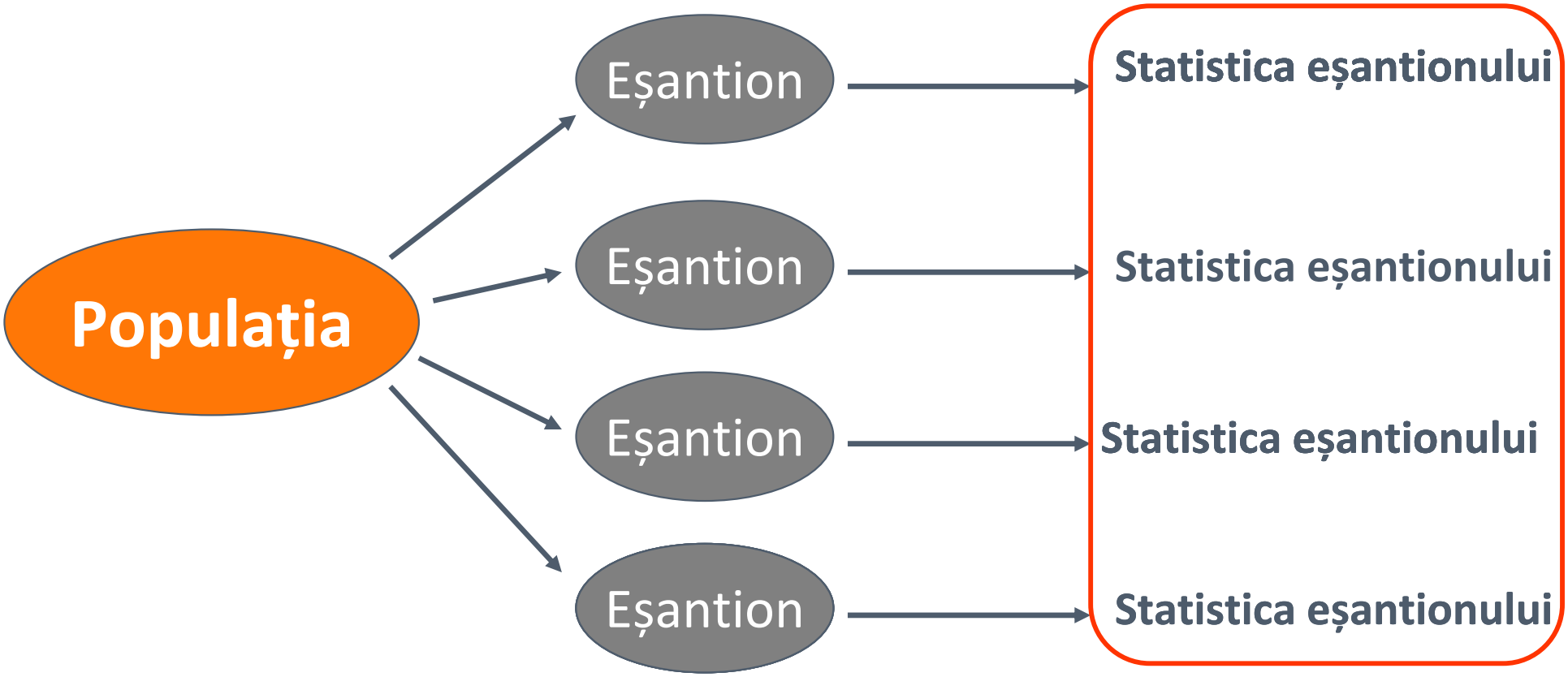
Estimarea parametrilor statistici

Sorana D. BOLBOACĂ

- Estimatorul punctual
- Intervalul de încredere
  - Medie
  - Proporție

Cuprins



Distribuția eșantionului ≠ **Distribuția de eșantionare**

N = volumul populației

Copii de 1 an din Ro

AB: $x_{AB,1}, x_{AB,2}, \dots x_{AB,1000}$

...

GJ: $x_{GJ,1}, xGJ_{,2}, \dots x_{GJ,1000}$

...

VN: $x_{VN,1}, xVN_{,2}, \dots x_{VN,1000}$

$\bar{x}_{AB}$

...

$\bar{x}_{GJ}$

...

$\bar{x}_{VN}$

$$\mu = \frac{x_{AB} + \dots + x_{GJ} + \dots + x_{VN}}{N}$$

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

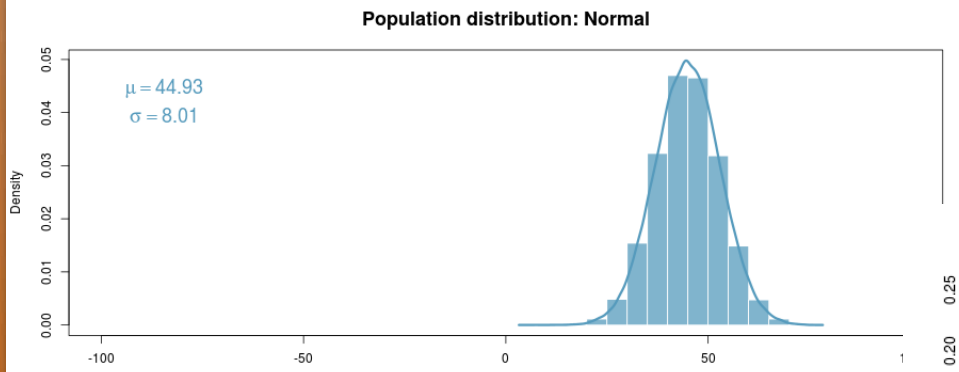
Distribuția de eșantionare

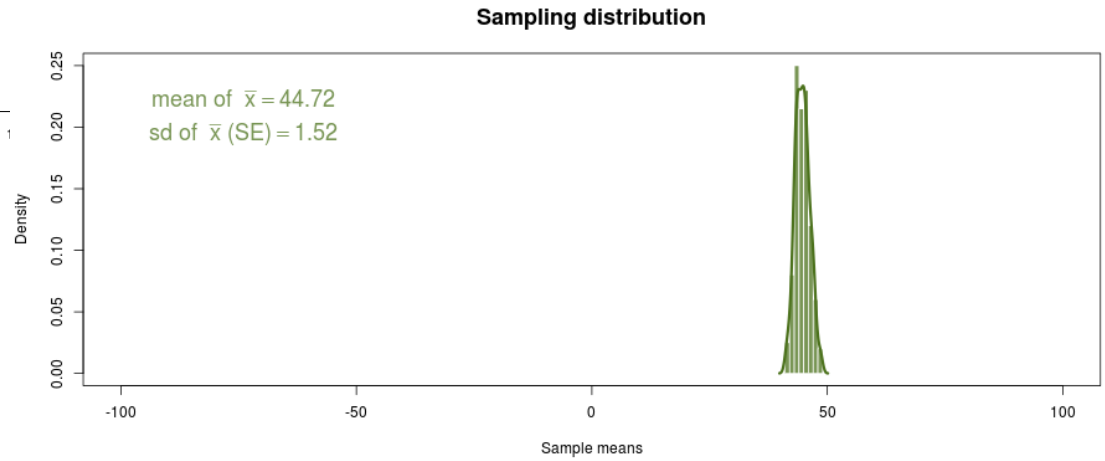
$$\text{media } (\bar{x}) \approx \mu$$

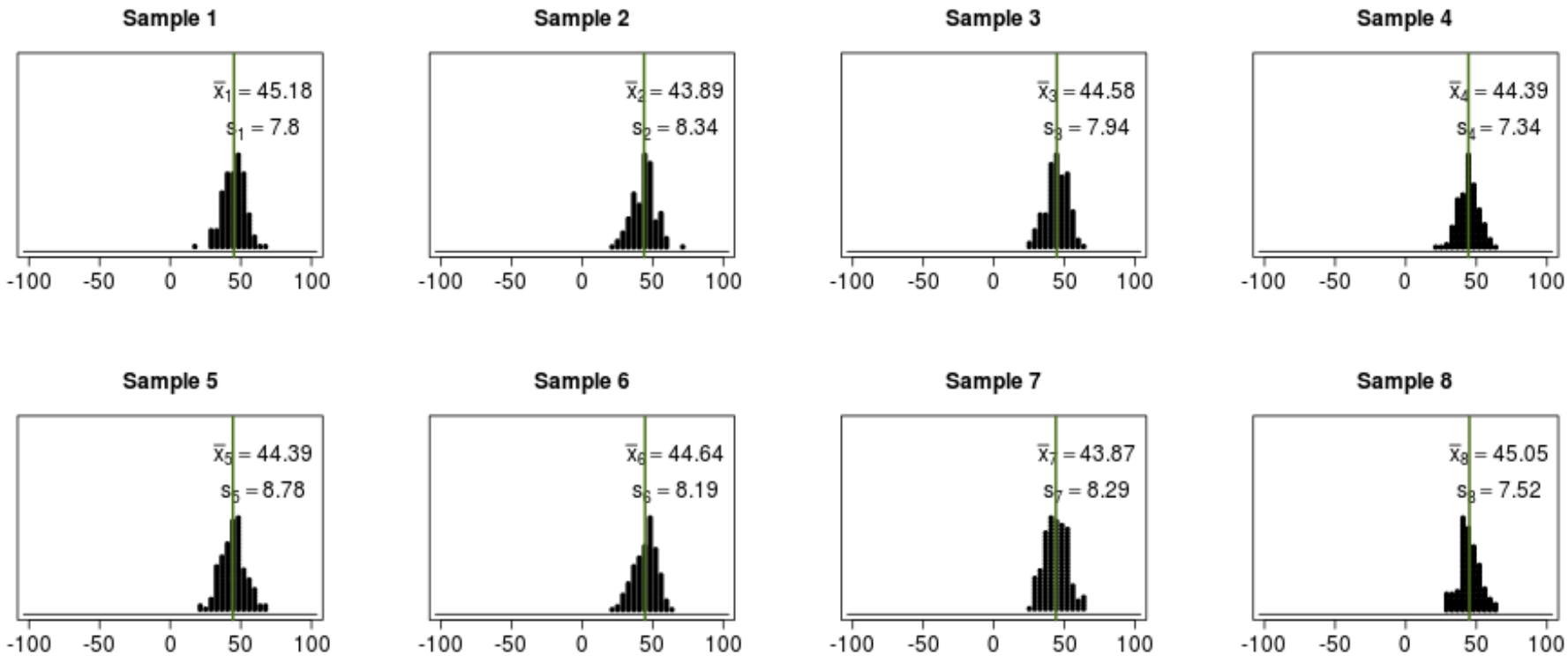
$$n \uparrow \Rightarrow \downarrow s < \sigma$$

n=30

200 eşantioane







» Distribuția statisticii eşantionului este aproape normală, cu media aproape egală cu cea a populației şi cu deviația standard egală cu eroarea standard (deviația standard a populației împărțită la radical din volumul eşantionului).

$$\bar{x} \sim \underbrace{\boxed{N}}_{\text{forma}}\left(\text{mean} = \underbrace{\boxed{\mu}}_{\text{centralitatea}}, SE = \underbrace{\boxed{\frac{\sigma}{\sqrt{n}}}}_{\text{dispersia}}\right)$$

» Condiții:

» Independența: eşantioanele trebuie să fie independente (eşantion randomizat / asignare randomizată). – în caz de eşantionare fără înlocuire, n < 10%×N.

» Volumul eşantionului/asimetrie: populația e normal distribuită / dacă e distribuția e asimetrică volumul eşantionului e mare (n > 30)

Teorema limită centrală

- Distribuția normală: De ce o folosim?
  - Multe variabile biologice urmează o distribuție normală
  - Distribuția normală este bine înțeleasă din punct de vedere matematic
- Estimarea punctuală
  - O valoare a parametrului teoretic estimat
    - m (media eşantionului) este un estimator punctual al mediei populației (μ)
  - Este influențată de fluctuațiile de eşantionare
  - Poate să fie foarte departe de valoarea reală a parametrului estimat

» Se recomandă ca estimarea unui parametru teoretic să se realizeze prin intermediul unui interval nu a unei singure valori

- Acest interval se numeşte interval de confidență/încredere
- Parametrul estimat aparține cu o probabilitate mare intervalului de confidență

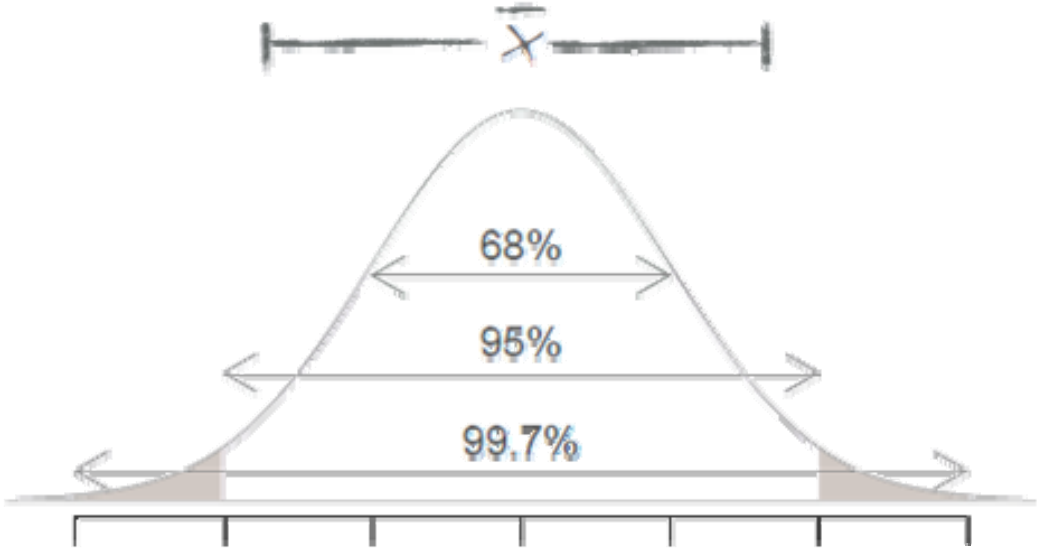
» Intervalul de confidență/încredere = un interval plauzibil de valori asociat unui parametru al populației

- Dacă raportăm un estimator punctual, cel mai probabil acesta nu va fi egal cu parametrul populației
- Dacă raportăm un interval avem o şansă ca acest interval să cuprindă valoarea parametrului populației

Intervalul de încredere - De ce?

Teorema limită centrală

$$\bar{x} \sim N\left(\text{mean} = \mu,\ SE = \frac{\sigma}{\sqrt{n}}\right)$$



Marja de eroare (ME)

~IC95% (medie): $\bar{x} \pm 1{,}96 \times SE$

Intervalul de încredere

Definiție

» Un şir de valori al unui estimator de interes calculat astfel încât pentru o probabilitate de eroare aleasă să includă valorile adevărate ale variabilei.

» **P[valoarea critică inferioară < estimatorul < valoarea critică superioară] = 1-α**

- unde α = nivelul de semnificație

» Intervalul definit de valorile critice va cuprinde estimatorul populației cu o probabilitate de 1-α

» Se aplică în cazul variabilelor distribuite normal!

» Unul din primele exemple de asimetrie comportamentală a omului este preferința de a întoarce capul spre dreapta nu spre stânga. Un studiu realizat pe un eşantion de 124 cupluri a pus în evidență că 64,5% din acestea întorc capul spre dreapta când se sărută. Eroarea standard asociată acestui estimator este egală cu aproximativ 4%. Care din următoarele sunt false?

1. Un volum de eşantion mai mare va determina o eroare standard mai mică.
2. Marja de eroare pentru un IC de 95% asociată procentului de cupluri care întorc capul la dreapta când se sărută e aproximativ 8%.
3. IC95% pentru procentul de cupluri care întorc capul spre dreapta când se sărută este 64,5%±4%
4. IC99,7% pentru procentul de cupluri care întorc capul spre dreapta când se sărută este 64,5%±12%

Intervalul de încredere pentru medie

» Eroarea standard a mediei este egală cu deviația standard împărțită la radicalul volumului eşantionului

- Dacă deviația standard este mare, şansa de eroare în estimator este mare
- Dacă volumul eşantionului este mare, şansa erorii în estimator este mică.

$$\left[\bar{x} - Z_{\alpha}\frac{s}{\sqrt{n}}, \bar{x} + Z_{\alpha}\frac{s}{\sqrt{n}}\right]$$

Valoare critică

$$\bar{x} \pm Z_{\alpha} \frac{s}{\sqrt{n}}$$

Z = valoare constantă pentru un prag de semnificație dat

Condiții:

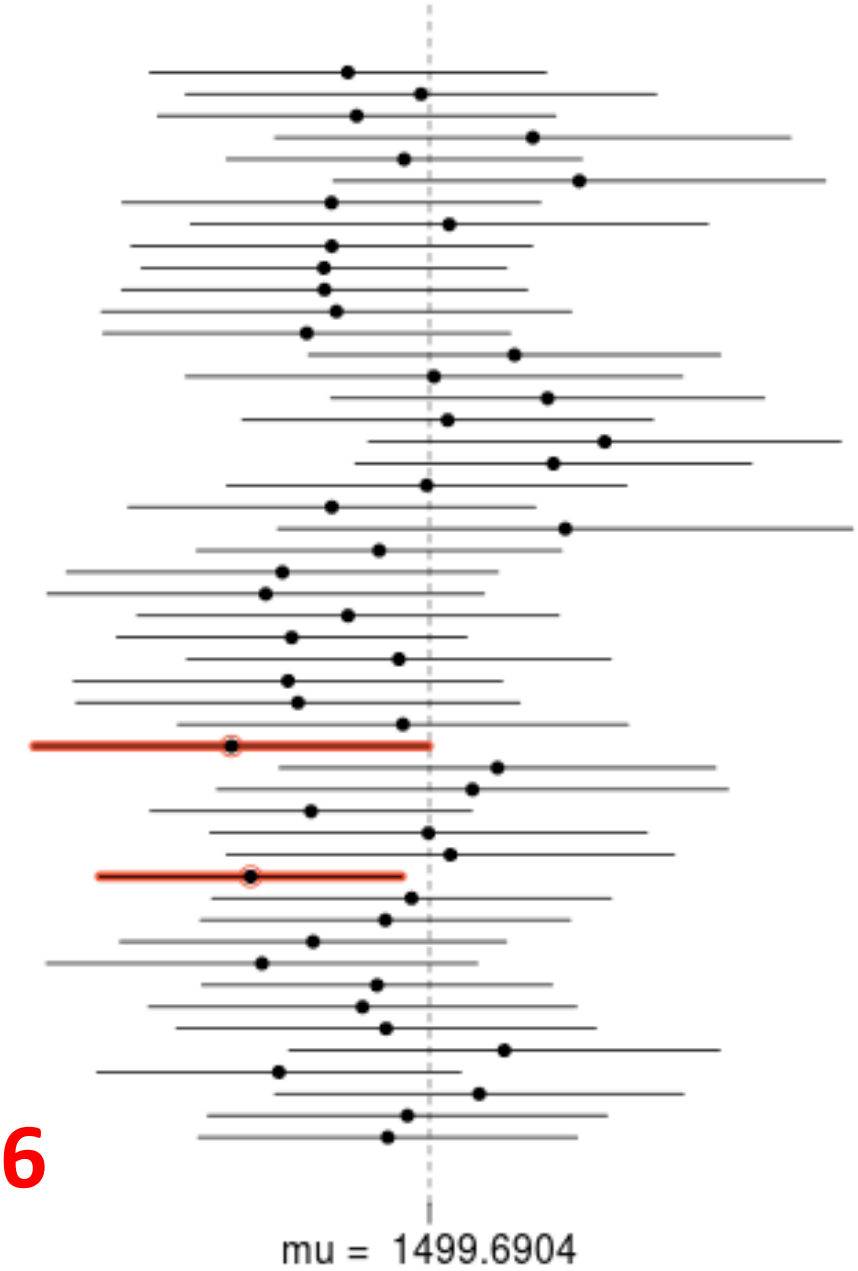
» Independența:

- Eşantioanele observate trebuie să fie independente
- Eşantion randomizat (fără înlocuire n<10%×N) sau asignare randomizată

» Asimetria distribuției: n ≥ 30, sau mai mare pentru distribuții asimetrice

Intervalul de încredere pentru medie

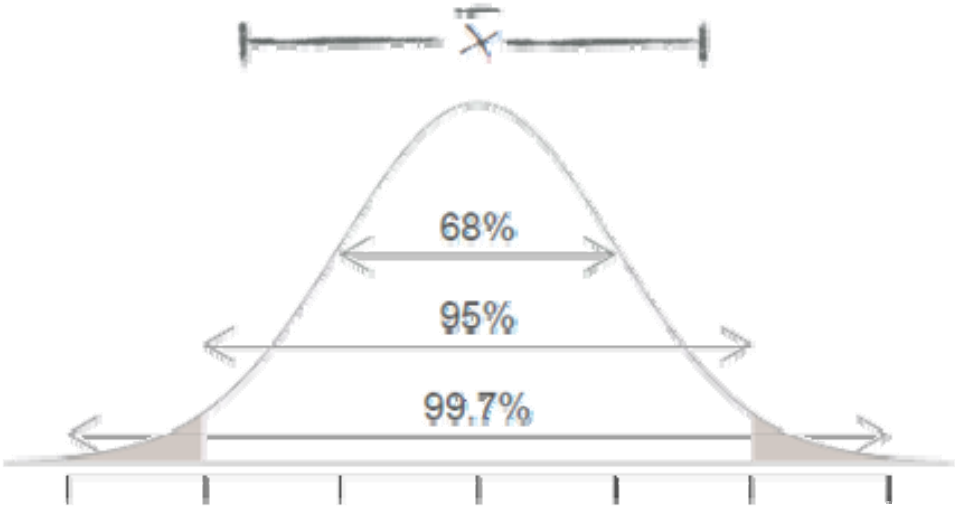
- Luăm mai multe eşantioane şi construi intervalele de confidență de 95%
- ~ 95% din aceste intervale vor conține media adevărată a populației (μ)
- Intervale de confident frecvent utilizate: 95% 98% şi 99%.

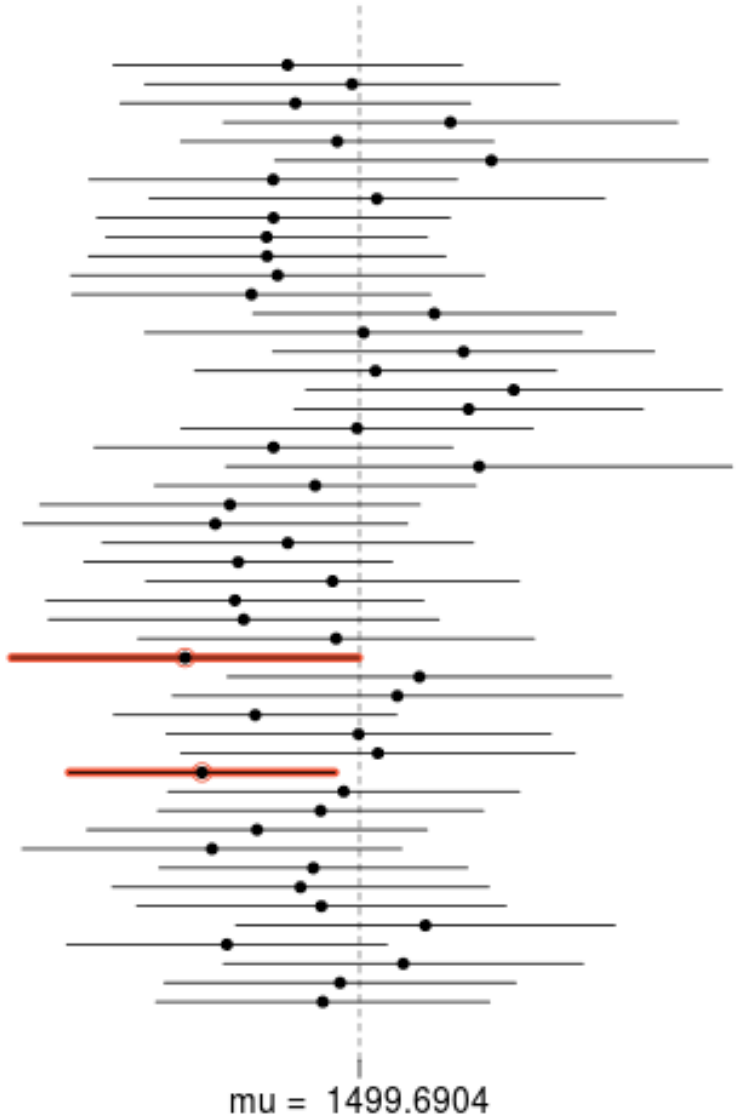


48/50 = 0,96

Acurateţe vs precizie

» Dacă dorim să fim siguri că media populației este cupri IC acesta trebuie să fie **<u>larg</u>** îngust?





Acurateț e vs precizie

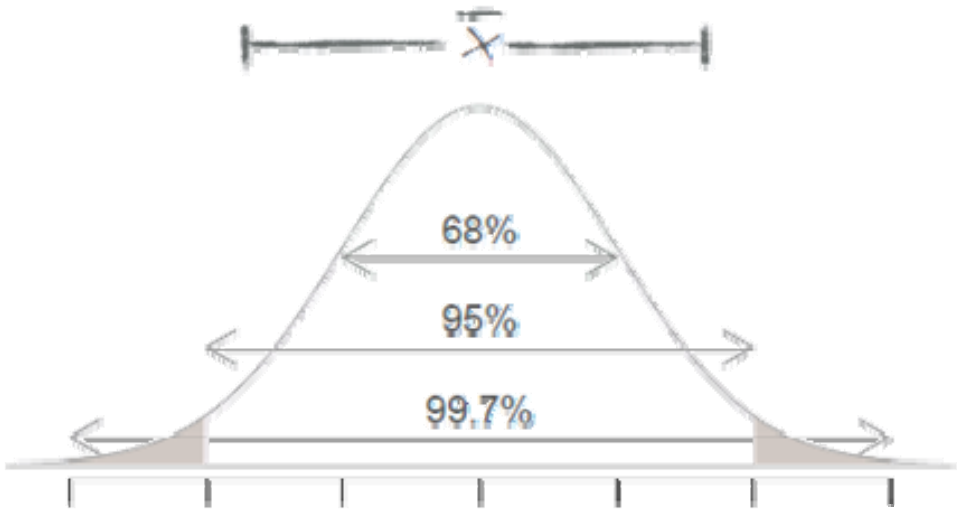
» Dacă dorim să fim siguri că media populației este cuprinsă în IC acesta trebuie să fie **larg** sau îngust?

- 99% (↑): lărgimea (↑) + acuratețe (↑) + precizie (↓)

» Cum putem avea precizie mare şi acuratețe mare?

- n↑

$$ME = Z_{\alpha} \frac{s}{\sqrt{n}} \rightarrow n = \left(\frac{Z_{\alpha} \times s}{ME} \right)^2$$



Acuratețe vs precizie

- Se doreşte testarea efectului unui medicament folosit în tratamentul epilepsiei la mamă asupra dezvoltării cognitive a copilului. Dezvoltarea cognitivă se testează prin estimarea indicelui de inteligență a copilului de 3 ani născut de femei care au urmat în timpul sarcinii tratament cu medicamentul de interes.
- Studii anterioare au arătat că deviația standard a indicelui de intelingență a copilului de 3 ani este egală cu 18 puncte.
- Care este numărul de copii în vârstă de 3 ani care trebuie incluşi în studiu pentru a obține un interval de confidență de 90% cu o margine a erorii mai mică sau egală cu 4 puncte?

ME ≤ 4
IC = 90%
Z = 1,65
Σ = 18

$$ME = Z_\alpha \frac{\sigma}{\sqrt{n}}$$

$$n = \left(\frac{Z_\alpha \times \sigma}{ME}\right)^2 = \left(\frac{1{,}65 \times 18}{4}\right)^2$$

$$n = 55{,}13 \rightarrow n = 56$$

- Se doreşte testarea efectului unui medicament folosit în tratamentul epilepsiei la mamă asupra dezvoltării cognitive a copilului. Dezvoltarea cognitivă se testează prin estimarea indicelui de inteligență a copilului de 3 ani născut de femei care au urmat în timpul sarcinii tratament cu medicamentul de interes.
- Studii anterioare au arătat că deviația standard a indicelui de intelingență a copilului de 3 ani este egală cu 18 puncte.
- Care este numărul de copii în vârstă de 3 ani care trebuie incluşi în studiu pentru a obține un interval de confidență de 90% cu o margine a erorii mai mică sau egală cu 4 puncte?

ME ≤ 4	**ME ≤ 2**	**ME ≤ 6**	**ME ≤ 4**	**ME ≤ 2**
IC = 90%	**IC = 90%**	**IC = 90%**	**IC = 95%**	**IC = 95%**
Z = 1,65	**Z = 1,65**	**Z = 1,65**	**Z = 1,96**	**n=311,17=312**
σ = 18	**σ = 18**	**σ = 18**	**σ = 18**	
n=56	**n=4×56=224**	**n=24,50**	**n=77,79=78**	**ME ≤ 5**
				IC = 95%
				n=49,79=50

- Un eşantion de 49 studenți au fost întrebați în câte relații exclusive au fost implicați până la data studiului. Studenții din eşantion au avut în medie 3 relații exclusive, cu o deviație standard de 1,2. estimați media adevărată a numărului de relații exclusive bazată pe rezultatele acestui eşantion utilizând intervalul de confidență de 95%. Distribuția de eşantionare a fost aproximativ normală.
- Pasul 1: verificarea condițiilor.
  - Numărul de relații exclusive ale unui student este independent de al altui student. $n = 49 < 10\% \times N$ (numărul de studenți din universitate)
  - $n>30$ → distribuția de eşantionare a numărului de relații exclusive dintr-un eşantion de volum egal cu 49 este aproximativ normală.

Exemplu

» Un eşantion de 49 studenți au fost întrebați în câte relații exclusive au fost implicați până la data studiului. Studenții din eşantion au avut în medie 3 relații exclusive, cu o deviație standard de 1,2. estimați media adevărată a numărului de relații exclusive bazată pe rezultatele acestui eşantion utilizând intervalul de confidență de 95%. Distribuția de eşantionare a fost aproximativ normală.

» Pasul 2: calculăm IC95%

$$n = 49$$

$$\bar{x} = 3$$

$$s = 1.2$$

$$SE = \frac{1.2}{\sqrt{60}} \approx 0.1714$$

$$ME = 3 \pm 1{,}96 * 0.1714$$

$$IC95\% \, [2{,}66;\ 5.66]$$

Suntem 95% siguri că studenții au fost implicați în medie în 3,66 - 5,66 relații exclusive.

Exemplu

Intervale de încredere pentru diferența între doi estimatori: Interpretare

» Dacă 0 este conținut în intervalul de încredere, diferența dintre cele două estimări (medii, proporții, rații, etc.) este zero

» Dacă zero nu este conținut în intervalul de încredere, diferența dintre cei 2 estimatori punctuali nu este egală cu zero.

- http://www.biomedcentral.com/1746-6148/8/68
- *BMC Veterinary Research* 2012, **8**:68 doi:10.1186/1746-6148-8-68

Results

After optimising the cut-off values in order to avoid doubtful results without deteriorating the concordance between the results of the two tests, the I-ELISA appeared to be slightly more sensitive than CFT ($Se_{I\text{-}ELISA}$=0.917 [0.822; 0.992], 95% Credibility Interval (CrI) compared to Se_{CFT}=0.860 [0.740; 0.967], 95% CrI). However, CFT was slightly more specific than I-ELISA (Sp_{CFT}=0.988 [0.947; 1.0], 95% CrI) compared to $Sp_{I\text{-}ELISA}$ =0.952 [0.901; 1.0], 95% CrI).

Intervale de încredere: Interpretare

» Când aceeaşi procedură se repetă pe mai multe eşantioane, intervalul de încredere (care va fi diferit pentru fiecare eşantion) va cuprinde in 95% din cazuri valoarea reală a estimatorului punctual.

Intervalul de încredere

- Se calculează în funcție de:
  - Talia eşantionului sau a populației
  - Tipul de variabilă (calitativă SAU cantitativă)
- Formula de calcul cuprinde 2 părți
  - Un estimator al calității eşantionului pe baza căruia estimatorul populației s-a calculat (eroarea standard)
    - Eroarea standard:
      - Cu cât *n* este mai mare cu atât eroarea standard este mai mică.
      - Este întotdeauna mai mică decât deviația standard
  - Gradul de încredere (confidență) al intervalului specificat (scorul Z_α)
- Se poate calcula pentru orice estimator

Intervalul de încredere pentru medie

» Media glicemiei la un eşantion de 121 pacienți este de 105 iar variația de 36. Care este intervalul de încredere al mediei glicemiei în populația din care s-a extras eşantionul cu un prag de semnificație α=0,05, considerând că glicemia este normal distribuită şi pentru acest prag Z = 1,96.

» n = 121

$$\overline{X} = 105$$

» $s^2 = 36$

» s = 6

$$\left[105 - 1{,}96\frac{6}{\sqrt{121}}; 105 + 1{,}96\frac{6}{\sqrt{121}}\right]$$

» [105-1.07, 105+1.07]

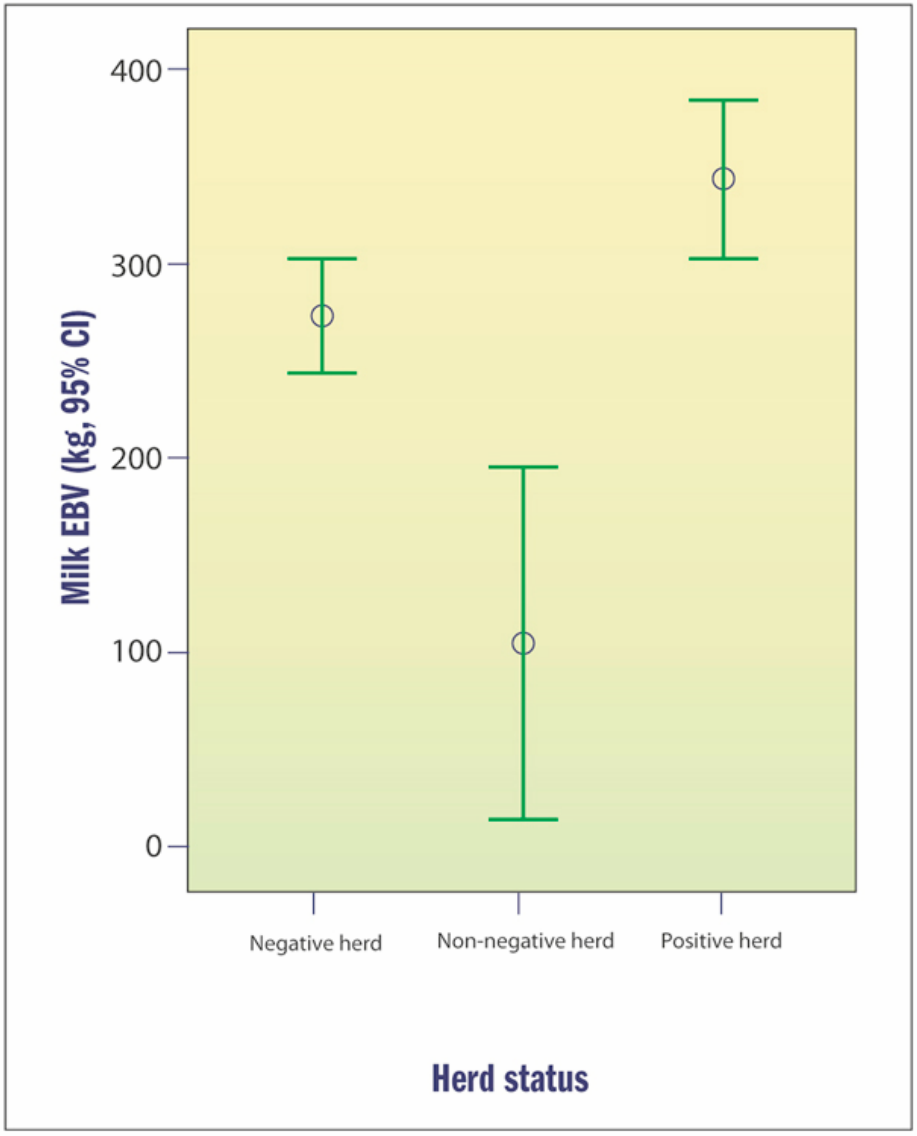
» [103.93 – 106.07]

» [104-106]

Compararea mediilor cu ajutorul intervalului de încredere

Figure 1.

Resolution: **standard** / high

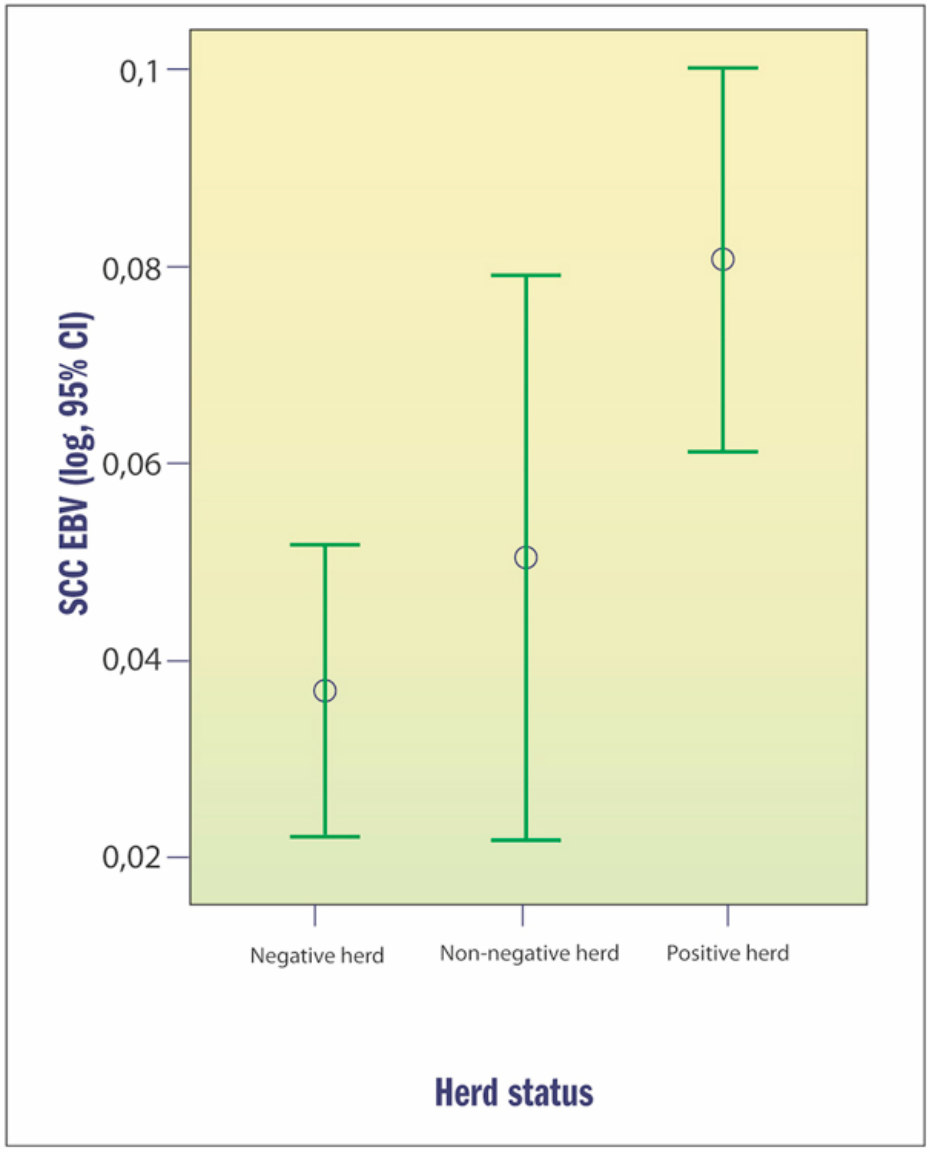


The Economic Breeding Value (EBV) for milk yield (kg) of paratuberculosis positive, non-negative or negative herds.

Hoogendam *et al. Irish Veterinary Journal* 2009 **62**(Suppl 4):265 doi:10.1186/2046-0481-62-4-265

Figure 2.

Resolution: **standard** / high



The Economic Breeding Value (EBV) for somatic cell count (SCC) of paratuberculosis positive, non-negative or negative herds.

Hoogendam *et al. Irish Veterinary Journal* 2009 **62**(Suppl 4):265 doi:10.1186/2046-0481-62-4-265

Intervalul de încredere pentru frecvențe

» Se calculează dacă:

- n*f > 10, unde n = talia eşantionului, f = frecvența

$$\left[f - Z_{\alpha}\sqrt{\frac{f(1-f)}{n}}; f + Z_{\alpha}\sqrt{\frac{f(1-f)}{n}} \right]$$

Intervalul de încredere pentru frecvențe

- Suntem interesați în estimarea frecvenței cancerului de sân la femeile între 50 şi 54 de ani care au antecedente familiale pozitive. Într-un studiu randomizat la care au participat 10000 de femei, s-a constatat că 400 dintre acestea au fost diagnosticate cu cancer de sân.
- Care este intervalul de încredere de 95% asociat frecvenței observate?

- f = 400/10000 = 0.04

$$\left[0{,}04 - 1{,}96\sqrt{\frac{0{,}04 \cdot 0{,}96}{10000}};\ 0{,}04 + 1{,}96\sqrt{\frac{0{,}04 \cdot 0{,}96}{10000}}\right]$$

- [0,04-0,004; 0,04+0,004]
- [0,036; 0,044]

$$\left[f - Z_\alpha\sqrt{\frac{f(1-f)}{n}};\ f + Z_\alpha\sqrt{\frac{f(1-f)}{n}}\right]$$

- Estimarea corectă a unui parametru statistic se face cu ajutorul intervalului de încredere.
- Intervalul de încredere depinde de volumul eşantionului şi de eroarea standard.
- Cu cât eroarea standard este mai mare cu atât intervalul de încredere este mai larg.
- Cu cât volumul eşantionului este mai mic cu atât intervalul de încredere este mai larg.

De reținut!