

# TESTE STATISTICE: FRECVENȚE

# Cuprins

- Testul Hi-pătrat (variabile calitative)
- Testul Fisher (variabile calitative)
- Testul Z pentru frecvențe
  - 1 frecvență observată vs. frecvența teoretică
  - 2 frecvențe observate

# Testul Hi-pătrat ( $\chi^2$ )

- 2 variabile calitative & fiecare variabilă cu cel puțin 2 categorii
- **National Longitudinal Study of Adolescent Health ([Add Health](#))**:
  - Statele Unite ale Americii - includerea în studiu a adolescenților s-a realizat în anul școlar 1994-1995 – subiecții au fost urmăriți ca și adulți tineri (4 interviuri – ultimul în 2008 – vârsta subiecților 24-32 ani)
  - Obezitate (IMC > 30) și starea civilă

	căsătorit	necăsătorit	Total
obezitate +	147	184	331
obezitate -	277	685	962
Total	424	869	1293

- **Există o relație între obezitate și starea civilă?**

# Testul Hi-pătrat ( $\chi^2$ )

	căsătorit	necăsătorit	Total
obezitate +	147	184	331
obezitate -	277	685	962
Total	424	869	1293

**Există o relație între obezitate și starea civilă?**

- Ipoteze statistice:
  - $H_0$ : Obezitatea și starea civilă sunt independente
  - $H_1$ : Obezitatea și starea civilă sunt dependente
- Evaluarea ipotezelor:
  - Cuantificăm cât de diferite sunt valorile observate față de cele expectate (/teoretice)
  - Diferențe mari dintre valorile expectate (/teoretice) și cele observate furnizează dovezi puternice în favoarea ipotezei alternative

# Testul Hi-pătrat ( $\chi^2$ )

	căsătorit	necăsătorit	Total
obezitate +	147	184	331
obezitate -	277	685	962
Total	424	869	1293

**Există o relație între obezitate și starea civilă?**

$$\chi^2 = \sum_{i=1}^{r \cdot c} \frac{(f_i^0 - f_i^t)^2}{f_i^t}$$

$\chi^2$  = statistica  $\chi^2$

$f_i^0$  = frecvența observată

$f_i^t$  = frecvența teoretică

r = numărul de rânduri

c = numărul de coloane

df = (r-1)·(c-1)

## Condiții de aplicare

1. Independența – observațiile trebuie să fie independente (dacă eșantionarea se face fără înlocuire  $n < 10\%$  din populație)
2. Volumul eșantionului – fiecare celulă trebuie să conțină cel puțin 5 observații

# Testul Hi-pătrat ( $\chi^2$ )

Observat	căsătorit	necăsătorit	Total
obezitate +	<b>AP</b> 147	<b>FP</b> 184	331
obezitate -	<b>FN</b> 277	<b>AN</b> 685	962
Total	424	869	1293 ( <b>n</b> )

$$\text{valoarea expectată} = \frac{(\text{total rând}) \cdot (\text{total coloană})}{n}$$

Expectat	căsătorit	necăsătorit	Total
obezitate +	= $331 \cdot 424 / 1293$ <b>=109</b>	= $331 \cdot 869 / 1293$ <b>=222</b>	<b>331</b>
obezitate -	= $962 \cdot 424 / 1293$ <b>=315</b>	= $962 \cdot 869 / 1293$ <b>=647</b>	<b>962</b>
Total	<b>424</b>	<b>869</b>	<b>1293</b>

# Testul Hi-pătrat ( $\chi^2$ )

<b>Observat</b>	căsătorit	necăsătorit
obezitate +	147	184
obezitate -	277	685

<b>Expectat</b>	căsătorit	necăsătorit
obezitate +	<b>109</b>	<b>222</b>
obezitate -	<b>315</b>	<b>647</b>

$$\chi^2 = \sum_{i=1}^{r \cdot c} \frac{(f_i^0 - f_i^t)^2}{f_i^t}$$

$$\chi^2 = (147-109)^2/109+(184-222)^2/222+(277-315)^2/315+(685-647)^2/647$$

$$\chi^2 = 26,57$$

$$p = 1,79 \cdot 10^{-7} \rightarrow \text{respingem } H_0$$

Putem însă să tragem concluzia că obezitatea este determinată de căsătorie?

Dacă  $\chi^2 \in [3,84, \infty)$  se respinge  $H_0$  cu un risc de eroare de tip I ( $\alpha$ ).

Dacă  $\chi^2 \notin [3,84, \infty)$  nu se respinge  $H_0$  cu un risc de eroare de tip II ( $\beta$ ).

# Testul Hi-pătrat ( $\chi^2$ )

- Testul Hi pătrat este valid dacă cel puțin 80% dintre frecvențele probabile depășesc 5 și toate frecvențele probabile depășesc 1
- Regula se aplică frecvențelor extectate/teoretice, nu celor observate
- Dacă criteriul nu este satisfăcut → se aplica corecția Yates (corecție de continuitate)

# Testul Hi-pătrat ( $\chi^2$ ): Corelția Yates

- Dacă volumul de eșantion este mic, testul Hi-pătrat are tendința să respingă  $H_0$  identificând diferențe acolo unde nu există
- Corecția Yates: micșorarea cu 0,5 unități a diferenței dintre frecvența observată și cea expectată în cadrul numărătorului lui  $\chi^2$  (din formulă) înainte de ridicarea la pătrat; astfel, valoarea lui  $\chi^2$  scade

$$\chi^2 = \sum_{i=1}^{r \cdot c} \frac{(|f_i^0 - f_i^t| - 0.5)^2}{f_i^t}$$

# Testul Hi-pătrat ( $\chi^2$ ): Corecția Yates

## Aplicarea corecției

- duce la scăderea valorii  $\chi^2$ 
  - ➔ șansele ca ipoteza nulă să fie respinsă ↓ ↓
  - ➔ riscul erorii de tipul I (respingerea ipotezei nule atunci când aceasta este în fapt veridică) ↓ ↓
- crește riscul erorii de tipul II (acceptarea unei false ipoteze atunci când ea este în fapt falsă).
- Unii statisticieni recomandă utilizarea întotdeauna a corecției de continuitate în cazul unui tabel de contingență  $2 \times 2$

# Testul exact Fisher

- = alternativă a testului Hi-pătrat în examinarea asociațiilor în cadrul unui tabel de contingență  $2 \times 2$ , atunci când frecvențele expectate/teoretice/probabile sunt mici.
- Se aplică când:
  - Frecvența expectată/teoretică  $< 2$
  - $> 20\%$  din frecvențele expectate/teoretice/probabile  $< 5$
- Valoarea  $p$  asociată parametrului ne dă probabilitatea ca valoarea observată de independență să fie atribuită doar șansei.
- O valoare  $p$  mică indică că există alte cauze decât șansa influențează rezultatul și astfel cele două variabile investigate nu sunt independente

$$p = \frac{(AP + FN)!(FP + AN)!(FN + AN)!(AP + FP)!}{n!AP!FP!FN!AN!}$$

# Testul exact Fisher: [OpenEpi](#)

[Expand All](#) | [Collapse](#)

- Home
- Info and Help
- Language/Options/Settings
- Calculator
- Counts
  - Std.Mort.Ratio
  - Proportion
  - Two by Two Table
  - Dose-Response
  - R by C Table
  - Matched Case Control
  - Screening
- Person Time
  - 1 Rate
  - Compare 2 Rates
- Continuous Variables
  - Mean CI
  - Median/%ile CI
  - t test
  - ANOVA
- Sample Size
- Power
- Random numbers
- Searches
  - Google--Internet
  - PubMed--MEDLARS
- Internet Links
- Download OpenEpi
- Development

Start | Enter | **Results** | Examples | Help

## 2 x 2 Table Statistics

### Single Table Analysis

		Disease		
		(+)	(-)	
Exposure	(+)	10	8	18
	(-)	8	2129	2137
		18	2137	2155

### Chi Square and Exact Measures of Association

Test	Value	p-value(1-tail)	p-value(2-tail)
Uncorrected chi square	3.677	0.02759	0.05518
Yates corrected chi square	2.588	0.05384	0.1077
Mantel-Haenszel chi square	3.599	0.02892	0.05784
Fisher exact		0.05415	0.1083
Mid-P exact		0.03360	0.06719

All expected values (row total\*column total/grand total) are  $\geq 5$   
OK to use chi square.

# Testul exact Fisher: [OpenEpi](#)

[Expand All](#) | [Collapse](#)

- Home
- Info and Help
- Language/Options/Settings
- Calculator
- Counts
  - Std.Mort.Ratio
  - Proportion
  - Two by Two Table
  - Dose-Response
  - R by C Table
  - Matched Case Control
  - Screening
- Person Time
  - 1 Rate
  - Compare 2 Rates
- Continuous Variables
  - Mean CI
  - Median/%ile CI
  - t test
  - ANOVA
- Sample Size
- Power
- Random numbers
- Searches
  - Google--Internet
  - PubMed--MEDLARS
- Internet Links
- Download OpenEpi
- Development

Start	Enter	Results	Examples	Help
-------	-------	---------	----------	------

  

**Single Table Analysis**

	Disease	
	(+)	(-)
Exposure (+)	10	5
Exposure (-)	1	6
	11	1122

  

**Chi Square and Exact Measures of Association**

Test	Value	p-value(1-tail)	p-value(2-tail)
Uncorrected chi square	5.238	0.01105	0.02210
Yates corrected chi square	3.352	0.03356	0.06711
Mantel-Haenszel chi square	5	0.01267	0.02535
Fisher exact		0.03173	0.06347
Mid-P exact		0.01683	0.03367

  

At least one expected value (row total\*column total/grand total) is < 5  
 Fisher or Mid-P exact tests are recommended rather than chi square.

# Testul Z pentru frecvențe

## 1 frecvență observată ( $f$ ) vs. frecvență teoretică ( $p$ )

- Condiții de aplicare:
  - Se cunoaște frecvența teoretică
  - $n \times p > 10$  &  $n \times (1-p) > 10$

$$z = \frac{f - p}{\sqrt{\frac{p(1-p)}{n}}}$$

## 2 frecvențe observate

- Condiții de aplicare:
  - $n_1$  &  $n_2 > 30$

$$z = \frac{(f_1 - f_2)}{\sqrt{f(1-f) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

$$f = \frac{f_1 n_1 + f_2 n_2}{n_1 + n_2}$$

# Testul Z pentru frecvențe

- Suntem interesați de investigarea frecvenței hepatitei B la personalul care lucrează în laboratoarele clinicilor de boli infecțioase din Transilvania. Se știe din studii anterioare că frecvența hepatitei B în populația generală din Transilvania este de 9%. S-a luat în studiu un eșantion de 100 persoane și s-a obținut o prevalență a hepatitei B de 6%. Există diferență semnificativă între frecvența hepatitei B la personalul care lucrează în laboratoarele spitalelor de boli infecțioase din Transilvania față de populația generală?
- $f = 0,06$ ,  $p = 0,09$ ,  $n = 100$
- $H_0$ : Nu există diferență semnificativă între frecvența hepatitei B la eșantionul studiat față de frecvența hepatitei B în populația generală.
- $H_1$ : Există diferență semnificativă între frecvența hepatitei B la nivelul eșantionului și prevalența hepatitei B în populația generală.

# Testul Z pentru frecvențe

$$z = \frac{f - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0,06 - 0,09}{\sqrt{\frac{0,09(1-0,09)}{100}}} = \frac{-0,03}{\sqrt{\frac{0,09 \cdot 0,91}{100}}} = \frac{-0,03}{\sqrt{\frac{0,0819}{100}}} = \frac{-0,03}{\sqrt{0,000819}} = \frac{-0,03}{0,029} = -1,05$$

Pragul de semnificație:  $\alpha = 0,05$ .

Regiunea critică test bilateral:  $(-\infty; -1,96] \cup [1,96; \infty)$

## ■ Concluzia testului:

- Deoarece parametrul statistic calculat al testului nu aparține regiunii critice, nu se respinge ipoteza nulă. Nu există suficiente dovezi pentru a afirma că frecvența hepatitei B la eșantionul studiat este semnificativ diferită față de frecvența hepatitei B în populația generală.

# Testul Z pentru frecvențe

- S-a studiat statutul HIV pe un eșantion de 170 femei cu vârste cuprinse între 18 și 40 de ani din Moldova, și respectiv un eșantion de 89 femei cu vârste cuprinse între 18 și 40 de ani din Transilvania. Pentru eșantionul din Moldova, Frecvența testelor HIV+ a fost de 10% în eșantionul din Moldova și 2,7% în eșantionul din Transilvania. Frecvența infecției cu HIV la femeile cu vârste cuprinse între 18 și 40 de ani din Moldova este diferită față de frecvența infecției la femeile de aceeași vârstă din Transilvania?

**Datele problemei:**  $p_1 = 0,10$ ;  $p_2 = 0,027$ ;  $n_1 = 170$ ;  $n_2 = 89$ .

**$H_0$ :** Nu există o diferență semnificativă între frecvența infecției HIV la femeile din Moldova față de frecvența infecției HIV la femeile din Transilvania.

**$H_1$ :** Există o diferență semnificativă între frecvența infecției HIV la femeile din Moldova față de frecvența infecției HIV la femeile din Transilvania.

# Testul Z pentru frecvențe

**Pragul de semnificație:**  $\alpha = 0,05$ .

**Regiunea critică:**

- Testul bilateral:  $(-\infty; -1,96] \cup [1,96; \infty)$
- Testul unilateral:  $[1,645, \infty)$
- 2,118 aparține regiunii critice  $\rightarrow$  respingem  $H_0$

$$z = \frac{(p_1 - p_1)}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0,10 - 0,027}{\sqrt{0,075(1-0,075)\left(\frac{1}{170} + \frac{1}{89}\right)}}$$
$$z = \frac{0,073}{\sqrt{0,075 \cdot 0,925 \cdot (0,006 + 0,011)}} = \frac{0,073}{\sqrt{0,001}} = \frac{0,073}{0,034} = 2,118$$

**AIM:** To determine whether there are gender differences in the epidemiological profile of atrial fibrillation (AF) and to characterise the clinical, biochemical, and therapeutic factors associated with AF.

### **Gender differences**

Table 2 shows the differences between the men and the women in epidemiological, biochemical, BP and therapeutic characteristics. The women exhibited more obe-

**Table 2 Epidemiological, clinical and therapeutic differences between genders (mean  $\pm$  SD) *n* (%)**

	<b>Females (<i>n</i> = 542)</b>	<b>Males (<i>n</i> = 486)</b>	<b><i>P</i> value</b>
Mean age, yr	72.7 $\pm$ 5.8	72.8 $\pm$ 5.8	NS
Abdominal circumference, cm	96.6 $\pm$ 11.8	100.4 $\pm$ 11.0	< 0.001
Weight, kg	71.4 $\pm$ 11.5	79.5 $\pm$ 11.5	< 0.001
Mean height, cm	155.2 $\pm$ 6.7	166.7 $\pm$ 6.7	< 0.001
BMI	29.6 $\pm$ 4.5	28.6 $\pm$ 3.6	< 0.001
Obesity	224 (41.4)	160 (32.9)	0.005
Years from the onset of HT	11.0 $\pm$ 8.2	10.8 $\pm$ 8.1	NS
Diabetes mellitus	134 (24.7)	150 (30.9)	0.03
Dyslipidaemia	267 (49.3)	230 (47.3)	NS
Smokers	17 (3.1)	76 (15.6)	< 0.001
Sedentariness	352 (70.5)	274 (56.4)	< 0.001
Regular alcohol intake	5 (0.9)	33 (6.8)	< 0.001

**Table 1. Characteristics of obese and not obese participants.**

	<b>Not Obese (N = 377,220) N (%)</b>	<b>Obese (N = 122,284) N (%)</b>	<b>P value*</b>
<b>Age (years)</b>			
31–40	31,493 (8.3)	8,508 (7.0)	<0.001
41–50	105,116 (27.9)	32,691 (26.7)	
51–60	152,997 (40.6)	52,308 (42.8)	
≥61	87,614 (23.2)	28,777 (23.5)	
<b>Gender</b>			
Female	207,664 (55.1)	64,326 (52.6)	<0.001
Male	169,556 (44.9)	57,958 (47.4)	
<b>Alcohol consumption</b>			
Never	27,556 (7.3)	12,552 (10.3)	<0.001
Occasionally	38,371 (10.2)	19,219 (15.8)	
1–3 per month	39,228 (10.4)	16,382 (13.4)	
1–2 per week	96,831 (25.7)	31,919 (26.2)	
3–4 per week	91,836 (24.4)	23,173 (19.0)	
Daily	82,646 (22.0)	18,700 (15.3)	
Missing	752	339	
<b>Hypertension</b>			
No	293,635 (78.1)	69,271 (56.9)	<0.001
Yes	82,356 (21.9)	52,481 (43.1)	
Missing	1,229	532	
<b>Diabetes</b>			
No	363,790 (96.8)	107,486 (88.5)	<0.001
Yes	12,090 (3.2)	13,997 (11.5)	
Missing	1,340	801	

N number.

\*chi square test for gender, hypertension and diabetes, chi-square test for trend for remainder.

# De reținut!

- Aplicarea corectă a unui test pe frecvențe se face după verificarea asumpțiilor testului care se dorește a fi aplicat
- Testul Hi-pătrat se poate aplica pe orice tabel de contingență de  $r \times n$
- Atenție la condițiile de aplicare a testului Hi-pătrat, a testului Hi-pătrat corectat sau Fisher exact!

# Tema II

- Probleme de statistică medicală
- Test cu întrebări cu răspunsuri multiple
- Testul e disponibil în perioada 18 Decembrie 2015 – 7 Ianuarie 2016 orele 8 p.m. [aici](#)
- Pentru a finaliza testul trebuie să răspundeți la toate întrebările
- Se ia în considerare doar prima submitere a testului!

