Universitatea de Medicină şi Farmacie Tîrgu Mureş
Disciplina de Informatică Medicală

Marius Măruşteri

Biostatistică

aplicaţii practice şi exerciţii recapitulative pentru studenţii Şcolii Doctorale

- 2005 -

Biostatistică

Biostatistica este o ramură a statisticii, specializată în studiul fenomenelor biologice, deci şi al celor medicale. Se ocupă de culegerea, centralizarea şi gruparea datelor, precum şi de prelucrarea şi determinarea unor indicatori pentru descrierea fenomenelor biomedicale studiate, pe baza evidenţierii unor regularităţi sau variabilităţi statistice. Totodată aplică şi dezvoltă tehnici statistico-probabilistice pentru analiza datelor biomedicale.

Începuturile biostatisticii au fost determinate de nevoia obţinerii unor informaţii cantitative dintre cele mai simple, formulate de regulă sub forma „câţi bolnavi ? ” , „câţi decedaţi ? ” , etc. Cu timpul s-a constatat însă că asemenea metode sunt insuficiente pentru caracterizările fenomenelor, că există o variaţie în răspunsurile care se obţin între diverse măsurători sau, cu alte cuvinte, că fenomenele biologice sunt caracterizate prin variabilitate. Dar şi în aceste condiţii, observându-se serii lungi de măsurători, s-a descoperit că se pot calcula indicatori simpli cu mare putere de sinteză, cum ar fi media (aritmetică, geometrică, etc), dispersia, etc.

Într-o etapă ulterioară, statistica a câştigat în puterea de analiză a fenomenelor. Pe această cale s-au descoperit legile care guvernează ceea ce înainte părea întâmplător. Această etapă, în care statistica trece de la descrierea fenomenelor la analiza lor, se caracterizează prin aplicarea în general a unui aparat matematic din ce în ce mai complicat şi a calculului probabilităţilor în special.

Indicatori statistici

Principalii indicatori care caracterizează un şir de date sunt fie indicatori de tendinţă centrală, fie indicatori ce caracterizează împrăştierea datelor în jurul unei valori medii.

O serie de date este alcătuită dintr-un şir de valori pe care le notăm :

$x_1,\ x_2,\ \dots\ , x_n$.

Indicatorii matematici mai importanţi ce caracterizează o serie de date sunt:

Media aritmetică - notată de regulă cu $\bar{x} = \dfrac{x_1 + x_2 + \dots + x_n}{n}$

Mediana - este acea valoare din şirul de date care împarte în două părţi egale şirul ordonat de valori (atenţie, şirul este ordonat crescător), situându-se la mijlocul seriei statistice. Dacă numărul de valori n este un număr impar, atunci mediana este valoarea $M_e = x_k$, unde $k = \dfrac{n}{2} + 1$. Dacă n este par, deci avem un număr par de valori, mediana este definită ca fiind $M_e = \dfrac{x_k + x_{k+1}}{2}$ unde

$k = n/2$.

Modul - constituie valoarea care apare cel mai des, deci valoarea cu numărul cel mai mare de apariţii.

Amplitudinea - este diferenţa dintre valoarea maximă şi cea minimă

$A = A_{max} - A_{min}$.

Amplitudinea relativă - notată $A\%$ este raportul dintre amplitudinea absolută şi media aritmetică a seriei de date.

Dispersia (varianța) notată s^2_x este un indicator de împrăştiere a datelor. Formula de calcul este: $s_x^2 = \frac{\sum x_i^2 - (\bar{x})^2}{n-1}$.

Abaterea standard sau deviația standard reprezintă rădăcina pătrată din varianță (dispersie) : $s_x = \pm\sqrt{s_x^2}$

Coeficientul de variație se calculează ca un raport procentual între abaterea standard şi valoarea medie a şirului de valori.

$$C.V.\% = \frac{s_x}{\bar{x}} \cdot 100$$

De remarcat că valoarea coeficientului de variație nu are unitate de măsură, se exprimă procentual. Acest fapt permite folosirea indicatorului la compararea a două sau mai multe serii de date, indiferent de ordinul de mărime al variabilelor (variantelor) şi de unitățile de măsură folosite. Se poate considera că un coeficient de variație sub 10% indică o dispersie mică (o împrăştiere), adică seria este omogenă. Un coeficient între 10% şi 30% indică dispersie mijlocie, iar peste 30% indică dispersie mare. Dacă dispersia este mare, media nu este un indicator reprezentativ.

Atunci când avem foarte multe date se recomandă includerea lor în clase egale ca mărime, ceea ce uşurează mult prelucrările statistice ulterioare. Spre exemplu sortăm pacienții pe grupe de vârstă: 21-24 de ani, 25-30 ani, etc… În acest caz apare noțiunea de frecvență a clasei.

Indicatori statistici pentru serii de date cu apariții frecvente ale aceleiaşi valori

Dacă datele pe care le studiem conțin valori care se repetă des, se obişnuieşte să se grupeze datele care au aceeaşi valoare . Numărul de apariții ale unei valori anume se numeşte frecvența de apariție şi se notează cu f_i.

Presupunem că în urma măsurătorilor am obținut şirul de valori:
x_1 cu frecvența f_1, x_2 cu frecvența f_2, … x_n cu frecvența f_n
Indicatorii statistici se calculează conform noilor formule:

Media aritmetică

$$\bar{x} = \frac{\sum\limits_{i=1,n} x_i \cdot f_i}{\sum\limits_{i=1,n} f_i} = \frac{x_1 \cdot f_1 + x_2 \cdot f_2 + ... + x_n \cdot f_n}{f_1 + f_2 + ... + f_n}$$

Mediana – este x_k unde $k = \frac{\sum\limits_{i=1,n} f_i + 1}{2}$

Dispersia (varianța) : $s_x^2 = \dfrac{\sum_{i=1,n}(x_i - \bar{x})^2 \cdot f_i}{\sum_{i=1,n} f_i - 1}$

Teste statistice

Pentru a stabili dacă există o legătură între două serii de date (deci între două variabile cantitative) sau între două variabile calitative se folosesc testele statistice.

Cele mai cunoscute sunt :

✓ testul **Student** - pentru compararea mediilor unei caracteristici la două populații.

✓ testul **Chi** - pentru a verifica dacă există o asociere sau o legătură semnificativă din punct de vedere statistic între două variabile calitative.

✓ testul **Z -** pentru compararea a două proporții (deci pentru compararea unor caracteristici calitative).

Testul STUDENT

Testul Student este utilizat în analiza statistică pentru compararea mediei unei caracteristici la două populații. Caracteristica studiată trebuie să fie o caracteristică cantitativă, măsurabilă.

Etapele aplicării testului STUDENT

Pentru aplicarea testului Student se parcurg următoarele etape:

- Se stabilesc două eşantioane de lucru: un grup de test extras din prima populație şi un grup martor, extras din a doua populație. Se culeg şi se înregistrează datele studiului. Se fac următoarele notații:

  X_i reprezintă valorile înregistrate în grupul de test

  Y_i reprezintă valorile înregistrate în grupul martor

  $\bar{X}$ reprezintă media caracteristicii în grupul de test

  $\bar{Y}$ reprezintă media caracteristicii în grupul martor

  n_1 reprezintă numărul de subiecți din grupul de test

  n_2 reprezintă numărul de subiecți din grupul martor

  s_1 reprezintă deviația standard în grupul de test

  s_2 reprezintă deviația standard în grupul martor

- Se formulează două ipoteze:

1. Ipoteza nulă (H_0) afirmă: „*media μ_1 a caracteristicii în populația din care face parte grupul de test este egală cu media μ_2 a caracteristicii în populația din care face parte grupul martor ($\mu_1 = \mu_2$)* ”

2. Ipoteza alternativă (H_1) afirmă: „*media* μ_1 *a caracteristicii în populația din care face parte grupul de test este diferită de media* μ_2 *a caracteristicii în populația din care face parte grupul martor (*$\mu_1 \neq \mu_2$*)*”

- Se calculează valoarea statistică a testului Student după formula:

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

unde s_p este:

$$s_p^2 = \frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 + n_2 - 2}$$

- Se calculează numărul de grade de libertate a testului Student:

$gl = n_1 + n_2 - 2$.

- Se calculează valoarea probabilității *p*. Probabilitatea *p* este probabilitatea să obținem întâmplător o valoare statistică egală sau mai mare decât valoarea *t* calculată, în condițiile în care ipoteza nulă H_0 este adevărată. În cazul în care probabilitatea *p* calculată este $\leq 0,05$ se respinge ipoteza H_0.

- Stabilirea concluziei testului Student.
  - Dacă probabilitatea *p* are o valoarea mai mică sau egală cu 0,5, atunci se repinge ipoteza nulă H_0 şi se acceptă ipoteza alternativă H_1, prin urmare există o diferență semnificativă între mediile caracteristicii în cele două populații.
  - Dacă probabilitatea *p* are o valoarea mai mare decât 0,5 , atunci se acceptă ipoteza nulă H_0, prin urmare nu există o diferență semnificativă între mediile caracteristicii în cele două populații.

Utilizarea funcției TTEST a utilitarului Excel

Testul STUDENT este mult mai uşor de aplicat cu ajutorul utilitarului *EXCEL* decât testul CHI. Pentru a obține probabilitatea finală *p* introducem valorile celor două serii pe o foaie de calcul. Funcția TTEST se introduce într-o celulă oarecare specificând în ordine:

-zonele care conțin datele celor două serii de valori

-valoarea: 1 sau 2 – pentru a indica dacă testul este cu un capăt sau cu două capete. Dacă testul este cu două capete, atunci în cazul respingerii ipotezei H_0 se consideră că există diferențe între mediile celor două caracteristici fără a se specifica care dintre cele două medii este mai mare. Dacă testul este cu un capăt, atunci în cazul în respingerii iptotezei H_0 este clar care dintre mediile celor două populații este mai mare. Cel mai des se utilizează testul cu 2 capete.

-tipul testului: 1, 2 sau 3

- **1** - dacă grupurile de date sunt dependente
- **2** - dacă grupurile de date sunt independente şi se presupune că populațiile au aceeaşi dispersie.

- **3** - dacă grupurile de date sunt independente şi se presupune că populaţiile au dispersii diferite.

Spre exemplu, dacă seriile de valori sunt conţinute în zonele B7:B25 şi E8:E35 şi grupurile de date sunt independente, atunci conţinutul funcţiei TTEST este:

= TTEST(B7:B25; E8:E35; 2; 2).

Exerciţii

1. Pentru stabilirea cantităţii de adenină dintr-o soluţie dată, s-au făcut măsurători spectrofotometrice şi s-au obţinut următoarele date:

Număr măsurătoare	Valoarea măsurată
1	64
2	71
3	73
4	82
5	87
6	95
7	100
8	101
9	102
10	105

Să se calculeze, cu ajutorul utilitarului Excel, valoarea medie, dispersia, amplitudinea, abaterea standard, amplitudinea relativă, coeficientul de variaţie.

2. Să se calculeze greutatea medie a 100 de copii născuţi la termen a căror greutate la naştere a fost următoarea:

Greutatea (în grame)	Frecvenţa
2800	10
2900	20
3000	40
3100	20
3200	10

De asemenea să se calculeze, cu ajutorul utilitarului Excel, mediana, modul, amplitudinea, amplitudinea relativă, dispersia şi coeficientul de variaţie.

3. Să se calculeze valoarea medie, amplitudinea, amplitudinea relativă, dispersia, abaterea standard şi coeficientul de variaţie al duratei de spitalizare în cazul unui grup de 200 de bolnavi internaţi cu hepatită virală. Datele sunt prezentate în tabelul următor:

Durata de spitalizare (zile)	Frecvența
20	2
22	6
24	10
26	18
28	30
30	80
32	26
34	10
36	8
38	6
40	4

Testul STUDENT

4. Într-un studiu al efectului bumetamidei în secreția de calciu în urină, 9 persoane alese aleator au primit fiecare câte o doză de 0,5 mg de medicament. S-a colectat în fiecare oră, timp de 6 ore, urina de la cele 9 persoane. La fel s-a procedat cu alte 10 persoane care nu au primit medicamentul. Pentru fiecare persoană s-a calculat o medie (prin calculul mediei celor 6 valori citite).

Datele obținute au fost următoarele:

Grupul de test	Grupul de control
2	3
4	4,5
5	5
3,5	6
7	6,5
10,5	6,5
16	7,5
18	8
1,5	8,5
	9,5

Să se determine dacă secreția de calciu în urină diferă la cele două grupuri, deci dacă administrarea medicamentului are efect în creşterea secreției de calciu. Pentru a realiza acest lucru, mai întâi introduceți datele de mai sus într-o foaie de calcul tabelar şi apoi aplicați testul Student cu un capăt (1 tails), de tipul 2 (two-sample equal variance) şi ipoteza 0. Dacă probabilitatea P obținută este mai mică decât 0,05 atunci medicamentul are efect.

Ipoteze:

H_0: medicamentul nu are efect în creşterea secreției de calciu.

H_1: medicamentul are efect în creşterea secreției de calciu.

$p > 0,05 \Rightarrow$ acceptăm iptoteza H_0

Rezultate

$p=0,31317 \Rightarrow$ acceptăm iptoteza H_0 , deci medicamentul nu are efect în creşterea secreției de calciu.

5. Se efectuează un studiu al nivelului de digoxin ser, după efectuarea rapidă a unei injecții intravenoase cu acest medicament. Să se stabilească dacă nivelul de digoxin ser la 4 ore după injectare diferă semnificativ de nivelul de la 8 ore după injectare. Datele obținute în urma studiului pe 10 subiecți sunt următoarele:

Nr. subiect	După 4 ore	După 8 ore
1	1	1
2	1,3	1,3
3	0,9	0,7
4	1	1
5	1	0,9
6	0,9	0,8
7	1,3	1,2
8	1,1	1
9	1	1
10	1,3	1,2

Pentru a obține rezultatul studiului, aplicați testul Student cu două capete şi de tipul 1 (grupuri dependente). Dacă probabilitatea *p* obținută este mai mică decât 0,05 atunci există diferențe semnificative.

6. Concentrația hemoglobinei în g/100 ml sânge, la un număr de 12 persoane cu anemie feriprivă, a crescut după tratament astfel:

Persoana	Hemoglobina (g./100 ml sânge)	
	Înainte de tratament	După tratament
1	3,4	4,9
2	3,0	2,3
3	3,0	3,1
4	3,4	2,1
5	3,7	2,6
6	4,0	3,8
7	2,9	5,8
8	2,9	7,9
9	3,1	3,6
10	2,8	4,1
11	2,8	3,8
12	2,4	3,3

Se poate afirma că tratamentul este eficace ? Pentru a putea răspunde la această întrebare utilizați testul STUDENT cu două capete şi pentru grupuri dependente (tipul 1) .

7. S-a măsurat glicemia la un lot de 5 persoane sănătoase, alese aleator. Apoi s-a măsurat glicemia la un lot de 8 persoane alese de asemenea aleator, dar bolnave de diabet zaharat. Rezultatele obținute sunt prezentate în tabelul următor.

Persoane	Sănă-toase	Bolnave de diabet
1	100	171
2	101	172
3	103	175
4	106	176
5	110	177
6		178
7		182
8		185

Să se stabilească dacă mediile celor două loturi diferă semnificativ, cu un risc de 0,05. Se va utiliza testul Student cu două capete, de tipul 2 (independente).

8. S-a măsurat urimia la două loturi de câte 10 bolnavi de gută, dintre care unii au fost tratați şi alții nu şi s-au obținut următoarele rezultate:

Nr. Subiect	Valoarea urimiei în lotul tratat (mg/l)	Valoarea urimiei în lotul netratat
1	42	48
2	45	54
3	48	60
4	52	66
5	55	72
6	58	78
7	60	84
8	63	90
9	67	96
10	70	102

Să se aprecieze cu un risc de 0,05 dacă medicamentul a avut efect. Se va aplica testul STUDENT cu două capete, de tipul 2 (grupuri independente).

9. La un lot de bolnavi cu hepatită cronică s-a efectuat proba Tymol şi apoi li s-a aplicat o rație alimentară hipercalorică, după care s-a repetat proba Tymol.

Rezultatele probelor sunt exemplificate în tabelul următor.

	Proba Tymol	
Bolnavul	Înainte	După rație
1	10	8
2	8	8
3	16	10
4	5	5
5	6	4
6	12	7
7	9	8
8	10	14

9	14	10
10	10	6

Să se stabilească dacă mediile celor două serii de date diferă semnificativ. Se va utiliza testul Student cu două capete, de tipul 1 (pentru grupuri dependente).

10. S-a măsurat valoarea sistolică la grup de pacienți dignosticați cu stenoză şi la un grup de pacienți asimptomatici. Să se determine dacă media celor două serii de date diferă seminificativ. Se va utiliza testul STUDENT cu două capete şi de tipul 2 (pentru grupruri independente).

Rezultatele măsurătorilor sunt exemplificate în tabelul următor.

	Valoarea sistolică	
Pacientul	**Simptomatici**	**Asimpto-matici**
1	160	150
2	155	160
3	170	155
4	170	150
5	170	150
6	185	155
7	190	165
8	195	165
9	205	165
10	210	170
11	210	175
12	220	175
13	220	180

Statistică Matematică

Testul CHI

estul **CHI** este utilizat în analiza statistică în următoarele cazuri:

T

➲ în studiile epidemiologice pentru identificarea unei **asocieri** între un **factor de risc** şi o **boală.** De exemplu, se poate aplica testul CHI pentru stabilirea unei eventuale legături între fumat şi moartea prematură ca urmare a unei boli cardiovasculare, sau a unei legături între expunerea la o anumită substanță chimică şi apariția malformațiilor congenitale la inimă.

➲ pentru a verifica o asociere semnificativă din punct de vedere statistic între două caracteristici calitative, cu alte cuvinte pentru stabilirea unei diferențe între proporții. De exemplu, se poate aplica testul CHI pentru a stabili dacă incidența cancerului la sân variază în concordanță cu cantitatea de grăsime din alimentație.

Etapele aplicării testului CHI

Pentru aplicarea testului CHI se parcurg următoarele etape:

➲ Se culeg şi se înregistrează datele studiului. Subiecții sunt clasificați ca bolnavi sau nu, expuşi la un anumit factor de risc sau nu, etc. Se stabileşte numărul de subiecți care fac parte din fiecare clasă. Numărul de subiecți care fac parte din clasa *i* relativ la prima caracteristică şi din clasa *j* relativ la a doua caracteristică se notează cu O_{ij} şi se numeşte frecvența observată a clasei *ij* .

➲ Se formulează două ipoteze:

(a) Ipoteza nulă (H_0) afirmă: „*între cele două caracteristici studiate nu există o asociere (o legătură)* ”

(b) Ipoteza alternativă (H_1) afirmă: „*există o asociere (o legătură) între cele două caracteristici studiate*”

➲ Se calculează frecvența relativă a fiecărei clase. Frecvențele relative se notează cu E_{ij} şi se calculează după formula:

➲ $$E_{ij} = \frac{\left(\sum_{i=1,n} O_{ij}\right)\cdot\left(\sum_{j=1,m} O_{ij}\right)}{\sum_{\substack{i=1,n\\ j=1,m}} O_{ij}}$$

➲ Se calculează valoarea statistică a testului CHI după formula:

➲ $$\chi^2_C = \sum_{\substack{i=1,n\\ j=1,m}} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

➲ Se calculează numărul de grade de libertate a testului CHI:

gl=(nr rânduri-1)*(nr coloane-1)

➲ Se calculează valoarea probabilității *p*. Probabilitatea *p* este probabilitatea să obținem întâmplător o valoare statistică egală sau mai mare decât valoarea χ_C^2 calculată, în

condițiile în care ipoteza nulă H_0 este adevărată. În cazul în care probabilitatea *p* calculată este ≤ 0,05 se respinge ipoteza H_0.

➲ Stabilirea concluziei testului CHI.

➲ Dacă probabilitatea *p* rezultată din calcul are o valoarea mai mică sau egală cu 0,05 , atunci se respinge ipoteza nulă H_0 şi se acceptă ipoteza alternativă H_1, prin urmare există o asociere între cele două caracteristici studiate.

➲ Dacă probabilitatea *p* rezultată din calcul are o valoarea mai mare decât 0,05 , atunci se acceptă ipoteza nulă H_0 , prin urmare nu există o asociere între cele două caracteristici studiate.

Utilizarea funcției CHITEST a utilitarului Excel

Utilitarul **Excel** oferă funcția CHITEST pentru calculul probabilității *p*. Argumentele acestei funcții sunt : zona care conține valorile de test şi zona care conține valorile estimate.

Pentru a putea utiliza această funcție trebuie să introducem datele aferente studiului nostru şi să calculăm valorile estimate E_{ij}.

	A	B		E	F
1					×
2					×
3			...		×
4	×	×	×	×	×

Figura 11.1

Spre exemplu, să presupunem că datele noastre sunt grupate în zona A-E, iar în căsuțele notate cu × sunt calculate totalurile pe linii, respectiv coloane (*figura 11.1*)

Valoarea unei celule din matricea valorilor estimate este egală cu produsul dintre suma valorilor de pe linia şi suma valorilor de pe coloana matricei datelor de test, totul împărțit la suma tuturor datelor de test.

Matricea valorilor estimate se contruieşte ca în figura următoare (*figura 11.2*)

	A	B		E	F
10	(F1*A4)/F4	(F1*B4)/F4		(F1*E4)/F4	
11	(F2*A4)/F4	(F2*B4)/F4		(F2*E4)/F4	
12	(F3*A4)/F4	(F3*B4)/F4	...	(F3*E4)/F4	
13					

Figura 11.2

După calculul matricei valorilor estimate se poate aplica funcția CHITEST şi anume: =CHITEST(A1:E3;A10:E12). Rezultatul întors de funcția CHITEST este valoarea *p*.

Exerciții.

1. O echipă de cardiologi au efectuat un studiu pentru a investiga o eventuală asociere între utilizarea medicamentelor contraceptive orale şi hipertensiune. Datele obținute au fost următoarele:

Tipul de contraceptive	*Tipul de tensiune*		
	Hipertensivitate	Tesiune normală	Total
Cu contraceptive orale	8	32	40
Cu alte contraceptive	15	45	60
	23	77	100

Să se stabilească dacă proporţia de femei hipertensive dintre cele care utilizează contraceptive orale diferă de proporţia de femei hipertensive dintre cele care folosesc alte medicamente anticoncepţionale.

Pentru a obţine rezultatul studiului, mai întâi introduceţi datele de mai sus într-o foaie de calcul tabelar. Apoi calculaţi frecvenţele estimate astfel:

E_{ij}=(Suma valorilor de pe linia **i**)*(suma valorilor de pe coloana **j**)/(suma tuturor valorilor).

După aceea aplicaţi testul CHI. Dacă probabilitatea P obţinută este mai mică decât 0,05 , atunci există diferenţă, deci există o legătură.

Ipoteze:

H_0: nu există nici o legătură între utilizarea medicamentelor contraceptive orale şi hipertensiune.

H_1: există o legătură între utilizarea medicamentelor contraceptive orale şi hipertensiune.

Dacă probabilitatea P obţinută este mai mică decât 0,05 , atunci există o legătură, se respinge ipoteza H_0 şi se acceptă iptoteza H_1.

Dacă probabilitatea P obţinută este mai mare decât 0,05 , atunci nu există nici o legătură, deci se acceptă ipoteza H_0.

Rezultate

Valorile estimate sunt:

9,2	30,8
13,8	46,2

p=0,560528 $\Rightarrow$ acceptăm ipoteza H_0

ipoteza H_0: nu există nici o legătură între utilizarea medicamentelor contraceptive orale şi hipertensiune.

2. Se efectuează un studiu pentru a stabili dacă există o asociere (legătură) între severitatea cancerului ovarian şi nivelul de stres. Datele obţinute sunt:

Severitatea bolii	*Nivelul de stres*				
	1	2	3	4	Total
Uşoară	362	60	141	317	880
Moderată	29	5	15	21	70
Severă	20	5	5	20	50
	411	70	161	358	1000

Stabiliţi dacă există o legătură sau nu.

3. Se studiază asocierea amigdalectomiei cu diferitele forme clinice de poliomielită, pe un lot de 461 de cazuri. Se pune întrebarea: diferențele sunt întâmplătoare? Prezența sau absența amigdalelor contribuie la determinarea formei de localizare a leziunilor de poliomielită ?

Datele studiate sunt prezentate în tabelul următor:

Tip boală	*Amigdale* prezente	absente
bulbară	16	99
dorsală severă	77	58
dorsală ușoară	76	85
neparalitică	24	26

4. Se studiază reacțiile locale produse de două tipuri de vaccin B.C.G. În acest scop s-au supus observației 348 de copii, dintre care la 177 s-a administrat vaccin de tip A, iar la 171 vaccin de tip B. Se doreşte să se afle dacă diferențele dintre reacțiile locale produse de aceste vaccinuri sunt semnificative din punct de vedere statistic sau dacă este vorba numai de o fluctuație de eşantion.

Datele rezultate din observarea reacțiilor locale sunt prezentate în tabelul următor:

Reacție locală	*tip vaccin* **A**	**B**
normală	12	29
intensă	156	135
ulcerație	8	6
abces	1	1

5. Să se testeze dacă există diferențe semnificative statistic între femei negravide, femei cu sarcini normale în luna a 9-a şi femei cu disgravidii tardive, privind valorile medii, în g/zi ale aldosteronului, cortizonului şi cortizolului.

Datele studiate sunt prezentate în tabelul următor:

Tip boală	*tip gravidă* negravide	sarcini normale	disgravidii tardive
aldosteron	4	79	24
cortizon	15	96	37
cortizol	25	55	33

6. În tabelul următor sunt trecute rezultatele unor observații asupra unui grup de 736 de persoane, în scopul stabilirii unei legături între persoanele supuse unui tratament împotriva holerei şi cele care suferă de această boală:

Persoane	suferă de holeră	nu suferă de holeră

supuse tratamentului	5	431
nesupuse tratamentului	9	291

Să se stabilească dacă tratamentul afectează numărul de persoane ce suferă de holeră, adică dacă există o legătură între numărul de persoane ce suferă de holeră şi numărul de persoane supuse tratamentului.

7. În urma aplicării unui vaccin, s-a înregistrat numărul de persoane care s-au îmbolnăvit şi care nu s-au îmbolnăvit. De asemenea, s-a înregistrat şi numărul persoanelor care s-au îmbolnăvit din rândul persoanelor nevaccinate. Se pune problema: diferenţele între bolnavii vaccinaţi şi cei nevaccinaţi sunt semnificative sau nu ?

Datele studiate sunt prezentate în tabelul următor:

	îmbolnăviţi	neîmbolnăviţi
vaccinaţi	20	74
nevaccinaţi	47	59

8. Se efectuează un studiu pentru a vedea dacă expunerea la un pesticid din agricultură are efect în avortul femeilor.

Datele studiate sunt prezentate în tabelul următor:

Tip boală	*femei gravide*	
	expuse la pesticid	neexpuse
cu avorturi spontane	30	10
fără avorturi	70	90

Să se stabilească existenţa unei eventuale legături între expunerea la pesticid şi avortul femeilor.

9. Se studiază efectul obţinut asupra numărului de carii prin efectuarea unui instructaj privind igiena orală unui număr de copii aleşi aleator. La 50 de copii li s-a făcut un instructaj privind igiena orală iar la 50 de copii aleşi la întâmplare nu li s-a făcut acest instructaj. Peste 6 luni s-au numărat cariile noi apărute. Se pune problema dacă aplicarea instructajului privind igiena orală are un efect asupra numărului de carii noi apărute.

Datele studiate sunt prezentate în tabelul următor:

	număr de carii noi		
	0-1	2-3	4-5
copii cu instrucţie	30	15	5
copii fără instrucţie	20	15	15

10. S-a studiat asocierea dintre prezența anemiei la un lot de subiecți şi grupa sanguină. Se pune problema: prezența anemiei este influențată de grupa sanguină?

Datele studiate sunt prezentate în tabelul următor:

grupa sanguină	*anemie*	
	prezentă	absentă
O	10	30
A	12	18
B	15	15
AB	13	12

11. Se efectuează un studiu pentru a stabili dacă există o legătură între nivelul de severitate al cancerului de plămâni şi starea de fumător sau nefumător. Stabiliți pe baza datelor următoare existența sau nu a unei legături:

Nivelul de severitate al cancerului	Fumători	Nefumători
Stadiul 1	60	40
Stadiul 2	75	25
Stadiul 3	80	20

Statistică Matematică

Corelație şi regresie

Interacțiunea dintre două variabile independente se referă la diferențele apărute în valorile măsurate ale unei variabile în funcție de nivelul celei de a doua variabile. De exemplu, este posibil ca un medicament să producă efecte mai bune dacă este utilizat în combinație cu un regim alimentar de reducere a greutății, decât dacă ar fi combinat cu un regim alimentar nesărat. În schimb, s-ar putea să nu obținem efecte semnificative ale medicamentului dacă se studiază toate grupurile alimentare la un loc. Studiul efectelor medica-mentului separat pe diferite regimuri alimentare ne conduce la concluzia că există o interacțiune între doi factori: regimul alimentar şi medicamentul.

Asocierea şi cauzalitatea – coeficientul de corelație

În acumularea şi evidența datelor ştiințifice apar o serie de probleme specifice, cum ar fi problema asocierii (dependenței) între două variabile. Se pune problema: există o dependență între sărăcie şi consumul de droguri ? Este stresul asociat cu boli cardiovasculare ?

Pentru a determina dacă există sau nu o astfel de dependență, trebuie mai întâi să cuantificăm, să măsurăm ambele variabile. De exemplu, stresul poate fi cuantificat prin utilizarea unor teste psihologice sau prin definirea clară, evaluarea şi scalarea factorului de

stres în situațiile din viața de zi cu zi. În ceea ce privește hipertensiunea, aceasta poate fi direct cuantificată prin măsurarea presiunii sanguine.

După ce variabilele au fost cuantificate, este necesară calcularea unei măsuri a dependenței dintre ele, adică a tăriei dependenței. De obicei se calculează **coeficientul de corelație** „*r*”. Coeficientul de corelație *r* este un număr calculat direct din datele observate şi poate varia între –1 şi +1.

Dacă x_i sunt valorile măsurate ale variabilei X şi y_i sunt valorile măsurate ale variabilei Y, $i=1,N$, atunci coeficientul de corelație se calculează astfel:

$$r = \frac{N\sum_i x_i \cdot y_i - \left(\sum_i x_i\right)\cdot\left(\sum_i y_i\right)}{\sqrt{N\sum_i x_i^2 - \left(\sum_i x_i\right)^2}\cdot\sqrt{N\sum_i y_i^2 - \left(\sum_i y_i\right)^2}}$$

Dacă coeficientul de corelație este $r = 0$, atunci înseamnă că nu avem nici o corelație între cele două variabile. De exemplu, nu există nici o legătură între presiunea sanguină şi numărul de fire de pe cap.

Dacă coeficientul de corelație este $r = +1$ înseamnă că avem o corelație pozitivă perfectă, adică există o dependență directă între cele două variabile. O persoană care are o valoare mare la prima variabilă va avea o valoare mare şi la cea de a doua. De asemenea, valoarea unei variabile poate fi prevăzută exact pe baza valorii celei de a doua variabile. Un exemplu de acest tip este corelația dintre vârsta unui copac şi numărul său de inele.

Dacă coeficientul de corelație este $r = -1$ atunci avem o dependență inversă perfectă. O valoare mare a unei variabile înseamnă o valoare mică a celeilalte variabile.

Dacă coeficientul de corelație este între 0 şi +1 sau între –1 şi 0 , atunci valoarea lui *r* ne dă tăria dependenței celor două variabile.

Aceste considerente se aplică în cazul în care dependența dintre cele două variabile este liniară. Dacă efectuăm de exemplu măsurători ale înălțimii şi greutății pentru un grup de persoane şi calculăm coeficientul de corelație, vom obține o valoare pozitivă, dar o valoare mai mică decât 1.

Corelație şi cauzalitate

Problema determinării tăriei corelației dintre variabilele aleatoare este o problemă relativ dificilă, ce depinde de domeniul aplicațiilor, precum şi de mulți alți factori. Variabilele psihologice sunt mai dificil de măsurat cu exactitate şi sunt afectate în general de multe alte variabile, fiind astfel dificil de stabilit corelațiile dintre ele. Corelațiile dintre variabilele biologice sunt în general mai tari, acestea având dealtfel şi avantajul că pot fi măsurate cu mai mare precizie.

Ca un exemplu, corelațiile dintre aptitudinile verbale şi cele non-verbale la copiii şcolari, măsurate la Philadelphia cu ajutorul unor teste standard naționale, variază între 0,44 şi 0,77 depinzând de rasă şi de clasa socială.

Pentru a stabili corelații cât mai semnificative, trebuie identificate situațiile care sunt responsabile, care cauzează aceste corelații.

ATENȚIE! *Existența unei corelații între două variabile nu implică în mod necesar cauzalitatea, se poate datora unor cauze comune. Prin urmare trebuie avut grijă la interpretarea acestor coeficienți de corelație.*

Reprezentarea grafică

Datele corespunzătoare celor două variabile studiate se pot reprezenta grafic sub forma unui sistem de coordonate bidimensionale. **Microsoft Excel** pune la dispoziție un astfel de grafic, numit *XY Scatter*.

Între cele două variabile există o corelație puternică dacă punctele reprezentate grafic sunt grupate de-a lungul unei drepte (*figura 12.1*). Cu cât punctele sunt mai alineate, cu atât corelația este mai puternică.

Valoarea critică a coeficientului de corelație

În studiul statistic al corelației a două variabile se pune întrebarea: sunt cele două variabile corelate semnificativ de tare din puncte de vedere statistic?

Pentru a răspunde la această întrebare trebuie calculat un prag critic. Corelația dintre două variabile se va estima cu o marjă de eroare numită nivel de semnificație, notat cu ***p*** . Cu cât ***p*** este mai mic, cu atât riscul (probabilitatea) de a greşi este mai mic, deci estimarea este mai sigură. Reamintim câteva noțiuni importante din *Probabilități şi Statistică Matematică*:

- **evenimentul sigur** - acel eveniment care va apărea întotdeauna, indiferent de situație.
- **1** - reprezintă probabilitatea ca să apară evenimentul sigur.
- **0** – reprezintă probabilitatea ca să nu apară evenimentul sigur.
- **probabilitatea** de apariție a oricărui alt eveniment, diferit de evenimentul sigur sau de evenimentul imposibil, variază ca valoare între 0 şi 1.

Numărul gradelor de libertate reprezintă numărul de perechi de date care se studiază, minus două.

Pragul critic reprezintă valoarea coeficientului de corelație peste care se consideră corelația ca fiind semnificativă. Dacă coeficientul de corelație depăşeşte acest prag critic, variabilele studiate se consideră corelate.

Pragul critic depinde de numărul gradelor de libertate şi de nivelul de semnificație.

Anexa 1 prezintă tabelul cu pragurile critice pentru nivelele de semnificație 0,10 ; 0,05 ; 0,02 şi 0,01.

Metoda practică de stabilire a corelației dintre două variabile

Pentru a afla dacă două variabile studiate sunt corelate sau nu, formulăm următoarele ipoteze statistice:

H_0: cele două variabile studiate nu sunt corelate.

H_1: cele două variabile studiate sunt corelate.

În continuare se efectuează paşii următori:

1. Calculăm coeficientul de corelație *r* asociat datelor x_i şi y_i ,cu ajutorul formulei prezentate mai sus sau cu ajutorul programului *Microsoft Excel* şi anume utilizând funcția CORREL(zona1; zona2).

2. Calculăm numărul gradelor de libertate: numărul perechilor de date – 2.

3. Analizăm datele din tabelul din anexa 1. În acest tabel, pentru numărul de grade de libertate calculat există mai multe praguri de semnificație: câte unul pentru fiecare nivel de semnificație şi anume: pentru 0.10 , 0.05 , 0.02 , 0.01. De exemplu, dacă *r* calculat este mai mare decât una dintre valorile din tabel, atunci cele două variabile sunt corelate cu nivelul de semnificație respectiv. Dacă *r* este mai mare decât pragul critic pentru 0,05 atunci cele două variabile sunt considerate corelate cu un nivel de semnificație de 0,05. În general se urmăreşte să se obțină o corelație cu un nivel de semnificație cât mai mic. Cu cât nivelul de semnificație este mai mic, cu atât corelația este mai sigură şi sunt mai puține şanse să greşim deoarece marja de eroare este mai mică.

4. În toate aceste cazuri se respinge ipoteza H_0 şi se acceptă ipoteza H_1 cu nivelul de semnificație respectiv.

Dacă *r* obținut este mai mic decât toate valorile din tabel, atunci cele două variabile sunt considerate necorelate. În acest caz se acceptă ipoteza H_0 .

Cel mai des se utilizează nivelul se semnificație 0,05 sau 0,01, care sunt considerate suficiente.

↳Exemplu

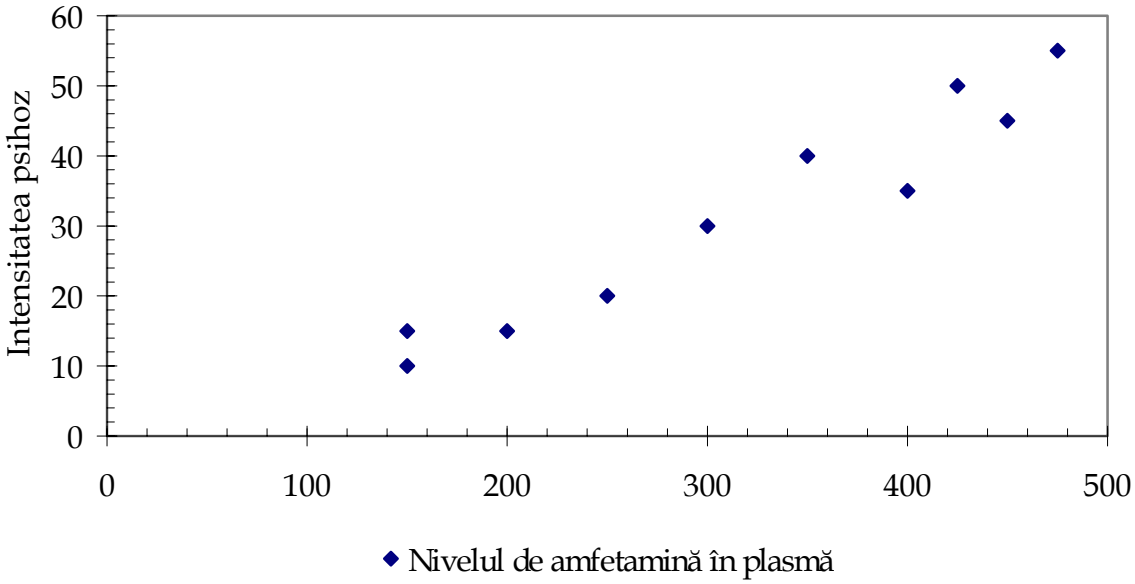
Studiem corelația dintre nivelul de Amfetamină în plasmă şi Intensitatea Psihozei.

Datele pe care efectuăm studiul sunt următoarele:

Nr. subiect	Intensitatea psihozei	Amfetamină în plasmă mg / ml
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

Reprezentarea grafică a datelor este ilustrată în *figura 12.1*

Coeficientul *r* rezultat din calcul este 0,96738 . Numărul gradelor de libertate este 10-2=8. În anexă, pragul critic pentru 8 grade de libertate şi nivelul de semnificație 0,05 este 0,6319, iar pentru nivelul de semnificație .01 este 0,7646.



REZULTAT

r > 0,7646 ⇒ se respinge ipoteza H_0 şi se acceptă ipoteza H_1 cu un nivel de semnificaţie de 0,01. Rezultă că intensitatea psihozei este corelată cu nivelul de Amfetamină din plasmă.

Exerciţii

1. Studiindu-se relaţia dintre doza unui medicament (exprimată în multipli ai unei doze minime) şi durata bolii (exprimată prin numărul de zile de boală), s-a obţinut următoarea relaţie:

Nr. subiect	doză	durată
1	1	23,5
2	2	20,0
3	3	14,9
4	4	8,1
5	5	7,5

Să se reprezinte grafic datele din tabel şi să se verifice dacă există o legătură între doza medicamentului şi durata bolii.

1. Să se aprecieze existenţa şi gradul legăturii dintre consumul de alcool (vin în litri) pe cap de locuitor, pe lună, şi vârsta medie de debut a cirozei hepatice. Datele de test sunt prezentate în tabelul următor.

Nr. subiect	l de vin pe lună	Vârsta de debut a cirozei
1	7	56
2	8	55
3	8	58
4	10	55
5	12	52
6	13	51
7	15	50
8	15	48
9	15	45
10	16	40
11	16	47
12	16	44
13	17	40
14	17	40
15	18	38
16	18	38
17	19	40
18	20	38
19	20	35
20	20	35

Să se reprezinte grafic datele din tabel.

1. Să se aprecieze gradul şi sensul legăturii dintre temperatură şi puls la un lot de 20 de bonavi. Datele de test sunt prezentate în tabelul următor.

Nr. subiect	temperatură	puls
1	36,5	68
2	36,6	72
3	36,7	70
4	36,7	74
5	36,8	72
6	36,8	75
7	36,8	70
8	37,0	78
9	37,0	78
10	37,4	80
11	37,8	82
12	38,0	82
13	38,3	84
14	38,4	85
15	38,8	86
16	38,9	86
17	39,0	100
18	39,2	94
19	39,4	110
20	39,6	120

Să se reprezinte grafic datele din tabel.

1. Următorul tabel conține informații despre un lot de paciente diagnosticate cu cancer de col uterin. Se cere să se precizeze dacă există o legătură între vârsta la care a fost depistat cancerul de col uterin şi menarha (vârsta de început a menstruației) şi să se reprezinte grafic datele din tabel.

Nr. subiect	vârsta	menarha
1	57	14
2	56	14
3	32	13
4	39	15
5	46	13
6	73	14
7	38	11
8	69	15
9	67	13
10	61	15
11	65	13
12	45	14
13	47	14
14	50	15
15	65	15
16	49	14
17	63	14
18	47	12
19	43	15
20	40	12
21	58	13
22	41	14
23	49	14
24	54	14
25	32	19
26	50	12
27	50	14
28	49	14
29	49	12
30	45	12
31	33	16
32	42	13
33	35	15
34	33	12
35	78	13
36	64	13
37	28	13
38	48	14
39	37	17
40	49	13
41	48	14
42	48	13
43	58	14
44	41	13
45	54	14

1. Într-un studiu cuprinzând cazuri de stenoză, s-au măsurat valorile IMT maxim şi valoarea sistolică la diverşi pacienți. Se pune problema există o legătură între valoarea sistolică şi valoarea IMT maxim ?

Datele de test sunt prezentate în tabelul următor.

1. La 7 subiecți de vârste diferite s-a înregistrat tensiunea arterială sistolică.

Datele obținute sunt prezentate în tabelul de mai jos. Se cere să se determine relația care există între variabila **vârstă** şi variabila **tensiune sistolică** şi să se reprezinte grafic datele din tabel.

Nr. subiect	vârsta	tensiunea arterială
1	35	114
2	45	124
3	48	125
4	55	143
5	57	144
6	65	158
7	75	166

.

Nr. subiect	IMT maxim	Valoarea sistolică a tensiunii arteriale
1	1,6	150
2	1,7	175
3	1,5	160
4	1,5	175
5	1,5	145
6	2,1	155

7	1,9	180
8	2,1	145
9	1,6	145
10	1,6	170
11	1,9	155
12	2,3	165
13	1,8	160

Să se reprezinte grafic datele din tabel.

1. O companie farmaceutică a încercat să evalueze relația dintre doza ingerată a unui nou medicament hipnotic şi durata somnului. Datele culese în urma studiului sunt prezentate în tabelul de mai jos. Există o legătură lineară între aceste două variabile?

Nr. subiect	**durata somnului (ore)**	**doza (mM/kg)**
1	4	3
2	6	3
3	5	3
4	9	10
5	8	10
6	7	10
7	13	15
8	11	15
9	9	15

Să se reprezinte grafic datele din tabel.

1. Într-un eşantion format din 10 persoane s-a măsurat înălțimea şi greutatea, pe baza cărora s-a atribuit fiecărei persoane un rang (poziție), în funcție de înălțime şi de greutate. Spre exemplu, a 8-a persoană ca înălțime este a 7-a ca şi greutate. Se pune problema există o legătură între înălțime şi greutate ?

Datele de test sunt prezentate în tabelul următor.

Nr. subiect	**înălțime**	**greutate**
1	3	1
2	1	2
3	2	3
4	8	7
5	5	6
6	9	8
7	10	10
8	6	5
9	7	9
10	4	4

Să se reprezinte grafic datele din tabel.

Într-un studiu cuprinzând multe cazuri, s-a descris relația dintre durata sarcinii exprimată în săptămâni şi greutatea la naştere (g). Prezentăm câteva dintre datele experimentale, care se referă la perioada între săptămâna a 26-a şi săptămâna a 37-a. Se cere să se studieze statistic relația dintre cele două variabile.

Datele de test sunt prezentate în tabelul alăturat.

Să se reprezinte grafic datele din tabel.

	săptămâna	**greutatea**
1	26	700
2	27	1050
3	28	1200
4	28	1230
5	29	1300
6	29	1325
7	30	1500
8	31	1600
9	31	1645
10	31	1640
11	32	1900
12	32	1920
13	32	1915
14	33	2100
15	33	2160
16	34	2300
17	34	2350
18	35	2500
19	35	2550
20	36	2700
21	37	2800

Soluții

Biostatistică

1. Media : 88
Amplitudinea : 41
Amplitudinea relativă: 46,59 %
Dispersia: 219,3333
Deviația standard: ±14,8099066
Coeficientul de variație:16,8294%

2. Media : 3000
Amplitudinea : 400
Amplitudinea relativă: 13,3333%
Dispersia: 12121,2121
Deviația standard: 110,096377
Coeficientul de variație: 3,6699%

3. Media : 29,84
Amplitudinea : 20
Amplitudinea relativă: 67,0241 %
Dispersia: 13,48180905
Deviația standard: 3,671758304
Coeficientul de variație: 12,3048%

Testul Student

4. p = 0,3131696 - nu diferă semnificativ

5. p = 0,0095346130 - diferă semnificativ

6. p = 0,13581046 - nu diferă

7. p = 0, 000000000011- diferă semnificativ

8. p = 0,008718356343 - diferă semnificativ

9. p =0,0601985648 - nu diferă

10. p = 0, 00080001 - diferă semnificativ

Testul CHI

1. p =0,5605275402 - nu există o legătură
2. p = 0,6749187157 - nu există o legătură
3. p = 0,0000000000252 - există o legătură
4. p = 0,032819435 - există o legătură
5. p = 0,00012499 - există o legătură
6. p = 0,070518030 - nu există o legătură
7. p = 0,000562803 - există o legătură
8. p = 0,000007744 - există o legătură
9. p = 0,030197383 - există o legătură
10. p = 0,089662498 - nu există o legătură
11. p = 0,004819151 - există o legătură

Corelație şi regresie

1. p = - 0,979383844 - corelate
2. p= - 0,94961689 - corelate
3. p = 0,900574197 - corelate
4. p = - 0,09190074 - necorelate
5. p = 0,984963074 - corelate
6. p = - 0,01283513 - necorelate
7. p = 0,900375235 - corelate
8. p = 0,915151515 - corelate
9. p = 0,994959179 - corelate

Anexa 1.

Valoarea criticã a coeficientului de corelaþie pentru nivele diferite de semnificaþie: 0,10 ; 0,05 ; 0,02 ; 0,01

gl	**0,10**	**0,05**	**0,02**	**0,01**
1	0,9877	0,9969	0,9995	0,9998
2	0,9000	0,9500	0,9800	0,9900
3	0,8054	0,8783	0,9343	0,9587
4	0,7293	0,8114	0,8822	0,9172
5	0,6694	0,7545	0,8329	0,8745
6	0,6215	0,7067	0,7887	0,8343
7	0,5822	0,6664	0,7498	0,7977
8	0,5494	0,6319	0,7155	0,7646
9	0,5214	0,6021	0,6851	0,7348
10	0,4973	0,5760	0,6581	0,7079
11	0,4762	0,5529	0,6339	0,6835
12	0,4575	0,5324	0,6120	0,6614
13	0,4409	0,5139	0,6923	0,6411
14	0,4259	0,4973	0,5742	0,6226
15	0,4124	0,4821	0,5577	0,6055
l6	0,4000	0,4683	0,5425	0,5897
17	0,3887	0,4555	0,5285	0,5751
18	0,3783	0,4438	0,5155	0,5614
19	0,3687	0,4329	0,5034	0,5487
20	0,3598	0,4227	0,4921	0,5368
25	0,3233	0,3809	0,4451	0,4869
30	0,2960	0,3494	0,4093	0,4487
35	0,2746	0,3246	0,3810	0,4182
40	0,2573	0,3044	0,3578	0,3932
45	0,2428	0,2875	0,3384	0,3721
50	0,2306	0,2732	0,3218	0,3541
60	0,2108	0,2500	0,2948	0,3248
70	0,1954	0,2319	0,2737	0,3017
80	0,1829	0,2172	0,2565	0,2830
90	0,1726	0,2050	0,2422	0,2673
100	0,1638	0,1946	0,2301	0,2540

***Gradele de libertate** (gl) = **(numãrul de perechi din datele studiate –** 2).