II.F

PRELUCRAREA DATELOR BIOMEDICALE PRIN METODE BIOSTATISTICE. PROBLEME

1. Introducere. Obiectivele şedinţelor de exerciţii practice de prelucrări de date

Biostatistica oferă metodele şi instrumentele de descriere şi analiză a datelor medicale, a rezultatelor experimentelor şi observaţiilor biomedicale. Totodată, abordarea problemelor medicale din punct de vedere statistic permite atât planificarea studiilor experimentale şi observaţionale, cât şi descrierea fenomenelor biologice prin metode matematice.

Domeniul este vast şi necesită un studiu aprofundat, dincolo de o introducere succintă pe care o putem acoperi în spaţiul rezervat din cadrul cursurilor pentru studenţii de la Universitatea de Medicină. În selecţia de probleme prezentată în continuare (ca o completare a secţiunii de biostatistică din cursul Mihalaş & Lungeanu 2009) ne-am propus familiarizarea cu noţiunile de bază privind prelucrările statistice, acoperind în acelaşi timp aspectele fundamentale:

- descrierea datelor calitative şi cantitative;
- estimarea statistică şi testarea statistică a ipotezelor de cercetare biomedicală;
- analiza şi corelarea;
- studii epidemiologice de risc.

Materialul începe cu prezentarea unor probleme rezolvate şi apoi propune o serie de probleme la care nu sunt date soluţiile, lăsând cititorilor plăcerea de a le explora pe cont propriu. Problemele propuse sunt preluate şi adaptate din: (a) Knapp & Clinton Miller 1992; (b) Rosner 2000.

Toate problemele propuse se pot rezolva cu *EpiInfo* şi/sau *Microsoft Excel*. Soluţiile sunt prezentate în varianta rezolvării lor cu unul dintre cele două pachete, dar aceste abordări nu sunt în nici un fel restrictive. La fel de bine, studenţii pot încerca utilizarea altor pachete statistice: *R*, *Statistica*, *SPSS*, etc.

2. Descrierea datelor. Estimarea statistică

Problema 2.1

Într-un studiu privind asocierea dintre fumat şi boala coronariană, o echipă de cercetători a obţinut următoarea distribuţie a valorilor presiunii arteriale sistolice (în mm Hg) pentru un grup de 37 de fumători.

Presiunea arteriala	Frecvente	Frecvente relative	Frecvente
89.5-109.5	5	13.5	13.5
109.5-129.5	15	40.5	54.0
129.5-149.5	**A**	27.0	**C**
149.5-169.5	3	8.1	89.1
169.5-189.5	2	**B**	94.5
189.5-209.5	**D**	5.4	99.9
TOTAL	**37**		**100.0**

1. Valoarea **A** din tabel este:
2. Valoarea **B** din tabel este:
3. Valoarea **C** din tabel este:
4. Probabilitatea ca un individ selectat la întâmplare din acest eşantion de 37 de fumători să aibă presiunea sistolică între 89.5 mm Hg şi 129.5 mm Hg este:

Soluţii 2.1

1. Valoarea **A** din tabel este: $37*0.27=9.99 \Rightarrow$ **A=10**

2. Valoarea **B** din tabel este: $\frac{2}{37}*100 = 5.4 \Rightarrow$ **B=5.4**

3. Valoarea **C** din tabel este: $54.0+27.0 = 81.0 \Rightarrow$ **C=81.0**

4. Probabilitatea ca un individ selectat la întâmplare din acest eşantion de 37 de fumători să aibă presiunea sistolică între 89.5 mm Hg şi 129.5 mm Hg este: **54.0%** (coloana de frecvenţe cumulate, rândul 2).

Problema 2.2

În procesul de pregătire pentru un test naţional, un lot de 200 de studenţi au dat un test simulat la care au avut de răspuns la un număr de 100 de întrebări. Fiecare student a răspuns corect la un număr de întrebări între 35 şi 59. Media

aritmetică a numărului de răspunsuri corecte a fost 47, iar deviaţia standard de 4. Presupunem că numărul de răspunsuri corecte urmează o distribuţie normală.

1. Procentul de studenţi care a răspuns corect la un număr de întrebări cuprins între 43 şi 51 este de aproximativ:

2. Procentul de studenţi care a răspuns corect la cel puţin 55 de întrebări este de aproximativ:

Soluţii 2.2

1. Dacă media este m=47, iar deviaţia standard s=4, atunci în intervalul $(47 \pm 4) = (43;51)$ se vor găsi aproximativ 68% dintre valori (în situaţia în care presupunem că valorile urmează o distribuţie normală).

Concluzie: 68% dintre studenţi au răspuns corect la un număr de întrebări cuprins între 43 şi 51.

2. Dacă media este m=47, iar deviaţia standard s=4, atunci în intervalul $(47 \pm 2 * 4) = (39;55)$ se vor găsi aproximativ 95% dintre valori (în situaţia în care presupunem că valorile urmează o distribuţie normală). Înseamnă că doar aproximativ $(1 - 0.95)/2 = 0.025$ dintre valori sunt mai mari sau egale cu 55.

Concluzia: aproximativ 2.5% dintre studenţi au răspuns corect la cel puţin 55 de întrebări.

Problema 2.3

Centrul de diabet şi boli de nutriţie din Timişoara are în observaţie un număr mare de pacienţi din care în atenţia noastră intră doar un eşantion de 32 de pacienţi. Glicemia a fost determinată cu un aparat portabil (glucotest, cu bandelete). Datele sunt prezentate în tabelul de mai jos.

	SEX	GLICEMIE mg%
1	M	127
2	F	130
3	F	126
4	M	111
5	F	109
6	M	99
7	F	116
8	F	116
9	F	150
10	M	95
11	F	120
12	M	100
13	M	116
14	M	116
15	M	114
16	M	108
17	M	112
18	M	98
19	M	115
20	F	109
21	F	106
22	F	103
23	M	97
23	F	95
25	M	119
26	M	117
27	M	105
28	M	137
29	F	106
30	F	101
31	F	129
32	M	132

1. Pentru un nivel de încredere 1-α=0.68, determinaţi intervalul de încredere pentru valorile de glicemie. Interpretaţi rezultatul obţinut (formulaţi concluzia în cuvinte).

2. Estimaţi media pentru valorile de glicemie ale acestor pacienţi, cu 1-α=0.95. Interpretaţi rezultatul obţinut (formulaţi concluzia în cuvinte).

Soluţii 2.3

Pentru simplificare, mai jos vom presupune că valorile de glicemie urmează o distribuţie normală. Sunt considerate corecte atât versiunile de

rezolvare riguroasă (utilizând distribuţia t), cât şi cele simplificate (utilizând distribuţia normală şi scorurile z).

Media valorilor de glicemie este m=113.56, iar deviaţia standard s=13.09 (se pot determina cu *Epi*, cu *Microsoft-Excel*, sau chiar prin calcul manual).

1. Pentru 1-α=0.68, z≅1, deci intervalul de încredere pentru valorile individuale de glicemie va fi aproximativ

$$(113.56 \pm 1 * 13.09) = (100.47;126.65)$$

Concluzia: putem spune că aproximativ 68% dintre valorile de glicemie se vor încadra între 100.47 şi 126.65

2. Pentru 1-α=0.95, z≅2, deci intervalul de încredere pentru media valorilor de glicemie va fi aproximativ

$$\left(113.56 \pm 2 * \frac{13.09}{\sqrt{32}}\right) = (108.94;118.19)$$

Concluzia: putem spune că media valorilor de glicemie (pentru populaţia din care a fost extras acest eşantion) se încadrează între 108.94 şi 118.19, cu o probabilitate de 95%.

3. Teste statistice

Problema 3.1

Cu datele de la Problema 2.3, ce test statistic se aplică pentru a vedea dacă există diferenţe între glicemia celor două sexe? Formulaţi ipotezele statistice.

Aplicaţi testul şi trageţi concluzia. (α= .05). Formulaţi în termeni statistici şi medicali.

Soluţii 3.1

Glicemia este o variabilă numerică despre care presupunem că urmează o distribuţie normală - rezultă că tendinţa centrală va fi descrisă de medie, iar ipotezele statistice se vor referi la acest parametru.

Ipotezele statistice sunt:

H_0: $\mu_M = \mu_F$

media glicemiei este aceeaşi pentru cele două sexe;

H_a: $\mu_M \neq \mu_F$

media glicemiei este diferită pentru cele două sexe;

Se aplică testul **t nepereche** (avem două eşantioane independente - unul de femei şi unul de bărbaţi) în varianta bi-direcţională (2-tailed).

Se aplică testul **t-nepereche *2-tailed***: se obţine statistica t=0.706, căreia îi corespunde valoarea p=0.486 > α=0.05.

Concluzie: Se acceptă H_0 şi vom spune că NU există diferenţe statistic semnificative între mediile glicemiei pentru cele două sexe.

Concluzia medicală: nivelul de glicemie este acelaşi pentru pacienţii din acest centru, indiferent de sex – diferenţele observate se încadrează în limitele variabilităţii biologice.

Problema 3.2

S-a făcut un studiu menit să investigheze efectele contraceptivelor orale (CO) asupra bolilor de inimă la femeile cu vârsta între 40 şi 44 de ani. S-a găsit că între cele 5000 de femei care iniţial erau utilizatoare curente de CO, 13 femei au dezvoltat infarct miocardic (IM) în decursul unei perioade de 3 ani. În acelaşi timp, dintre cele 10000 de femei care nu utilizau CO, doar 7 au dezvoltat IM în perioada de 3 ani.

Cum putem determina dacă aceste diferenţe sunt semnificative sau ele se datorează doar şansei?

1. Sintetizaţi rezultatele studiului într-un tabel de contingenţă.

2. Ce test statistic se utilizează pentru a determina semnificaţia presupusei asocieri dintre utilizarea CO şi bolile de inimă?

3. Formulaţi ipotezele statistice.

4. Aplicaţi testul ales. Trageţi concluzia statistică.

5. Formulaţi concluzia medicală.

Soluţii 3.2

Tabelul de contingenţă va fi:

	IM +	**IM -**	Totaluri
utiliz CO	13	4987	5000
NU utiliz CO	7	9993	10000
Totaluri	20	14980	15000

Se aplică **testul χ^2** – se compară proporţiile celor care dezvoltă infarct miocardic în rândul utilizatoarelor CO, respectiv a celor care nu utilizează CO.

Ipoteza de zero: proporţia celor care dezvoltă IM este aceeaşi în cele 2 grupe de femei (utilizatoare, respectiv ne-utilizatoare CO).

Ipoteza alternativă: există proporţii diferite de cazuri de IM pentru cele două grupe de femei.

Se aplică testul χ^2 (Epi aplică implicit varianta bi-direcţională).

Se obţine:

χ^2 (*uncorrected*) =9.04, p=0.0026

χ^2 (*Yates corrected*) =7.67, p=0.0056

(corecţia de continuitate – în această situaţie concluziile statistice sunt aceleaşi).

Concluzie: H_0 este respinsă (riscul erorii de tipul I este sub pragul de semnificaţie α), iar H_a va fi acceptată.

Vom spune că **sunt diferenţe statistic semnificative** între proporţia celor care dezvoltă IM pentru cele două grupe de femei. Cum $p < 0.01$, putem spune că **diferenţele sunt chiar foarte semnificative din punct de vedere statistic**. Concluzia medicală care se impune este ca aceste diferenţe (implicit riscul respectiv) nu pot fi ignorate, ele nu se datorează doar şansei.

Observaţi însă ce loturi mari au fost implicate în acest studiu pentru a se atinge puterea statistică necesară punerii în evidenţă a acestor diferenţe între proporţii.

Problema 3.3

S-a făcut un studiu de comparare a efectului analgezic al unor produse farmaceutice utilizate de pacienţi cu dureri lombare: ibuprofen (400 mg), codeină (60 mg), codeină (30 mg) şi placebo. Cei 20 de participanţi la studiu au fost distribuiţi aleatoriu în patru grupe (n=5 pentru fiecare grup). La două ore după administrarea produsului, pacienţilor li s-a cerut să descrie senzaţia de îndepărtare a durerii pe o scară între 0 şi 100. Rezultatele sunt prezentate în tabel:

Scorul atribuit			
Ibuprofen	**Codeină**	**Codeină**	**Placebo**
82	80	77	65
89	70	69	75
77	72	67	67
72	90	65	55
92	68	57	63

1. Care sunt **ipotezele statistice**? Formulaţi-le în termeni statistici şi în cuvinte.

2. Ce **test statistic** se aplică pentru a vedea dacă există diferenţe semnificative între efectele celor patru substanţe? Motivaţi-vă decizia.

3. Aplicaţi testul ales şi formulaţi concluzia (în termeni statistici şi medicali).

4. Comparaţi apoi separat doar **Ibuprofen (400 mg)** cu **Placebo**.

Care sunt **ipotezele statistice**? Formulaţi-le în termeni statistici şi în cuvinte.

Aplicaţi testul statistic potrivit şi trageţi concluzia (în termeni statistici şi medicali).

Soluţii 3.3

Ipoteza de zero: toate cele 4 produse farmaceutice au acelaşi efect - media scorurilor va fi aceeaşi (aceeaşi tendinţă centrală)

H_0: $\mu_A=\mu_B=\mu_C=\mu_D$

Ipoteza alternativă: cel puţin unul dintre produse diferă de celelalte - cel puţin o medie diferă de celelalte

H_a: $\mu_i \neq \mu_j$ pentru cel puţin una din cele $C_4^2 = 6$ combinaţii posibile.

Se aplică testul ANOVA (Analiza Variantei) – se compară mediile pentru 4 eşantioane.

După aplicarea testului ANOVA se obţine statistica F=5.16, căreia îi corespunde valoarea $p=0.011 < \alpha=0.05$.

Concluzie: Se respinge H_0 şi vom spune că există diferenţe statistic semnificative în efectul analgezic al celor 4 produse farmaceutice – cel puţin unul dintre cele 3 produse medicamentoase (ibuprofen, codeină-60 mg, respectiv codeină-30mg) are un efect semnificativ diferit de placebo.

Ipoteza de zero: cele 2 produse farmaceutice au acelaşi efect - media scorurilor va fi aceeaşi (aceeaşi tendinţă centrală).

H_0: $\mu_{ibuprofen}=\mu_{placebo}$

Ipoteza alternativă: cele două produse au efecte diferite - mediile vor fi diferite.

H_a: $\mu_{ibuprofen} \neq \mu_{placebo}$

Presupunem că întrebările 9-10 sunt separate de punctele anterioare (pentru simplificare).

Se aplică testul **t- nepereche *2-tailed***: se obţine statistica t=3.547, căreia îi corespunde valoarea p=0.008 < α=0.05.

Concluzie: Se respinge H_0 şi vom spune că există diferenţe statistic semnificative în efectul analgezic al celor 2 produse farmaceutice –ibuprofenul are un efect semnificativ diferit de placebo.

Observaţie (suplimentar):

Dacă suntem riguroşi şi aplicăm un *test post-hoc* (Bonferroni de exemplu - un test t la care α a fost împărţit între cele 6 teste t posibile) nivelul de semnificaţie atins va fi p=0.02 < α=0.05, deci concluzia va fi aceeaşi.

Problema 3.4

Tabelul de mai jos prezintă concentraţia de digoxină în sânge la 4 ore şi la 8 ore după injectarea acesteia intravenos pe un lot de 9 bărbaţi sănătoşi.

Concentraţia de digoxină (mg/l)			
Pers.	4 ore	8 ore	Diferenţe (d_i)
1	1.0	1.0	0.0
2	1.3	1.3	0.0
3	0.9	0.7	-0.2
4	1.0	1.0	0.0
5	1.0	0.9	-0.1
6	0.9	0.8	-0.1
7	1.3	1.2	-0.1
8	1.1	1.0	-0.1
9	1.0	1.0	0.0
	m_1=1.056	m_2=0.99	m_d=-0.067
	s_1=0.15	s_2=0.18	s_d=0.017

1. Ce test se aplică pentru a vedea dacă digoxina se elimină – concentraţia de digoxină măsurată la 4 ore de la injectare diferă semnificativ faţă de cea măsurată la 8 ore? Explicaţi de ce.

2. Formulaţi ipotezele statistice, atât în termeni matematici, cât şi în cuvinte.

3. Aplicaţi testul statistic şi trageţi concluzia (α= .05) ? Formulaţi în termeni statistici şi medicali.

Soluţii 3.4

Se aplică **testul t-pereche** – se verifică existenţa unui fenomen măsurat printr-o variabilă numerică + valorile din cele două eşantioane de valori (la 4 ore şi la 8 ore) se pun în corespondenţă – ne interesează tendinţa diferenţelor dintre valoarea concentraţiei la 4 ore şi cea de la 8 ore.

Tendinţa pentru variabile numerice se exprimă prin medie – deci ipotezele statistice se vor referi la medie**.**

Ipoteza de zero: digoxina nu se elimină – concentraţia în sânge rămâne constantă $\Rightarrow$ pentru populaţie diferenţele sunt nule.

$\mathbf{H_0}$**:** $\mu_d=0$

Ipoteza alternativă: diferenţele nu sunt nule – mai mult, dacă digoxina se elimină, ne aşteptăm ca diferenţele pentru populaţie să fie negative.

Valoarea p=0.01 (sau p=0.02) < α=0.05 reprezintă probabilitatea ca H_0 să fie adevărat în condiţiile în care pentru acest eşantion s-au obţinut valorile precizate.

Concluzie: Se **respinge** $\mathbf{H_0}$ şi vom spune că diferenţele de concentraţie sunt statistic semnificativ diferite de zero.

Concluzia în termeni medicali: Eliminarea digoxinei în intervalul de 4 ore (măsurat la 4 ore, respectiv la 8 ore de la injectare) este statistic semnificativă, deci nu poate fi neglijată.

4. Analiza corelaţiei şi regresiei

Problema 4.1

Obstetricienii obişnuiesc să ceară sumarul de urină pentru a testa nivelul de estriol la femeile gravide care se apropie de termen deoarece s-a descoperit că nivelul de estriol este relaţionat cu greutatea fătului. Testul poate furniza o dovadă indirectă pentru o greutate mică (sub nivelul normal) a fătului.

În 1963, doi obstetricieni (J. Greene şi J. Touchstone) au făcut un studiu în care au înregistrat nivelul de estriol în apropierea termenului, respectiv greutatea copilului la naştere. Tabelul de mai jos prezintă datele publicate de ei, iar pe pagina următoare sunt prezentate rezultate la prelucrarea statistică.

i	Estriol (mg/24 hr) x_i	Greutatea la naştere (g/100) y_i	i	Estriol (mg/24 hr) x_i	Greutatea la naştere (g/100) y_i
1	7	25	17	17	32
2	9	25	18	25	32
3	9	25	19	27	34
4	12	27	20	15	34
5	14	27	21	15	34
6	16	27	22	15	35
7	16	24	23	16	35
8	14	30	24	19	34
9	16	30	25	18	35
10	16	31	26	17	36
11	17	30	27	18	37
12	19	31	28	20	38
13	21	30	29	22	40
14	24	28	30	25	39
15	15	32	31	24	43
16	16	32			

1. Ce tip de studiu este cel prezentat. Motivaţi-vă afirmaţia şi explicaţi care sunt caracteristicile acestui studiu.

2. Formulaţi ipotezele statistice ale acestui studiu în termeni matematici şi în cuvinte.

3. Reprezentaţi grafic datele. Alegeţi cea mai potrivită reprezentare grafică, astfel încât ea să fie în concordanţă cu afirmaţiile făcute la punctele 1 şi 2.

4. Prelucraţi datele din punct de vedere statistic, potrivit afirmaţiilor pe care le-aţi făcut la punctele 1, 2 şi 3.

Indicaţie. Puteţi încerca rezolvarea problemei cu Microsoft-Excel, dar şi cu alte pachete statistice.

5. Interpretaţi rezultatele obţinute. Formulaţi în termeni statistici şi medicali.

Soluţii 4.1

Studiul şi-a propus *analiza corelaţiei* dintre nivelul de estriol la femeile gravide care se apropie de termen şi greutatea nou-născutului. În studiu au fost implicate 31 de viitoare mame şi pentru fiecare dintre ele s-a consemnat valoarea pentru cele două variabile (sub formă de perechi de valori). Interesează investigarea relaţiei dintre cele două variabile numerice şi exprimarea tăriei acesteia printr-un *coeficient de corelaţie*.

Dacă se doreşte să se poată face o predicţie a greutăţii nou-născutului (respectiv a fătului) în funcţie de nivelul de estriol, atunci se face *analiza regresiei* – nivelul de estriol va fi considerat ca fiind variabila independentă. De obicei, pentru două variabile numerice (cum este cazul aici), se investighează o posibilă relaţie liniară exprimabilă printr-o *dreaptă de regresie*. Panta acestei drepte se numeşte *coeficient de regresie* şi va permite calculul greutăţii fătului în funcţie de nivelul de estriol, chiar pentru valori ale acestuia care nu au apărut în studiul iniţial (dar între limitele min. şi max. din studiul făcut).

Din punctul de vedere al tipurilor de studii epidemiologice, nu ni se dau suficiente informaţii ca să putem identifica tipul de studiu. Am putea presupune că este un *studiu transversal*.

Ipotezele statistice sunt:

H_0: Ipoteza de zero: cele două variabile sunt independente - nu există nici o relaţie între nivelul de estriol şi greutatea nou-născutului

H_0**:** $\rho=0$

H_a: Ipoteza alternativă: cele două variabile sunt corelate – există o relaţie între nivelul de estriol şi greutatea nou-născutului

H_a: $\rho\neq0$

Grafic de tip "*scatter*" (nor de puncte) – variabila independentă pe abscisă (Figura 1).

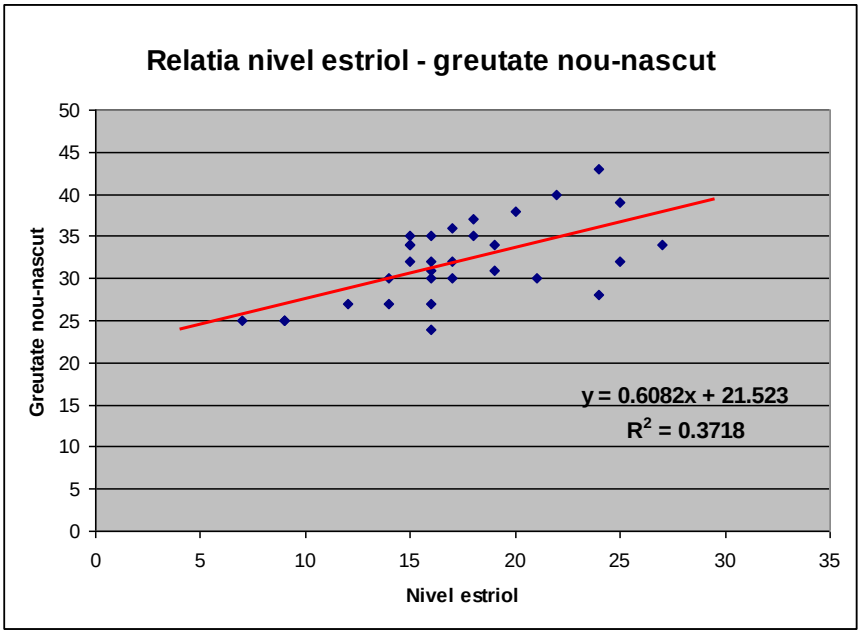

Figura 1. Cea mai potrivită reprezentare grafică este cea de tip „*scatter*", care scoate în evidenţă relaţia dintre cele doua variabile.

În Microsoft-Excel coeficientul de corelaţie se poate determina cu funcţia CORREL sau cu *Tools → Data analysis → Regression.*

Coeficientul de corelaţie Pearson este **r** = 0.6097 cu N=31-2=29 grade de libertate.

Dacă se face calculul statisticii t se obţine:

r= 0.609731
t= 4.142656
p= 0.000271

În Microsoft-Excel:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$ adică t=r*SQRT(7/(1-r*r))

p<0.001 s-a obţinut cu funcţia TDIST.

Rezultat similar se obţine şi cu *Tools → Data analysis → Regression.*

O altă alternativă de rezolvare - se utilizează comanda *Regress* în programul *Analysis* din **EpiInfo**. În acest caz se obţine un interval de încredere (cu 1-α=0.95) pentru coeficientul de corelaţie şi se interpretează acest interval.

Judecând doar după **r**, putem spune că există o corelaţie directă (pozitivă) şi relativ puternică (**r** peste 0.5) – corespunzătoare unei corelaţii puternice între nivelul de estriol şi greutatea nou-născutului.

Pentru o discuţie mai nuanţată şi pentru a ne pronunţa asupra semnificaţiei statistice, trebuie luată în considerare **valoarea lui p** (indiferent prin ce metodă s-a obţinut).

Concluzie: corelaţia între nivelul de estriol şi greutatea nou-născutului este extrem de semnificativă din punct de vedere statistic (p<0.001).

Din punct de vedere medical, există dovezile că nivelul de estriol al mamei la apropierea termenului este corelat direct şi semnificativ cu greutatea fătului. Cunoscând nivelul de estriol, se poate estima greutatea fătului şi pot fi luate din timp măsurile medicale necesare.

Problema 4.2

Un grup de cercetători şi-au propus să investigheze relaţia dintre nivelul plasmatic de amfetamină şi psihoza indusă de această substanţă. În studiul făcut au fost cuprinşi 10 consumatori cronici, cărora li s-a determinat nivelul plasmatic de amfetamină, determinare urmată imediat de o evaluare psihiatrică (în urma

căreia s-a obţinut un scor al intensităţii psihozei). Datele rezultate sunt sintetizate în tabelul următor:

Pacient	Intensitatea psihozei (Y)	Nivelul de amfetamină (mg/ml) (X)
1	10	150
2	30	300
3	20	250
4	15	150
5	45	450
6	35	400
7	50	425
8	15	200
9	40	350
10	55	475

1. Reprezentaţi grafic datele din acest studiu – alegeţi tipul de grafic care ilustrează cel mai bine relaţia dintre cele două variabile din studiu.
2. Determinaţi coeficientul de corelaţie Pearson.
3. Interpretaţi valoarea obţinută - trageţi concluzia pentru studiul în discuţie.
4. Este această corelaţie statistic semnificativă? Argumentaţi.

Soluţii 4.2

Se face un grafic de tip "*scatter*" (nor de puncte) – variabilă independentă pe abscisă (Figura 2).

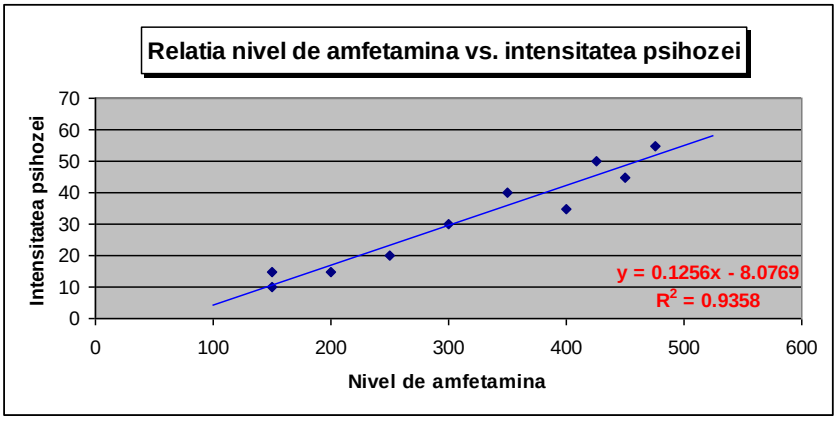

Figura 2. Cea mai potrivită reprezentare grafică este cea de tip „*scatter*", care scoate în evidenţă relaţia dintre cele doua variabile.

În Microsoft-Excel coeficientul de corelaţie se poate determina cu funcţia CORREL sau cu *Tools → Data analysis → Regression.*

Coeficientul de corelaţie Pearson este **r = 0.97** cu N=10-2=8 grade de libertate.

Judecând doar după **r**, putem spune că există o corelaţie directă (pozitivă) şi puternică (**r** aproape de valoarea 1 – corespunzătoare unei corelaţii "perfecte") între nivelul plasmatic de amfetamină şi intensitatea psihozei.

Pentru o discuţie mai nuanţată – punctul 2.4.

Dacă se face calculul statisticii t se obţine:

r= 0.967381949

t= 10.80114078

p= $4.76491*10^{-6}$

În Microsoft-Excel:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$ adica t=r*SQRT(8/(1-r*r))

p=4.76491E-06 obtinut cu functia TDIST

Concluzie: corelaţia între nivelul plasmatic de amfetamină şi intensitatea psihozei este extrem de semnificativă din punct de vedere statistic (p<0.0001).

Rezultat similar se obţine şi cu *Tools → Data analysis → Regression.*

O altă alternativă de rezolvare - se utilizează comanda *Regress* în secţiunea *Analysis* din **EpiInfo**. În acest caz se obţine un interval de încredere (cu 1-α=0.95) pentru coeficientul de corelaţie şi se interpretează acest interval.

5. Studii observaţionale. Analiza riscului

Problema 5.1

Într-un studiu privind frecvenţa de apariţie a efectelor secundare la un medicament anti-artritic, s-au observat două grupuri de pacienţi: unii care urmau tratament cu noul medicament şi alţii care nu urmau acest tratament. Au fost observaţi 400 de pacienţi selectaţi dintr-un spital. Rezultatele sunt prezentate în tabelul ce urmează.

TRATAMENT	EFECTE SECUNDARE	FARA EFECTE SECUNDARE	**TOTALURI**
Nu au urmat tratament	39	271	310
Au urmat tratament	7	83	90
TOTALURI	46	354	**400**

1. Ce tip de studiu a fost utilizat? Explicaţi care sunt caracteristicile acestui studiu.
2. Calculaţi **RR** de apariţie a efectelor secundare, asociate cu medicamentul anti-artritic aflat în studiu.
3. Interpretaţi **RR** calculat la punctul anterior – luaţi în considerare intervalul de încredere pentru RR şi exprimaţi concluzia din punct de vedere medical.

Soluţii 5.1.

Studiul este de tip cohort prospectiv (longitudinal; "*follow-up*") – în studiu au fost implicate 2 grupuri de subiecţi (unii "expuşi" tratamentului în discuţie, ceilalţi nu) care au fost urmăriţi un timp pentru a se depista eventualele efecte secundare. S-au numărat cazurile cu probleme adverse pentru ambele grupuri şi s-a determinat riscul de apariţie a efectelor asociate cu boala artritică la cei care au urmat tratamentul în discuţie în comparaţie cu pacienţii care nu au urmat acel tratament.

Dacă se face calculul utilizând programul Epi-STATCALC – tabelul de contingenţă va arăta astfel:

Tratament - *Exposure*		**Efecte secundare - *Disease***		
		(+)	**(-)**	**Totaluri**
	(+)	7	83	90
	(-)	39	271	310
	Totaluri	46	354	**400**

Se obţine:

$RR = 0.62$

Intervalul de estimare pentru riscul relativ este:

$0.29 < RR < 1.33$

$1-\alpha = 0.95$

Intervalul de estimare îl conţine pe 1 $\Rightarrow$ nu putem spune că tratamentul în discuţie ar implica un risc mai crescut sau mai scăzut în apariţia efectelor secundare la pacienţii ce suferă de artrită şi urmează tratamentul anti-artritic în discuţie.

(RR se defineşte ca fiind raportul între incidenţa efectelor secundare pentru cei ce urmează tratamentul în discuţie şi incidenţa efectelor pentru cei care nu urmează tratamentul respectiv, urmând probabil alte tratamente).

6. Probleme propuse

Până în anii 1970, emisia de compuşi de plumb în aer nu era considerată o problemă de sănătate publică. O contribuţie importantă în evidenţierea efectelor poluării cu plumb a adus-o un studiu de la începutul anilor 1970 făcut în Texas pentru a investiga efectele expunerii la compuşi de plumb asupra funcţiilor neurologice şi psihologice [1, 2]. Datele au fost furnizate de către profesorul Bernard Rosner [3], căruia i-au fost date de către dr. Philip Landrigan [4].

Pe scurt, a fost studiat un grup de copii care au locuit în apropierea topitoriei de plumb din El Paso, Texas, cărora li s-a determinat nivelul de plumb din sânge. A fost determinat un grup de 46 copii "expuşi" care aveau nivelul de plumb în sânge ≥ 40 μg/ml în 1972 (pentru câţiva dintre ei determinarea s-a făcut în 1973). În acelaşi timp, cei 78 de copii cu nivel de plumb în sânge < 40 μg/ml (atât în 1972 cât şi în 1973) au format grupul de control.

[1] Landrigan, PJ, Whitworth, RH, Baloh, RW, Staehling, NW, Barthel, WF, Rosenblum, BH (1975). Neuropsychological dysfunction în children with chronic low-level lead absorption. *Lancet* 1: 708-715.

[2] Morse, DL, Landrigan, PJ, Rosenblum, BH, Hubert JS, Housworth J (1979). El Paso revisited. Epidemiologic follow-up of an environmental lead problem. JAMA 249 (8). http://jama.ama-assn.org/cgi/content/abstract/242/8/739 (Last access 13th April 2008).

[3] Bernard, R (2000). *Fundamentals of biostatistics* (5th Ed.). Duxbury: Pacific Grove.

[4] CDC. epidemic Intelligence Service. http://www.cdc.gov/EIS/about/landrigan.htm (Last access 13th April 2008).

O parte din datele obţinute în acest studiu sunt furnizate în fişierul **lead_hw2345.xls** ce se găseşte pe CD-ul ce însoţeşte această carte. Semnificaţia câmpurilor este:

Id – codul de identificare;

Area – distanţa faţă de topitorie

1=0-1 mile;

2=1-2.5 mile;

3=2.5-4.1 mile.

Sex – sexul (1=masculin, 2=feminin);

Lead_type – descrie cele trei grupuri de copii în funcţie de nivelul de plumb în sânge în 1972 şi 1973:

1 – grupul de CONTROL, cu nivel de plumb în sânge < 40 μg/ml atât în 1972, cât şi în 1973;

2 – grupul EXPUS CURENT, cu nivelul de plumb în sânge ≥ 40 μg/ml în 1973;

3 – grupul EXPUS ÎN TRECUT, cu nivelul de plumb în sânge ≥ 40 μg/ml în 1972, dar cu nivel normal în 1973.

Pica şi Colic – exprimă prezenţa simptomelor de pica şi colici (1=da, 2=nu);

FWT_R şi FWT_L - performanţa copiilor la un test de tipul "*finger-wrist tapping*";

MAXFWT – scorul obţinut pentru mâna dominantă (maximul dintre FWT_R şi FWT_L);

Age_y – vârsta exprimată în ani;

Group – identifică două grupuri de intoxicaţie:

0=neintoxicaţi;

1=intoxicaţi (cei pentru care Lead_type este 2 sau 3);

Acolo unde nu apar valori, înseamnă că datele lipsesc.

Întrebările 1-5

Distribuţia valorilor pentru vârstă este prezentată în tabelul de mai jos, unde intervalele sunt considerate închise la dreapta şi deschise la stânga:

Vârsta ani	Frecvenţe absolute	Frecvenţe relative	Frecvenţe cumulate%
0 - 2	0	0.000	0.00%
2 - 4	8	0.065	6.45%
4 - 6	A	0.153	B
6 - 8	31	C	46.77%
8 - 10	21	0.169	63.71%
10 12	14	0.113	75.00%
12 - 14	19	0.153	90.32%
14 - 16	12	0.097	D
TOTAL	**124**	**1**	

Graficul din Figura 3sintetizează distribuţia valorilor.

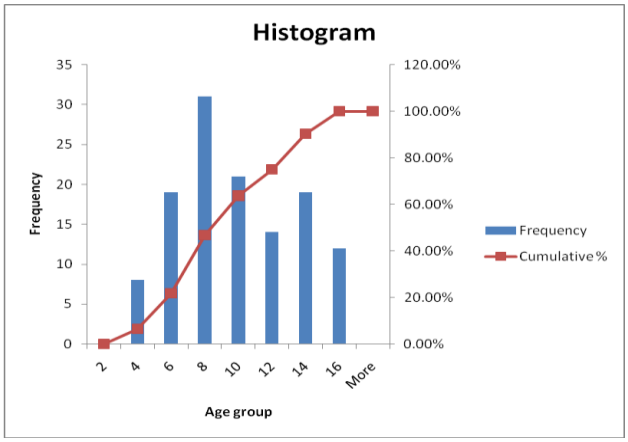

Figura 3. Distribuţia valorilor de vârstă pentru datele din fişierul **lead_hw2345.xls**

1. Valoarea **A** din tabel este:
2. Valoarea **B** din tabel este:
3. Valoarea **C** din tabel este:
4. Valoarea **D** din tabel este:
5. Probabilitatea ca un individ selectat la întâmplare din acest eşantion de 124 de copii să aibă vârsta între 6 şi 12 ani este:

Întrebările 6 - 8

Pentru următoarele trei întrebări, consideraţi eşantionul reprezentativ pentru zona respectivă.

6. Pentru un nivel de încredere $1-\alpha=0.68$, determinaţi intervalul de încredere pentru vârsta copiilor din zona respectivă. Interpretaţi rezultatul obţinut (formulaţi concluzia în cuvinte).

7. Estimaţi media pentru vârsta copiilor din zona respectivă, cu $1-\alpha=0.95$. Interpretaţi rezultatul obţinut (formulaţi concluzia în cuvinte).

8. Estimaţi procentul de copii care au vârsta sub 5 ani (cu aproximaţie).

Întrebările 9-13

9. Pentru un nivel de încredere $1-\alpha=0.95$, estimaţi media pentru vârsta copiilor de sex feminin. Interpretaţi rezultatul obţinut (formulaţi concluzia în cuvinte).

10. Estimaţi media vârstei pentru pacienţii de sex masculin, cu $1-\alpha=0.95$. Interpretaţi rezultatul obţinut (formulaţi concluzia în cuvinte).

11. Ce test statistic se aplică pentru a vedea dacă există diferenţe între vârstele copiilor de cele două sexe? Formulaţi ipotezele statistice.

12. Aplicaţi testul şi trageţi concluzia statistică (α=0.05).

13. Formulaţi concluzia în termeni medicali.

Întrebările 14 - 15

Investigaţi asocierea dintre intoxicaţia cu plumb şi prezenţa colicilor la copiii din studiul descris. Estimaţi riscul pe care-l constituie intoxicaţia cu plumb (descrisă prin variabila GROUP) pentru apariţia colicilor (descrisă prin variabila COLIC).

Informaţii ajutătoare: în grupul copiilor neintoxicaţi cu plumb (GROUP=0) au fost 78 de copii, din care 11 au prezentat colici; în grupul copiilor intoxicaţi (GROUP=1) au fost 46 de copii, din care 12 au prezentat colici.

14. Ce tip de studiu statistic este recomandat pentru această investigaţie?

15. Investigaţi asocierea dintre intoxicaţia cu plumb şi prezenţa colicilor.

15.1. Formulaţi ipotezele statistice (matematic şi în cuvinte);

15.2. Precizaţi ce test statistic aplicaţi şi de ce;

15.3. Aplicaţi testul ales;

15.4. Formulaţi decizia statistică;

15.5. Formulaţi concluzia medicală în cuvinte.

Întrebările 16-17

Investigaţi performanţa copiilor la testul "*finger-wrist tapping*" (FWT), exprimată de variabila numită MAXFWT, pentru cele trei grupuri de copii identificaţi prin variabila LEAD_TYPE.

16. Determinaţi intervalele de încredere pentru estimarea mediei performanţei pentru fiecare grup în parte.

17. Comparaţi cele trei grupuri din punctul de vedere al performanţei la testul FWT (variabila MAXFWT). Analizaţi atât din punct de vedere statistic, cât şi al interpretării medicale a rezultatelor statistice:

17.1. Formulaţi ipotezele statistice (matematic şi în cuvinte);

17.2. Precizaţi ce test statistic aplicaţi şi de ce;

17.3. Aplicaţi testul ales;

17.4. Formulaţi decizia statistică;

17.5. Formulaţi concluzia medicală în cuvinte.

Întrebările 18-19

Reluaţi investigarea asocierii dintre intoxicaţia cu plumb şi prezenţa colicilor la copiii din studiul descris. Estimaţi riscul pe care-l constituie intoxicaţia cu plumb (descrisă prin variabila GROUP) pentru apariţia colicilor (descrisă prin variabila COLIC).

[Informaţii ajutătoare: în grupul copiilor neintoxicaţi cu plumb (GROUP=0) au fost 78 de copii, din care 11 au prezentat colici; în grupul copiilor intoxicaţi (GROUP=1) au fost 46 de copii, din care 12 au prezentat colici.]

18. Rediscutaţi tipul de studiu statistic recomandat pentru această investigaţie, într-un context mai complex (luând în considerare şi studiile epidemiologice).

19. Estimaţi riscul pe care-l constituie intoxicaţia cu plumb pentru apariţia colicilor:

19.1. Formulaţi ipotezele statistice pentru această abordare;

19.2. Aplicaţi metoda pe care aţi ales-o şi precizaţi rezultatul obţinut;

19.3. Trageţi concluzia statistică;

19.4. Formulaţi concluzia medicală în cuvinte.

Întrebările 20-22

Similar cu modul în care aţi investigat asocierea dintre intoxicaţia cu plumb şi prezenţa colicilor la copiii din studiul descris (atât în tema anterioară, cât şi în cea prezentă), investigaţi şi asocierea intoxicării (descrisă prin variabila GROUP) cu prezenţa simptomelor de pica (descrisă prin variabila PICA).

20. Ce tip de studiu statistic este recomandat pentru aceasta investigaţie?

21. Investigaţi asocierea dintre intoxicaţia cu plumb şi prezenţa simptomelor de pica.

22. Estimaţi riscul pe care-l constituie intoxicaţia cu plumb pentru apariţia simptomelor de pica.

7. Concluzii

În această şedinţă aţi învăţat noţiuni de bază privind prelucrările statistice ale datelor biomedicale:

- descrierea datelor obţinute în urma unor studii experimentale sau observaţionale;
- estimarea statistică;
- testarea statistică a ipotezelor de cercetare;
- analiza regresiei şi corelaţiei;
- studii epidemiologice, estimarea riscului.

Referințe

Vernic CV, Apostol SA, Frandes M, Mada L, Lungeanu D. APLICATII PRACTICE DE INFORMATICA SI BIOSTATISTICA MEDICALA IN NURSING. Editura Eurostampa, ISBN 978-606-32-0487-6, Colectia Derzelas, 2017:1-216

Vernic CV, Mada L, Lungeanu D, Muntean C, Apostol SA, Catu CO, Ursoniu S. Aplicaţii practice de Informatică Medicală şi Biostatistică. Editura Victor Babes, ISBN 978-606-8054-09-4, 2010: 1-229.

Mihalaş GI, Lungeanu D. Informatică medicală şi biostatistică. Timişoara: EVB, 2009.

Sheskin DJ. Handbook of parametric and nonparametric statistical procedures (3rd Edition.). Boca Raton: Chapman & Hall/CRC, 2004.

Rosner B. Fundamentals of biostatistics (5th Edition). Pacific Grove: Duxbury. Thomson Learning, 2000.

Knapp RG, Clinton Miller M: Clinical epidemiology and biostatistics. Baltimore: Williams & Wilkins, 1992.