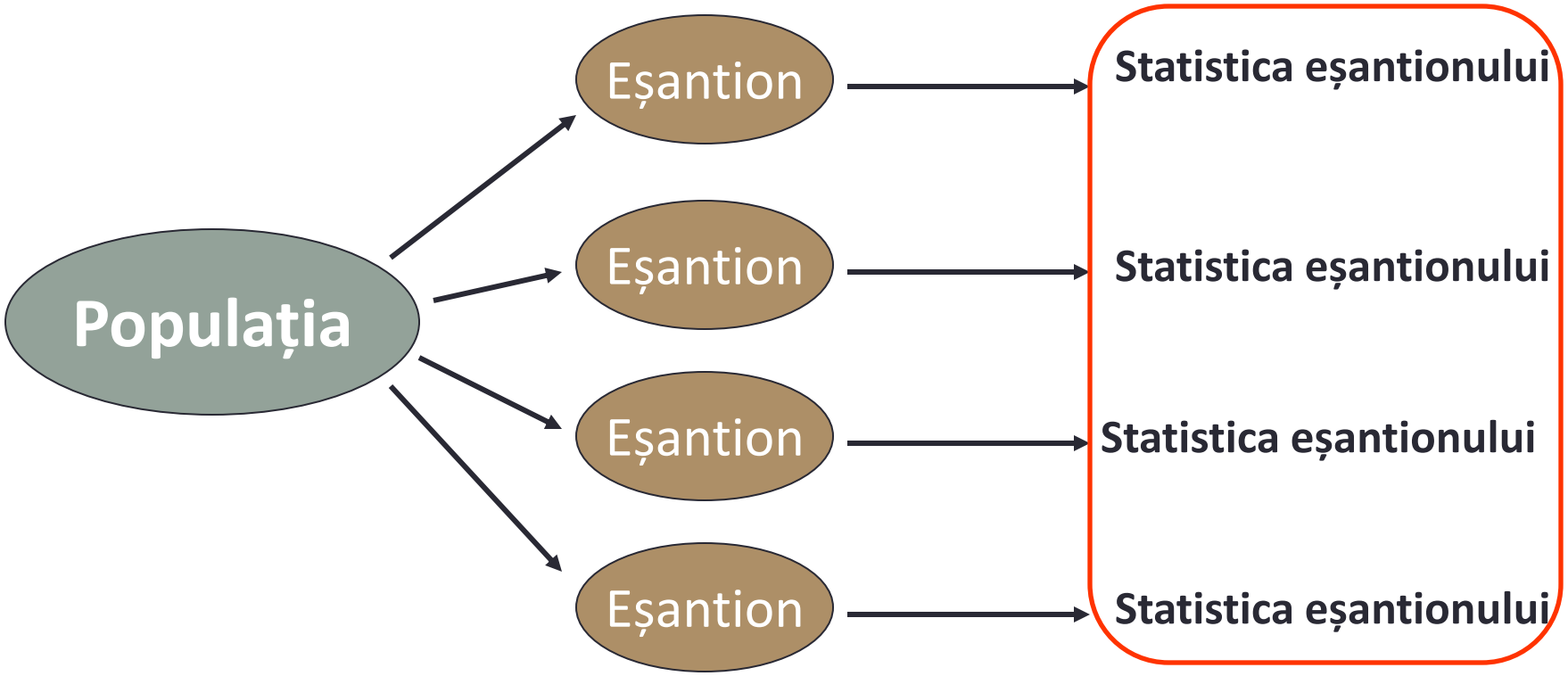
Estimatori Punctuali şi Intervale de Încredere

Despre ...

- Estimatorul punctual
- Intervalul de încredere:
  - Definiţii
  - Pentru
    - medie
    - diferența dintre medii
    - frecvență
    - diferența dintre frecvențe
    - riscul relativ

©2016 - Sorana D. BOLBOACĂ



Distribuția eșantionului ≠ **Distribuția de eșantionare**

©2016 - Sorana D. BOLBOACĂ

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

$\bar{X}_{AB}$

$\bar{X}_{GJ}$

$\bar{X}_{VN}$

$$\mu = \frac{\bar{X}_{AB} + \dots + \bar{X}_{GJ} + \dots + \bar{X}_{VN}}{N}$$

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

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

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

©2016 - Sorana D. BOLBOACĂ

Estimarea punctuală

- O valoare a parametrului teoretic estimat
  - media eşantionului ($\bar{X}$) este un estimator punctual al mediei populaţiei (μ)
- Este influenţată de fluctuaţiile 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 încredere
  - Parametrul estimat aparţine cu o probabilitate mare intervalului de confidenţă

©2016 - Sorana D. BOLBOACĂ

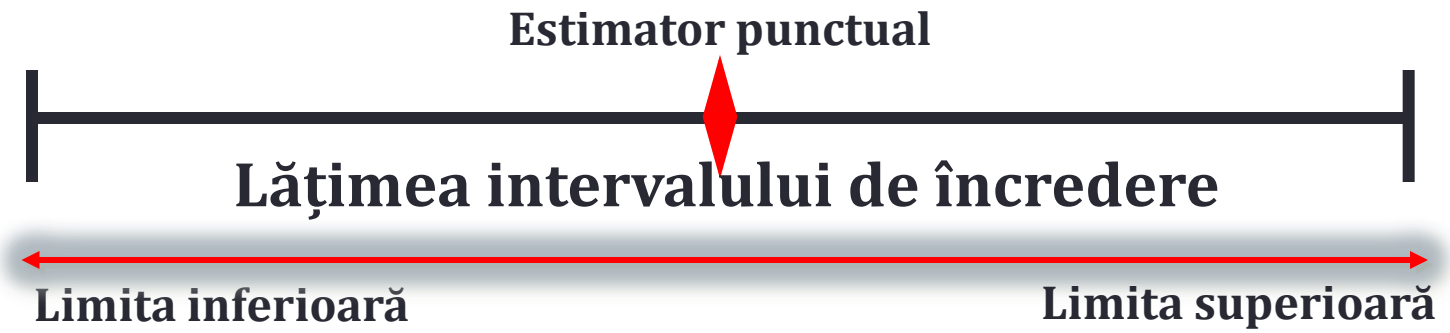
Definiţie

- 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 semnificaţie
- Intervalul definit de valorile critice va cuprinde estimatorul populaţiei cu o probabilitate de 1-α

©2016 - Sorana D. BOLBOACĂ

Estimatorul punctual vs. intervalul de încredere

- Estimatorul punctual = valoarea unei statistici obținută pe un eșantion
  - Cât de multă incertitudine este asociată estimatorului punctual?
- Un interval oferă mai multe informații despre o caracteristică a populației decât un estimator punctual



©2016 - Sorana D. BOLBOACĂ

Intervalul de încredere

- **Intervalul de încredere**:
  - Ia în considerare variabilitatea de eșantionare ➔ are valori diferite pentru fiecare eșantion
  - Se poate calcula pe baza observării unui singur eșantion
  - Oferă informații despre parametrul necunoscut al populației

- **Formula generală:**

  Estimator punctual ± (Valoare critică)×(Eroarea standard)

Marja de eroare

Valori din tabel

Depinde de α şi df (grade de libertate)

©2016 - Sorana D. BOLBOACĂ

Intervalul de încredere

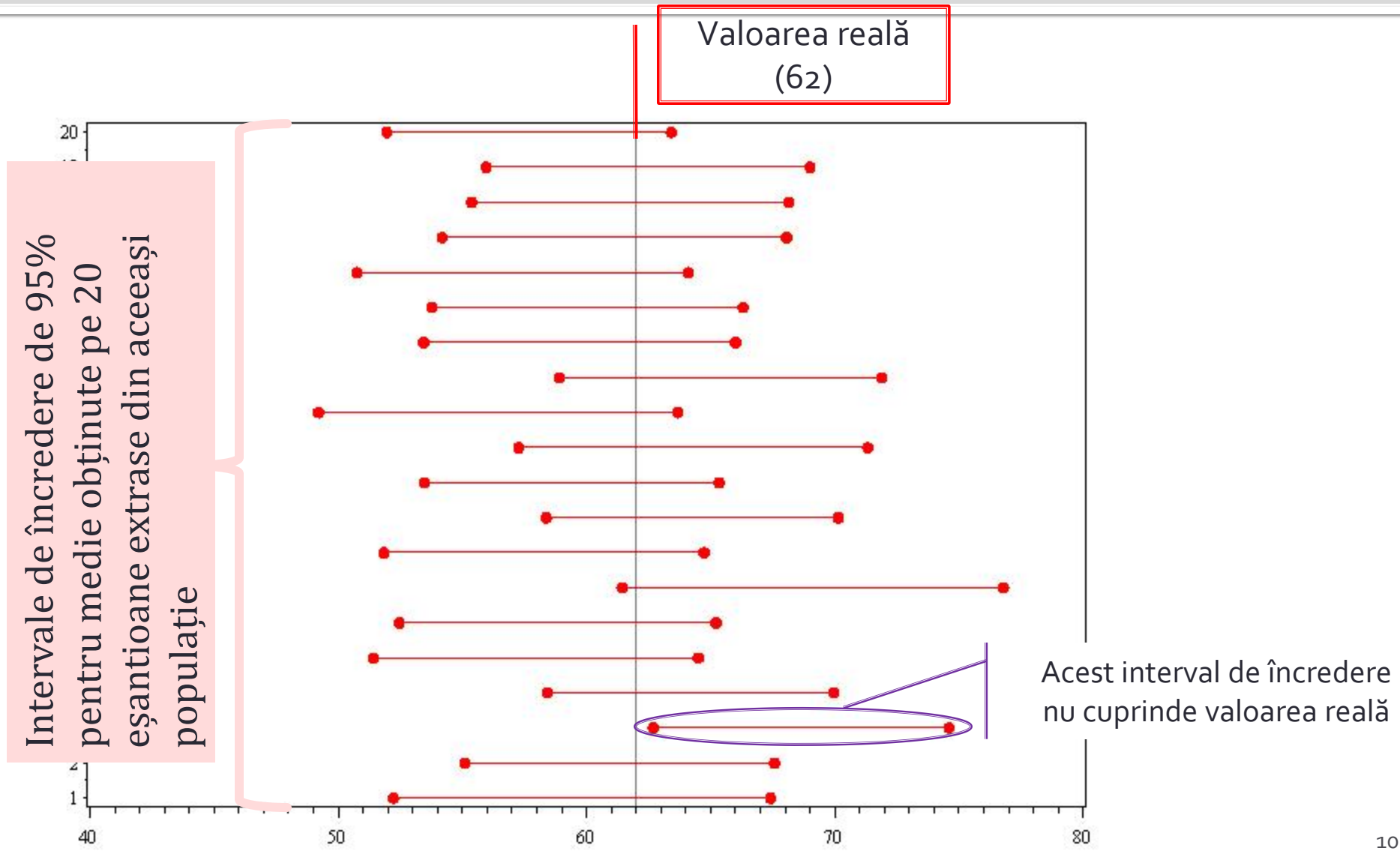
- Marja de eroare și respectiv lățimea intervalului de încredere
  - este cu atât mai mică cu cât volumul eșantionului e mai mare.
  - variază cu valoarea nivelului de semnificație (α)

- Nivel de semnificație $\alpha = 5\%$ → interval de încredere este de 95% (IC95%) - IC = $(1 - \alpha) = 0,95$

- Interpretare:
  - 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 vor conține valoarea reală a parametrului populației
  - Un interval de încredere poate să conțină sau poate să nu conțină valoarea reală a parametrului (datorită riscului de 5%)

©2016 - Sorana D. BOLBOACĂ

Intervalul de încredere



©2016 - Sorana D. BOLBOACĂ

Intervalul de încredere

- Se calculează în funcție de:
  - Talia eşantionului
  - 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
- Se poate calcula pentru orice estimator
- Nivele de confidență utilizate curent: 90%, 95%, 98%, și 99%

©2016 - Sorana D. BOLBOACĂ

Intervalul de încredere pentru medie

- Eroarea standard a mediei este egală cu deviaţia standard împărţită la radicalul volumului eşantionului
  - Dacă deviaţia standard este mare, şansa de eroare în estimator este mare
  - Dacă volumul eşantionului este mare, şansa erorii în estimator este mică

$$\bar{X} \pm Z_{\alpha} \frac{S}{\sqrt{n}}$$

- Condiţii de aplicare a intervalului de încredere:
  - Independența: Observațiile trebuie să fie independente
    - Eșantion randomizat / Asignare randomizată
    - Pentru un eșantion fără înlocuire, n < 10%*N
  - Volumul eșantionului: n ≥30 sau mai mare dacă distribuția în populație nu este simetrică

©2016 - Sorana D. BOLBOACĂ

Dacă dorim să fim siguri că intervalul conține parametrul populației, acesta va trebui să fie îngust sau larg?

IC larg ➔ acuratețea ↑ precizie ↓
n ↑ ↑ ➔ Acuratețe ↑ & precizie ↑

- Parametrul populației: nivelul colesterolului la femeile din România cu HTA și obezitate
- S-a extras din această populație un eșantion randomizat de volum 40 și s-a determinat o statistică a estimatorului punctual medie egală cu 220 mg/dl, deviația standard (S) egală cu 18 mg/dl
  - Media 220 mg/dl este estimatorul punctual al parametrului necunoscut al populației
  - Datorită variabilității de eșantionare, media va fi însoțită de intervalul de încredere asociat pentru estimarea corectă a parametrului populației

$$IC95\% = \left[220 - 1{,}96\frac{18}{\sqrt{49}}; 220 + 1{,}96\frac{18}{\sqrt{49}}\right] = [215; 225]$$

Lățimea = 225-215 = 10

$$IC99\% = \left[220 - 2{,}58\frac{18}{\sqrt{49}}; 220 + 2{,}58\frac{18}{\sqrt{49}}\right] = [213; 227]$$

Lățimea = 227-213 = 14

©2016 - Sorana D. BOLBOACĂ

- 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 inteligenţă a copilului de 3 ani născut de femei care au urmat în timpul sarcinii tratament cu medicamentul de interes.
- Studii anterioare au arătat că deviaţia standard a indicelui de intelingenţă 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 obţine un interval de confidenţă de 90% cu o margine a erorii mai mică sau egală cu 4 puncte?

ME ≤ 4
IC = 90%
Z = 1,65
s = 18

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

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

Avem nevoie de cel puțin 56 subiecți pentru a obține o margine de eroare de până la 4 puncte

©2016 - Sorana D. BOLBOACĂ

- Un eşantion de 49 studenţi au fost întrebaţi în câte relaţii exclusive au fost implicaţi până la data studiului. Studenţii din eşantion au avut în medie 3 relaţii exclusive, cu o deviaţie standard de 1,2. Estimaţi media adevărată a numărului de relaţii exclusive bazată pe rezultatele acestui eşantion utilizând intervalul de confidenţă de 95%. Distribuţia de eşantionare a fost aproximativ normală.

©2016 - Sorana D. BOLBOACĂ

Pasul 1: verificarea condiţiilor.

- Numărul de relaţii exclusive ale unui student este independent de al altui student. n = 49 < 10%×N (numărul de studenţi din universitate)
- n>30 → distribuţia de eşantionare a numărului de relaţii exclusive dintr-un eşantion de volum 49 este aproximativ normală.

- Pasul 2: calculăm IC95%

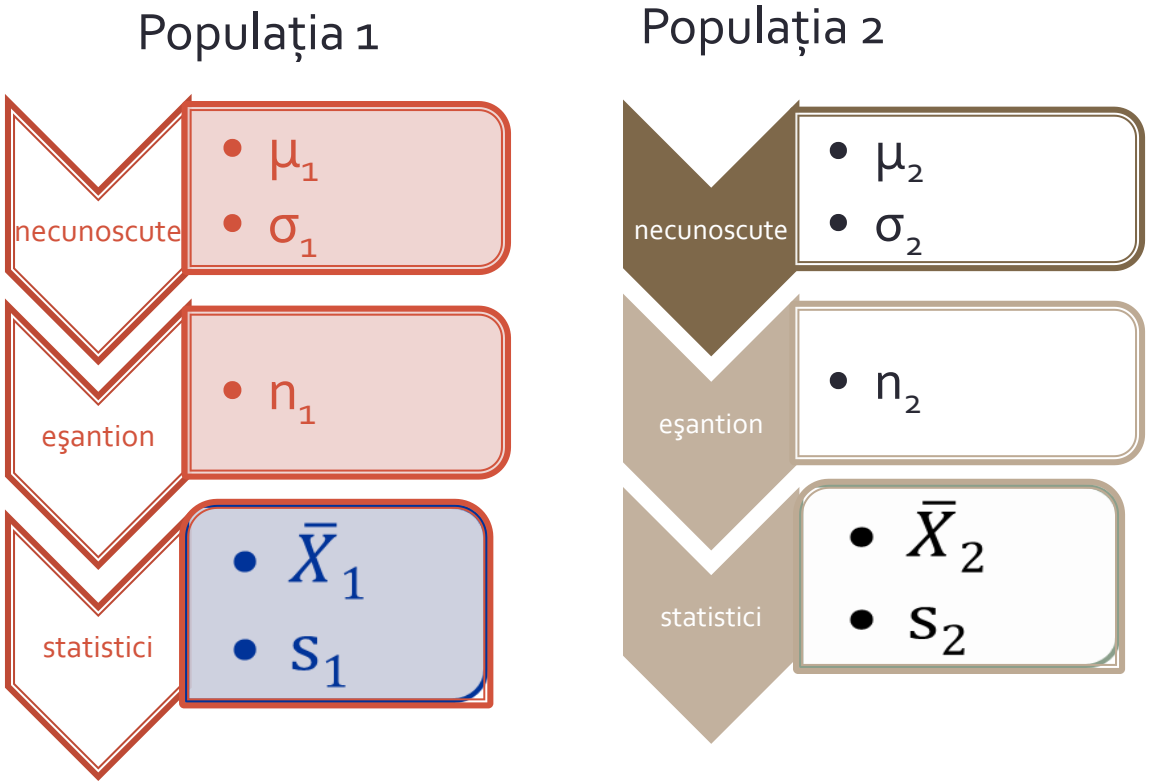
$$n = 49$$
$$\bar{x} = 3$$
$$s = 1{,}2$$

$$ES = \frac{1{,}2}{\sqrt{60}} \approx 0{,}1714$$
$$ME = 3 \pm 1{,}96 * 0{,}1714$$
$$IC95\%\,[2{,}66;\ 5{,}66]$$

- Suntem 95% siguri că studenţii au fost implicaţi în medie în 3 până la 6 relaţii exclusive.

©2016 - Sorana D. BOLBOACĂ

Intervale de încredere pentru diferența dintre medii



Estimăm $(\mu_1 - \mu_2)$ cu $\overline{X_1} - \overline{X_2}$

Interpretare:

- Dacă valoarea zero este în intervalul de încredere ➔ diferența dintre medii nu este semnificativ diferită de zero
- Dacă valoarea zero este în intervalul de încredere ➔ diferența dintre medii este semnificativ diferită de zero

©2016 - Sorana D. BOLBOACĂ

Intervale de încredere pentru diferența dintre medii

$$(\bar{X}_1 - \bar{X}_2) \pm t_{critic} \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

Grupa 10	7	7	8	8	8	6	9	6	5
Grupa 3	8	10	9	6	10	8	9	7	8

	Grupa 1	Grupa 2
Media	7,11	8,33
Variația	1,27	1,32
Deviația standard	1,61	1,75

df=15,97

pentru $\alpha = 0,05 \longrightarrow t_{15,97} = 2,13$

$(7,11-8,33) \pm 2,13\sqrt{(0,18+0,19)}$

$-1,22 \pm 2,13 * 0.61$

$-1,22 \pm 1,30 \longrightarrow [-2,52; 0,08]$

©2016 - Sorana D. BOLBOACĂ

Intervalul de încredere pentru frecvenţe

- O frecvenţă:
  - n*f > 10, unde n = talia eşantionului, f = frecvenţa

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

- Diferenţa dintre două frecvenţe

$(f_1-f_2)\pm Z_{critic}\times \text{sqrt}((f_1*(1-f_1)/n_1)+(f_2*(1-f_2)/n_2))$

©2016 - Sorana D. BOLBOACĂ

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]

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

©2016 - Sorana D. BOLBOACĂ

Intervalul de încredere pentru RR

	Boală+	Boală-
Expunere +	a	b
Expunere -	c	d

$$RR = \frac{a/(a+b)}{c/(c+d)}$$

$$\ln(\widehat{RR}) \pm Z_{crit}\sqrt{\frac{b/a}{a+b} + \frac{d/c}{c+d}} \rightarrow$$

$$\left[\exp\left(\ln(\widehat{RR}) - Z_{crit}\sqrt{\frac{\frac{b}{a}}{a+b} + \frac{\frac{d}{c}}{c+d}}\right); \exp\left(\ln(\widehat{RR}) + Z_{crit}\sqrt{\frac{\frac{b}{a}}{a+b} + \frac{\frac{d}{c}}{c+d}}\right)\right]$$

	CCPulmonar+	CCPulmonar-
Fumat+	30	20
Fumat-	15	35

$$RR = \frac{30/(30+20)}{15/(15+35)} = 2$$

ln(RR) = 0,69

Z_{critic} = 1,96

ES = sqrt(0,06) = 0,24

[exp(0,69-1,96*0,24); exp(0,69+1,96*0,24)] → [1,24; 3,23]

©2016 - Sorana D. BOLBOACĂ

Compararea Mediilor cu Ajutorul Intervalului de Încredere

http://www.biomedcentral.com/content/pdf/1471-2458-12-1013.pdf

Table 1 Living conditions of the MS-MV and the immigrant population (CASEN survey 2006)

	IMMIGRANT POPULATION 1% total sample, n = 154 431 weighted population (1877 real observations)		**MS-MV GROUP 0.67% total sample, n = 108 599 weighted population (1477 real observations)**	
	% or mean	**95% CI**	**% or mean**	**95% CI**
DEMOGRAPHICS				
Mean age**	*X* = 33.41	31.81–35.00	X = 26.13	23.41–28.26
Age categories:				
<16 years old**	13.60	11.29–16.28	45.25	39.53–51.10
16-65 years old**	79.08	75.92–81.93	47.26	41.64–52.94
>65 years old	7.32	5.33–9.97	7.49	5.31–10.46
Sex (female = 1)	45.21	41.74–48.72	51.27	47.99–55.41
Marital status:				
Single**	45.81	42.06–49.62	64.30	59.36–68.95
Married**	45.49	41.66–49.36	29.39	25.09–34.10

©2016 - Sorana D. BOLBOACĂ

De reţinut!

- 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.

$$\left[\overline{X} - Z_{\alpha}\frac{S}{\sqrt{n}}; \overline{X} + Z_{\alpha}\frac{S}{\sqrt{n}}\right]$$

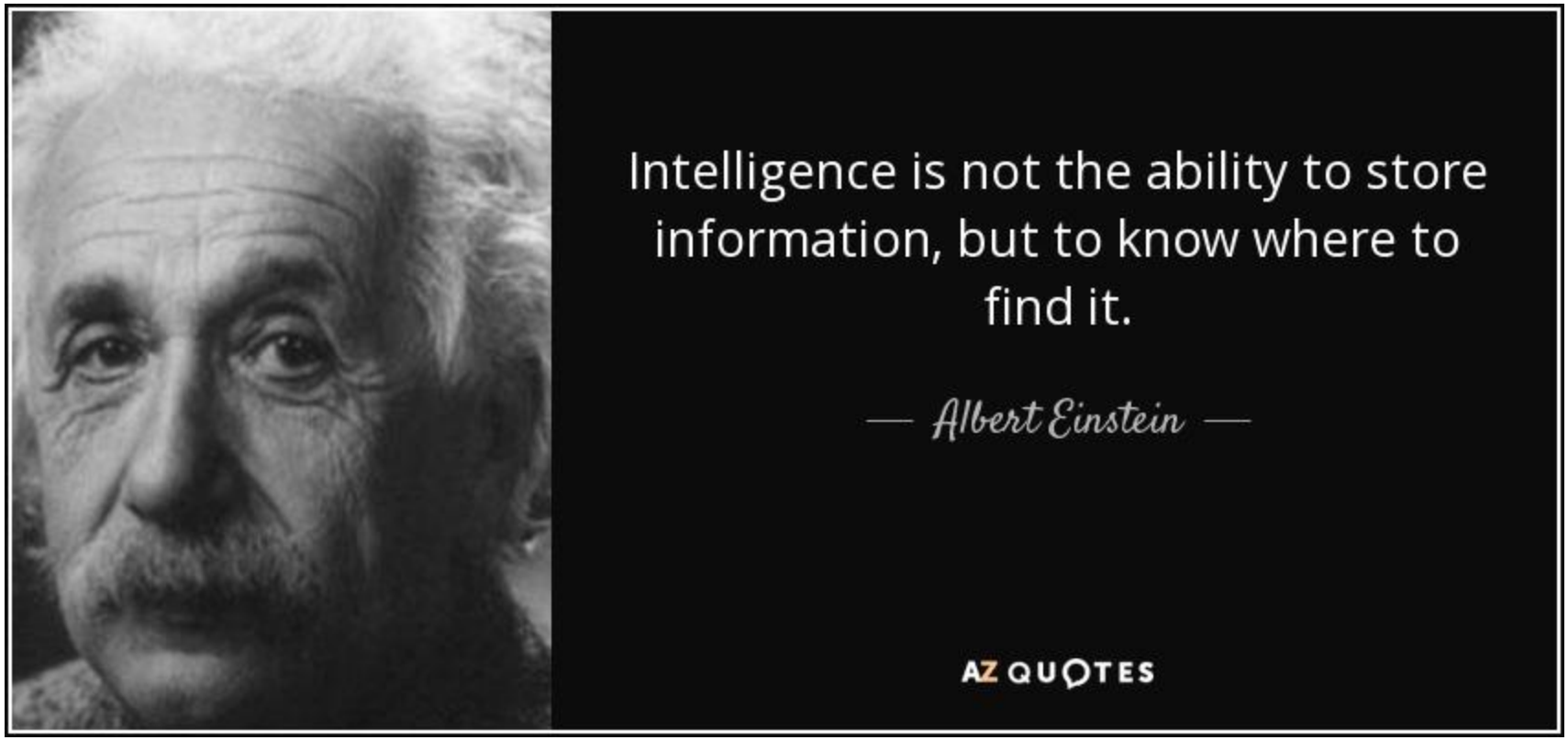
$$(\bar{X}_1 - \bar{X}_2) \pm t_{critic}\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}$$

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

$$(f_1\text{-}f_2)\pm Z_{criticc}\times\text{sqrt}((f_1*(1\text{-}f_1)/n_1)+(f_2*(1\text{-}f_2)/n_2))$$

©2016 - Sorana D. BOLBOACĂ

MULȚUMESC PENTRU ATENȚIE!



©2016 - Sorana D. BOLBOACĂ