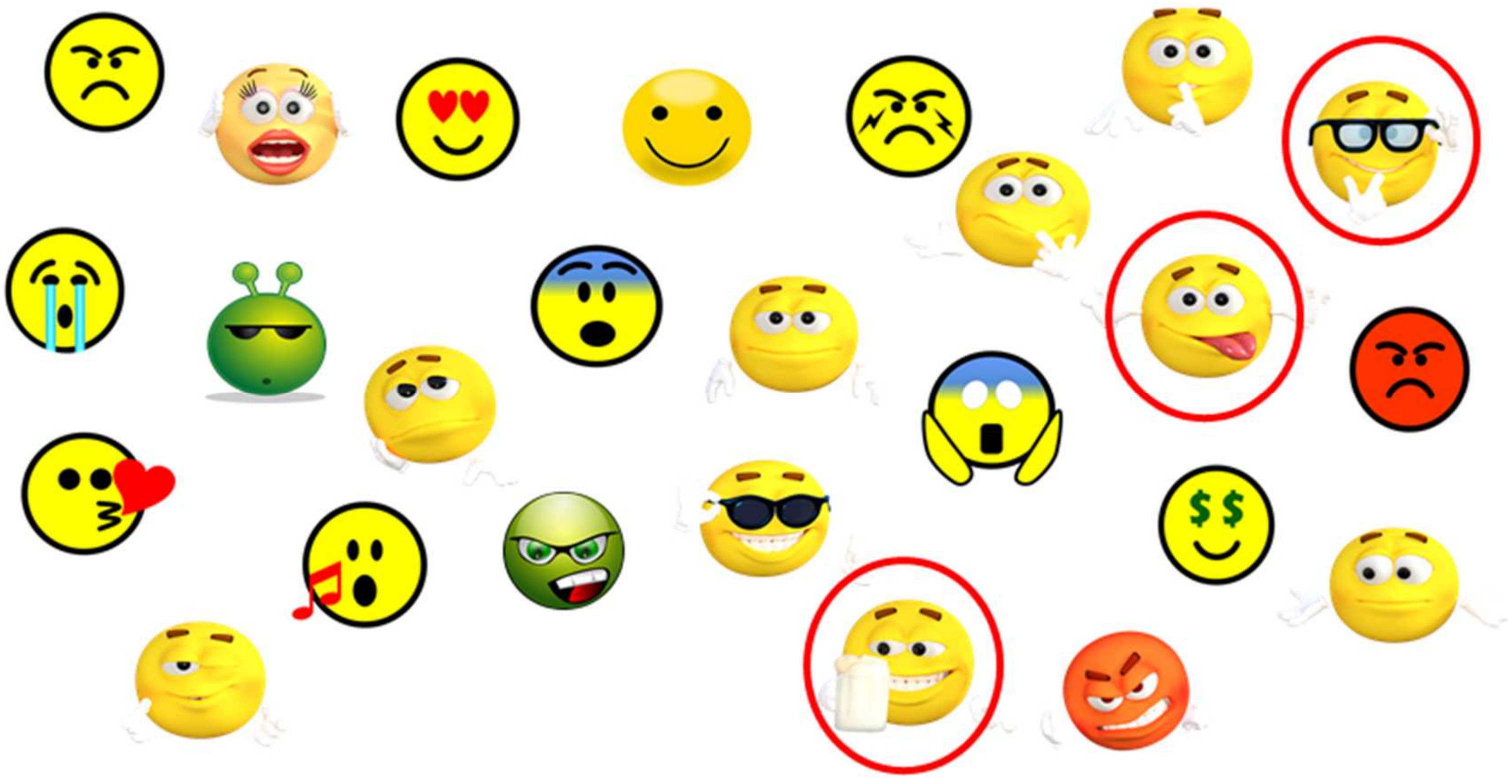
select any members of the population who are conveniently and readily available



Bulgăre de zăpadă



Eșantionarea non-probabilistică



Deliberat

appy emoji → n=3

Cazul unei variabile calitative (X)

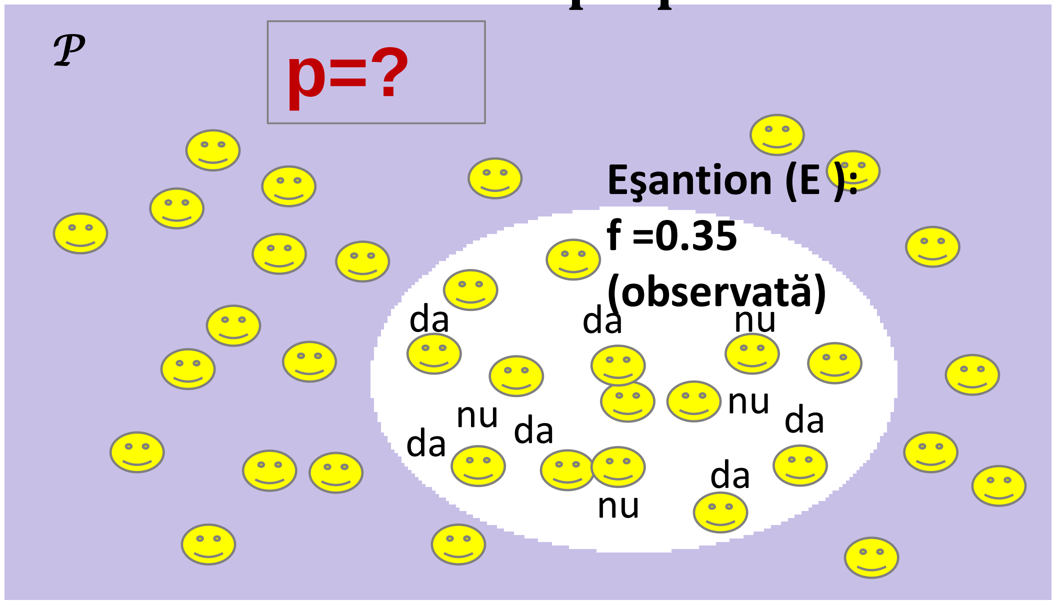
Frecvența relativă teoretică p a variabilei Y în populația $\mathcal{P}$ este necunoscută.

- Din populația P se extrage la întâmplare eşantionul E reprezentativ.
- În eşantionul reprezentativ E pentru variabila Y se observă o frecvență **f**.
- Se încearcă să se estimeze frecvența teoretică p cu ajutorul lui frecvenței observate (f).

Cazul unei variabile calitative (Y)

Populația ($\mathcal{P}$): **frecvența relativă p (necunoscută) a variabilei Y**

Variabila calitativă: Supraponderabilitate (da,nu)



Ce proporție dintre bolnavii de diabet sunt supraponderali?

Într-un eșantion de 100 diabetici, 35% au fost supraponderali.

Tipuri de Estimări

Populaţie

Eşantion

Frecvenţa relativă observată (**f**) → Frecvenţa relativă **p** pe populatie:
f - estimare punctuală
[a, b] – estimare prin interval de încredere

Media observată (**m**) → Media **μ** pe populaţie:
m – estimare punctuală
[a, b] – estimare prin interval de încredere

Tipuri de estimări: estimarea punctuală şi estimarea prin intervalul de încredere

Estimatorul punctual = o statistică a eșantionului ce poate estima fără bias valoarea parametrului pe populație.

- Media m eșantionului este estimatorul punctual al mediei populației μ.
- Frecvența relativă (f) pe eşantion este estimatorul punctual al frecvenţei relative (p) pe populație

Intervalul de încredere (engl. *confidence interval, notatie: IC*) = un interval marginit de valori [a, b] care va conţine parametrul pe populatie cu o probabilitate ridicată.

Estimatorul punctual cu bias respectiv fără bias

- Definiție: Un estimator T al unui parametru **θ** pe populatie este **estimator punctual fără bias** dacă speranța matematică M(T) = **θ** (estimarea este corectă),

Exemple:

- media **m pe eșantion**: **estimator fără bias** al mediei μ

$$\mathbf{M(m) = \mu}$$

- frecvența relativă **f**: **estimator fara bias** al frecvenței teoretice p

$$\mathbf{M(f) = p}$$

- Daca M(T) ≠ **θ** (estimator cu bias, estimarea este incorectă),

Exemplu: variația standard descriptiva (s^2): **estimator cu bias** al lui σ^2 pentru că **$E(s^2) \neq \sigma^2$ unde** $E(s^2) = \frac{n}{n-1}\sigma^2 \neq \sigma^2$

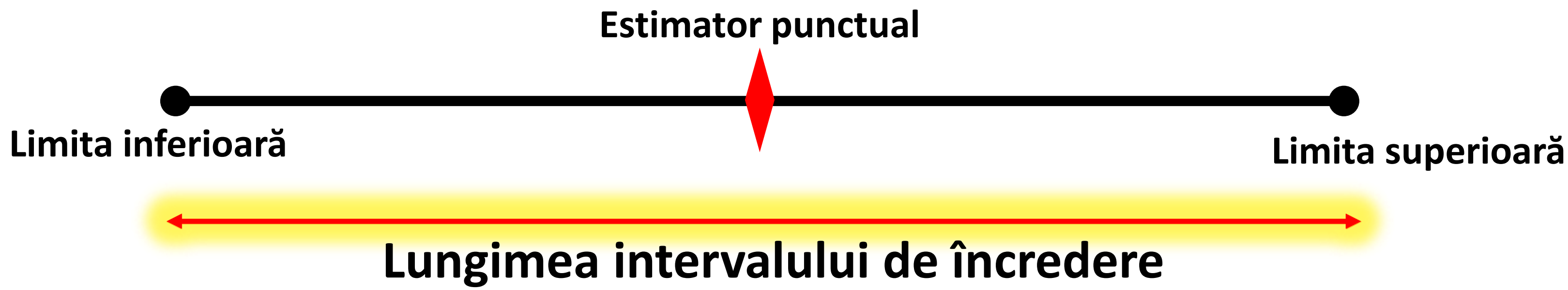
Estimatori punctuali

Parametrul pe populație ce se dorește a fi estimat	**Estimatorul punctual al parametrului**
Media (μ)	m
Variația (σ^2)	$S^2 = \frac{n}{n-1} s^2 = \frac{\sum_{i=1}^{n}(x_i - m)^2}{n-1}$
Deviatia standard (σ)	$S = \sqrt{S^2}$=deviatia standard de eșantionare
Frecvența relativă (p)	f

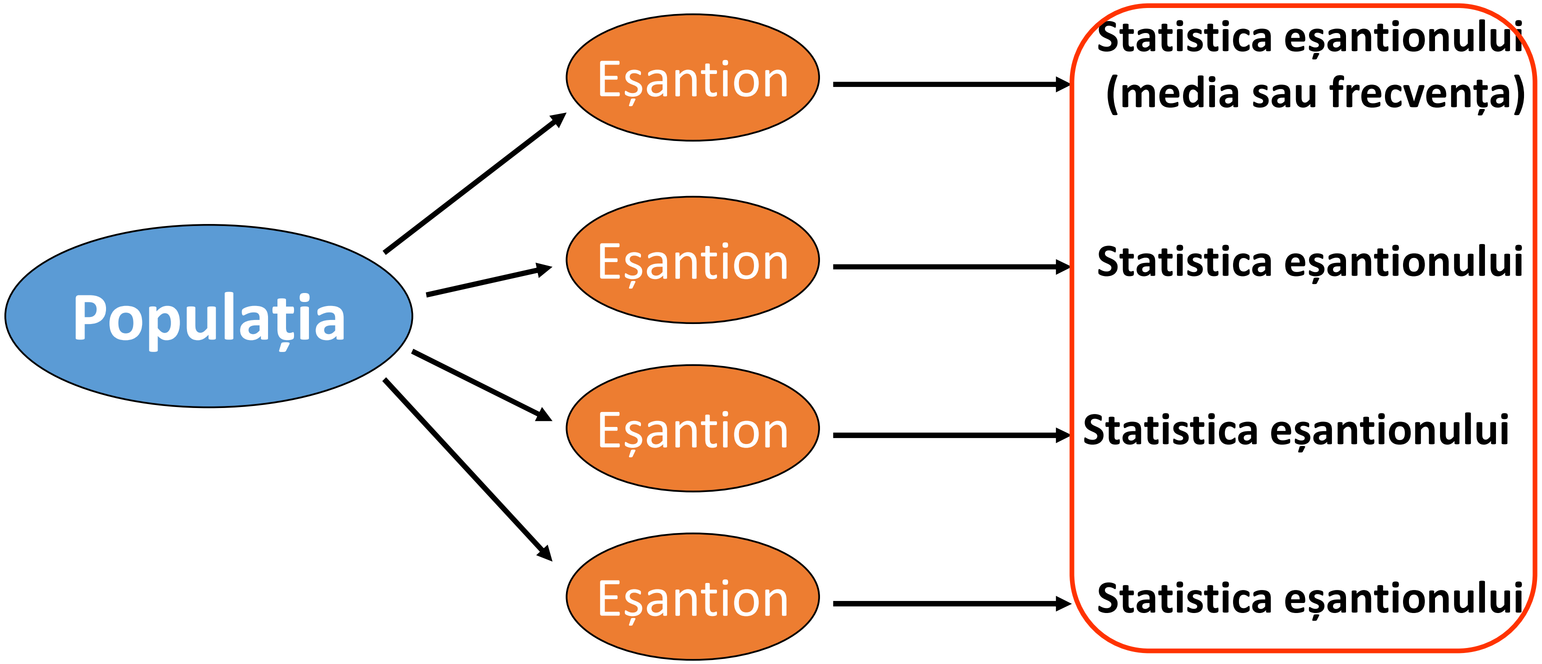
Notații: m=media calculată pe eșantion; s² =variatia descriptiva calculată pe eșantion; f=frecvența relativă observată pe eșantion

Estimatorul punctual vs. intervalul de încredere

- Estimatorul punctual = o singură valoare ce depinde de fluctuaţiile de eşantionare şi poate fi la o distanţă mai mare sau mai mică de valoarea parametrului estimat.
- Este recomandabil să se estimeze un parametru nu printr-o singură valoare ci printr-un interval (nu orice interval) ci prin
- ***<u>intervalul de încredere</u>*** despre care să se poată afirma că va contine parametrul estimat cu o probabilitate ridicată.



Distribuția de eșantionare



Distribuția de eşantionare a statisticii

Ex.: Media din eşantion≈ variabila aleatoare

->distribuţia mediilor esantioanelor se numeşte **distribuția de eşantionare a mediei**.

.

TEOREMA LIMITĂ CENTRALĂ (TLC)

- **Cazul unei variabile cantitative:**

Distribuția mediilor eșantioanelor mari ($n \geq 30$) de talie n, extrase dintr-o distribuție oarecare de medie μ și variație σ se poate aproxima printr-o distribuție normală de medie μ și deviație standard $\sigma_{\mathbf{m}} = \frac{\sigma}{\sqrt{n}}$ (numită eroare standard notată cu ES).

- **Cazul unei variabile calitative:**

Distributia frecvențelor eșantioanelor mari (*np sau n*(1-p) ≥ 10)* de talie *n*, se poate aproxima printr-o distribuție normală de medie p și deviație standard $\sigma_{\mathbf{f}} = \sqrt{\frac{pq}{n}}$ (numită eroare standard, notată cu ES).

Curba Normală (curba lui Gauss)

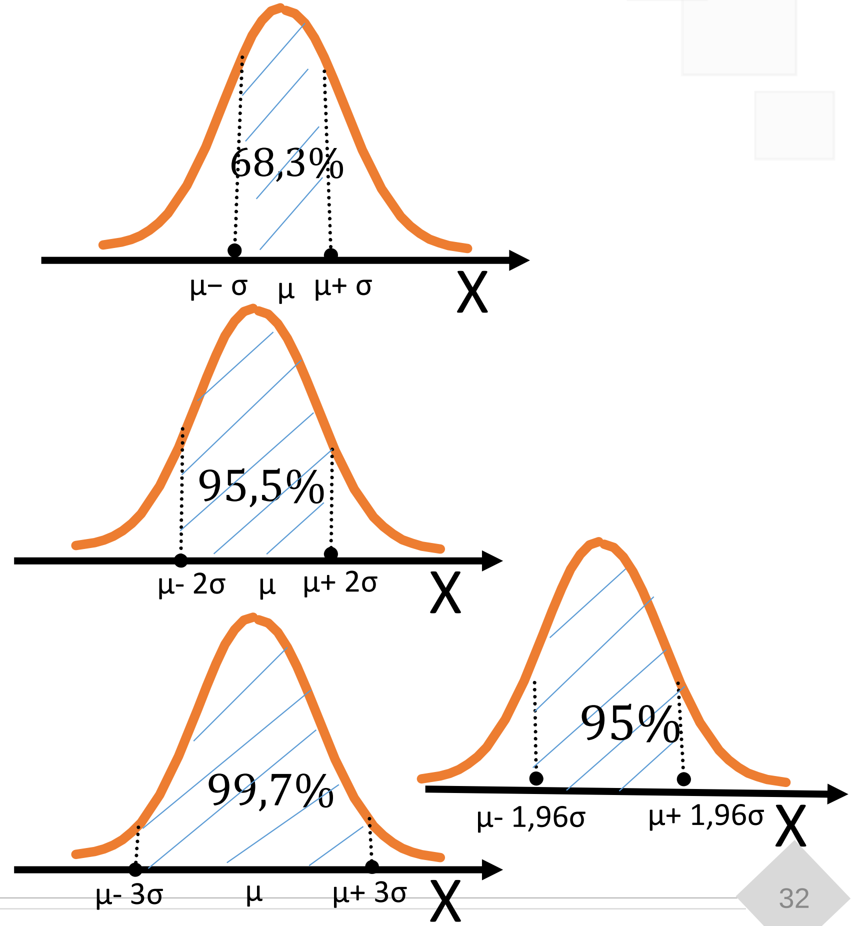
✓Distribuție simetrică în jurul mediei

✓ Intervalul de valori **[μ –σ; μ +σ]** conține aproximativ **~ 68**,3 **%** din valorile distribuției.

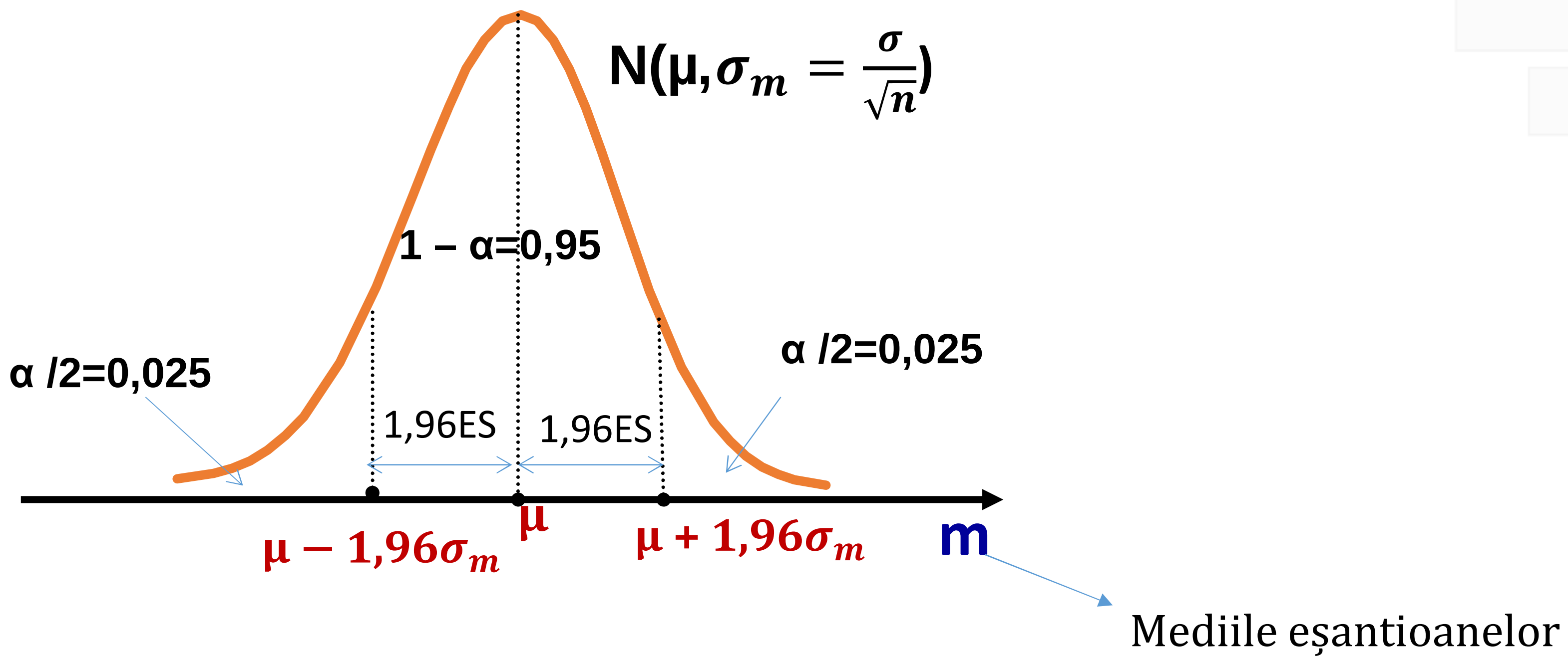
✓ Intervalul de valori **[μ –2σ; μ +2σ]** conține aproximativ **~95**,5 **%** din valorile distribuției.

✓ Intervalul de valori **[μ –1,96σ; μ +1,96σ]** conține **95 %** din valorile distribuției.

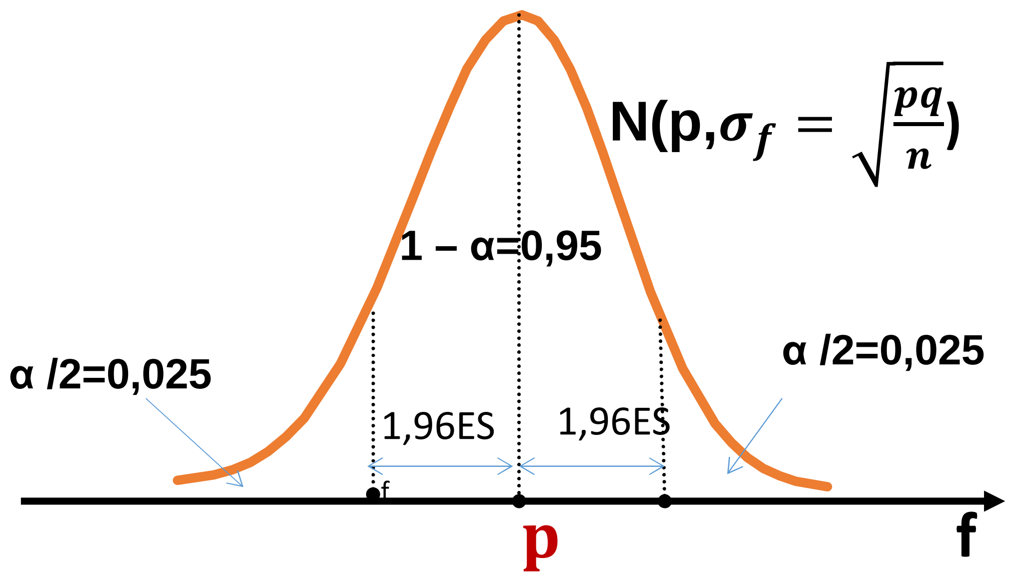
✓ Intervalul de valori **[μ –3σ; μ +3σ]** conține aproximativ **~ 99**,7 **%** din valorile distribuției.



Intervalul de încredere al mediei



Intervalul de încredere al frecvenței



Intervalul de încredere

* interval mărginit [a, b] <u>centrat</u> în estimatorul punctual T al parametrului θ.

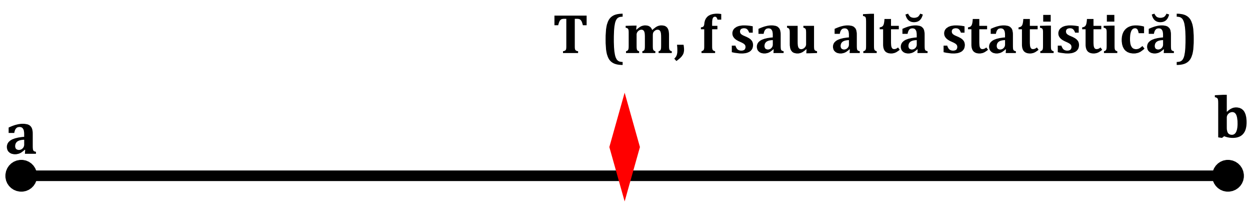
* interval care va contine parametrul θ *cu o probabilitate de* 1 – α:

$$Pr(a \leq \theta \leq b) = 1 - \alpha$$

* **<u>Nivel de încredere (1- α)</u>** fixat *<u>a priori</u>*: de obicei egal cu 0,95 (dar poate fi si 0,90; 0,99).

* Nivel de semnificație: α=0,05

* Dacă α = 0,05 => **95% IC: [a; b]**



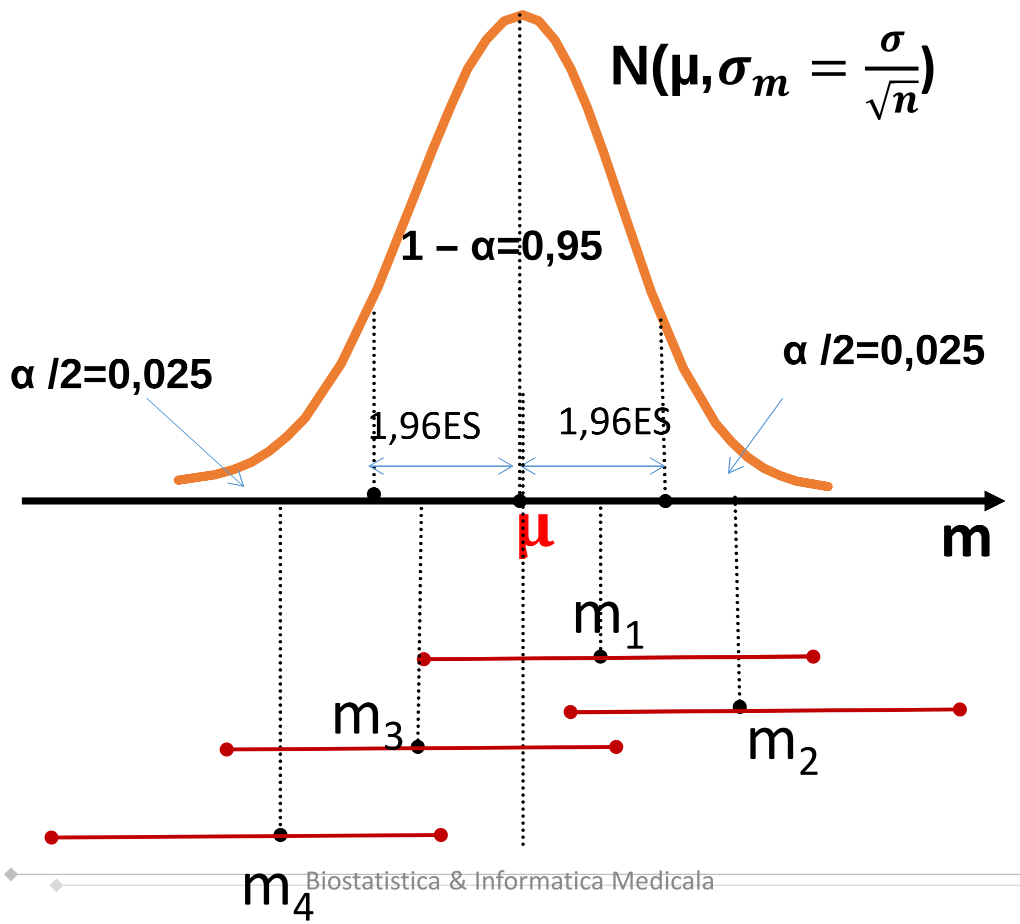
Intervalul de încredere

Lungimea intervalului de încredere:

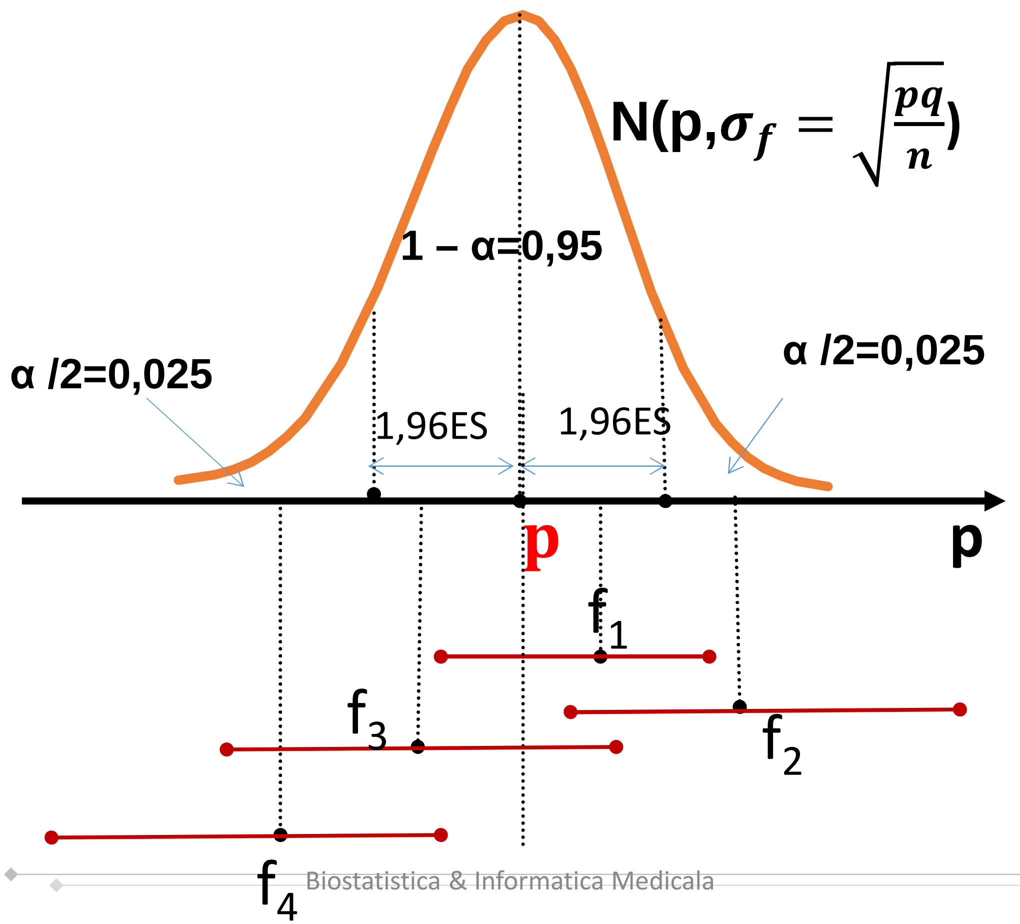
- este cu atât mai mică cu cât volumul eșantionului e mai mare.
- variază în functie de nivelul de semnificație (α)

- **Nu** poate fi estimată niciodată cu încredere de 100%.
- **Interpretare probabilistică:**
  - Dacă toate eșantioanele posibile de volum *n* s-ar extrage din populație și mediile și intervalele de încredere asociate ar fi calculate, 95% din intervalele de încredere conțin parametrul pe populație.
  - **Interpretare practică:** Suntem 95% siguri ca intervalul de încredere [a, b] conține parametrul (media, frecvența,...) pe populație.

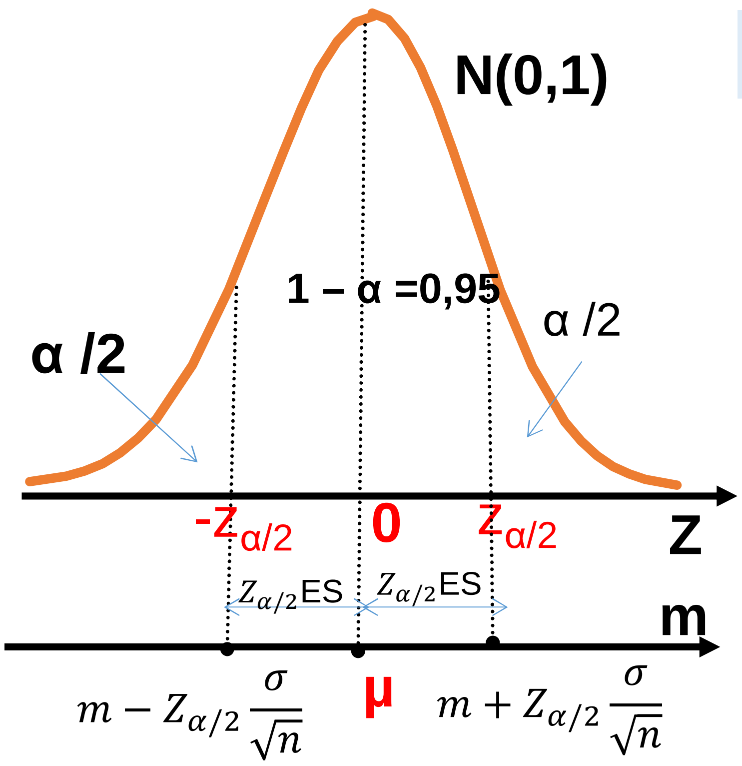
Intervalul de încredere al mediei



Intervalul de încredere al frecvenței



ESTIMAREA MEDIEI PE POPULAȚIE
CAZUL EȘANTIOANELOR MARI (n≥30)



$$\Pr(a \leq \mu \leq b) = 1 - \alpha = 0{,}95$$

$$-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}$$

$$-Z_{\alpha/2} \leq \frac{\mathrm{m} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq Z_{\alpha/2}$$

$$\underbrace{m - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}_{a} \leq \mu \leq \underbrace{m + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}_{b}$$

ESTIMAREA MEDIEI PE POPULAŢIE CAZUL EŞANTIOANELOR MARI (n≥30)

- Cel mai frecvent se utilizează un prag de semnificație $\alpha = 0{,}05$. Atunci $Z_{\alpha/2}=1{,}96$ și deci intervalul de încredere este:

Deviația standard (σ) pe populație este cunoscută?

DA

95% IC:

$$\left[m - 1{,}96\frac{\sigma}{\sqrt{n}}, m + 1{,}96\frac{\sigma}{\sqrt{n}}\right]$$

NU

$$\sigma \approx S = \sqrt{\frac{n}{n-1}}s$$

95% IC:

$$\left[m - 1{,}96\frac{s}{\sqrt{n-1}}, m + 1{,}96\frac{s}{\sqrt{n-1}}\right]$$

$$\left[m - 1{,}96\frac{S}{\sqrt{n}}, m + 1{,}96\frac{S}{\sqrt{n}}\right]$$

ESTIMAREA MEDIEI PE POPULAŢIE
CAZUL EŞANTIOANELOR MARI ($n \geq 30$)
EXEMPLU DE CALCUL

- S-a extras din populația femeilor din Ro cu HTA un eșantion aleatoriu de talie 49. Media colesterolului a fost egală cu 220 mg/dl, iar deviația standard de eșantionare S =18 mg/dl.
- Care este estimarea mediei colesterolului la femeile din Ro cu HTA folosind intervalul de încredere de 95%?
- Se știe că m= 220 mg/dl este estimatorul punctual al parametrului necunoscut (media colesterolului) pe populație

$$95\%\ IC: \left[220 - 1,96\frac{18}{\sqrt{49}}; 220 + 1,96\frac{18}{\sqrt{49}}\right] = [215; 225]$$

Estimarea unui parametru (medie, frecvența) pe populație

Interval de încredere pentru

Dacă n*f ≥10 and n*(1-f) ≥10

Media **μ** pe populație

Frecvența **p** pe populație

$$\left[f - Z_{\alpha/2}\sqrt{\frac{f \times (1-f)}{n}}, f + Z_{\alpha/2}\sqrt{\frac{f \times (1-f)}{n}}\right]$$

σ cunoscut și n≥30

$$\left[m - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, m + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

σ necunoscut și n≥30

$$\left[m - Z_{\alpha/2}\frac{S}{\sqrt{n}}, m + Z_{\alpha/2}\frac{S}{\sqrt{n}}\right]$$

$$\left[m - Z_{\alpha/2}\frac{s}{\sqrt{n-1}}, m + Z_{\alpha/2}\frac{s}{\sqrt{n-1}}\right]$$

σ necunoscut și n<30

$$\left[m - t_{n-1;\alpha/2}\frac{s}{\sqrt{n-1}}, m + t_{n-1;\alpha/2}\frac{s}{\sqrt{n-1}}\right]$$

distribuția Student (t)-nu face obiectul acestui curs

Pentru 1-α =0,95 ⇒ $Z_{\alpha/2}$ = 1,96

ESTIMAREA FRECVENTEI PE POPULATIE CAZUL EȘANTIOANELOR MARI($np \geq 10$ sau $n(1-p) \geq 10$)

100 (1-α)% IC: $$[f - Z_{\alpha/2}\sqrt{\frac{f \times (1-f)}{n}}; f + Z_{\alpha/2}\sqrt{\frac{f \times (1-f)}{n}}]$$

95% IC: $$[f - 1{,}96\sqrt{\frac{f \times (1-f)}{n}}; f + 1{,}96\sqrt{\frac{f \times (1-f)}{n}}]$$

- f este frecvența relativă observată pe eșantion

ESTIMAREA FRECVENTEI PE POPULATIE CAZUL EȘANTIOANELOR MARI(np≥10 sau n(1-p)≥10 EXEMPLU DE CALCUL:

Într-un eșantion aleator de 100 diabetici s-a observat ca 35% au fost supraponderali.
Care este estimarea prevalentei supraponderabilității pe populație folosind intervalul de încredere de 95%?

Soluție: f=0,35
Verificarea condiției: np=100*0,35=35≥10

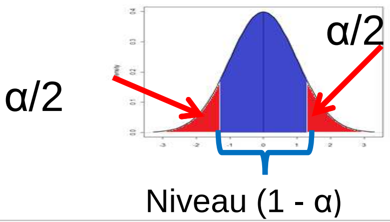
95% IC: $$[f - 1{,}96\sqrt{\frac{f \times (1-f)}{n}}; f + 1{,}96\sqrt{\frac{f \times (1-f)}{n}}]$$

95% IC: [0,26;0,44]

Factori care influențează lărgimea/precizia unui interval de încredere (IC)

- Lărgimea unui interval de încredere (IC) este determinată de: **n, σ** și **α**:
- DACĂ:
  * **« n »** crește, lărgimea lui IC scade
  * **Deviația standard** (**σ**): variabilitatea crește ⇒ lărgimea crește (relatie de directă proportionalitate),
  * **Nivelul de încredere** crește (deci **α** diminuează) ⇒ lărgimea crește
  * Riscul **α** crește ⇒ lărgimea scade

- Lărgimea unui interval de încredere (IC) cuantifică **precizia estimării**:
  - Dacă IC est larg ⇒ estimare imprecisă/slabă
  - Dacă IC est îngust ⇒ estimare precisă



Exemple de probleme

E1. * Dorim sa calculam intervalul de incredere al mediei circumferintei abdominale intr-o populatie cunoscand datele pe un esantion de 26 subiecti. Valoarea critica a cuantilei de interes este 2,06. Media circumferintei abdominale este 63 cm si deviatia standard a esantionului este 7 cm.

Care este intervalul de încredere de 95% pentru circumferința abdominală pe populație?

A. [60.12; 65.88]
B. [60.26; 65.74]
C. [42.41; 83.59]
D. [60.06; 65.94]
E. [60.17; 65.83]

R1.A

Exemple de probleme

E2. * Dorim sa calculam intervalul de incredere al mediei circumferintei abdominale intr-o populatie cunoscand datele pe un esantion de 16 subiecti. Valoarea critica a cuantilei de interes este 2,13. Media circumferintei abdominale este 63 cm si deviatia standard de esantionare este 7 cm.

Care este intervalul de incredere de 95% pentru circumferinta abdominala pe populatie?

A. [60.12; 65.88]
B. [59.27; 66.73]
C. [42.41; 83.59]
D. [60.06; 65.94]
E. [60.17; 65.83]

R2.B

Exemple de probleme

E3. * Dorim sa calculam intervalul de incredere al frecventei gripei intr-o populatie cunoscand datele pe un esantion de 100 pacienti. In esantion, 10 dintre subiecti au avut gripa.

Care este intervalul de incredere de 95% pentru prevalența gripei ?

A. [0.04; 0.16]

B. [-0.01; 0.21]

C. [-0.01; 0.21]

D. [0; 0.2]

E. [0.07; 0.13]

R3. A

MULȚUMESC PENTRU ATENȚIE!