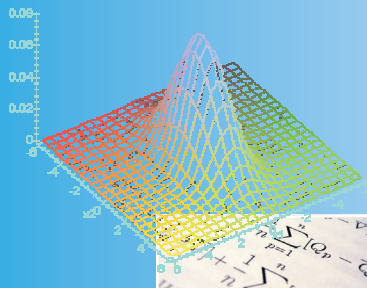


Marius Mărușteri

Noțiuni fundamentale de biostatistică



- Note de curs -



2006

Marius Ștefan Mărușteri

NOȚIUNI FUNDAMENTALE DE BIOSTATISTICĂ

- NOTE DE CURS -



- 2006 -

MĂRUȘTERI, ȘTEFAN MARIUS

Noțiuni fundamentale de biostatistică : note de curs

Referenți:

Prof. Dr. Șchiopu Alexandru

Prof. Dr. Nicolaescu Ioan

Descrierea CIP a Bibliotecii Naționale a României

CIP nr. 5537/ 12.04.2006

MĂRUȘTERI, ȘTEFAN MARIUS

Noțiuni fundamentale de biostatistică : note de curs / Marius Mărușteri

- Târgu-Mureș : University Press, 2006

Bibliogr.

ISBN (10) 973-7665-11-2 ; ISBN (13) 978-973-7665-11-9

519.22:57(075.8)

Editor:

©University Press, Târgu Mureș, 2006

Str. Gh. Marinescu, 38,

Târgu Mureș, MS, 540139

Director de editură: Prof. Dr. Șchiopu Alexandru

Tel: +40-265-215551/126

Fax: +40-265-210407

CUPRINS

I. INTRODUCERE	6
☞ POPULAȚIA STATISTICĂ ȘI PROBELE STATISTICE	7
☞ VARIABILITATE BIOLOGICĂ	9
☞ FACTORI DE EROARE	10
II. DISTRIBUȚII DE FRECVENȚĂ	16
☞ CURBA DE DISTRIBUȚIE NORMALĂ	21
☞ DISTRIBUȚII ANORMALE (NON-GAUSSIENE)	25
III. STATISTICĂ DESCRIPTIVĂ. CONSIDERAȚII GENERALE. INDICATORI STATISTICI.....	28
☞ A. INDICATORI AI TENDINȚEI CENTRALE	28
☞ B. INDICATORI AI DISPERSIEI (ÎMPRĂȘTIERII) DATELOR ÎN JURUL VALORII MEDII ..	29
☞ III.1. MEDII	31
☞ III.2. DISPERSIA (VARIANȚA). ABATEREA STANDARD	34
☞ III.3. EROAREA STANDARD	35
☞ III.4. EROAREA PROCENT	36
☞ III.5. COEFICIENTUL DE VARIAȚIE.....	37
☞ III.6. GRADE DE LIBERTATE	37
☞ III.7. LIMITE FIDUCIALE (INTERVAL DE ÎNCREDERE)	39
IV. STATISTICĂ INFERENȚIALĂ. INTERPRETAREA STATISTICĂ A REZULTATELOR UNUI EXPERIMENT.....	41
☞ IV.1. TESTE DE VALABILITATE	42
<i>Noțiuni introductive</i>	<i>42</i>
<i>IV.1.1. Testul Grubbs.....</i>	<i>43</i>
<i>IV.1.2. Criteriul de eliminare a lui Chauvenet</i>	<i>47</i>
☞ IV.2. TESTE DE CONCORDANȚĂ (GOODNESS-OF-FIT).....	50
<i>Noțiuni introductive</i>	<i>50</i>
<i>IV.2.1. Testul de concordanță Kolmogorov-Smirnov.....</i>	<i>52</i>
☞ IV.3. TESTE DE SEMNIFICAȚIE	55
<i>Valoarea lui P.....</i>	<i>56</i>
<i>Ipoteza nulă (H_0).....</i>	<i>56</i>
<i>Dependența/independența eșantioanelor studiate</i>	<i>58</i>
<i>Statistică parametrică/neparametrică</i>	<i>59</i>
<i>IV.3.1. Teste de semnificație parametrice.....</i>	<i>60</i>
<i>IV.3.2. Teste de semnificație neparametrice. Statistica ordinii.....</i>	<i>72</i>
<i>Testul Wilcoxon</i>	<i>72</i>
<i>Testul Mann-Whitney U.....</i>	<i>76</i>
V. ANALIZA DISPERSIONALĂ (ANALIZA DE VARIANȚĂ) - ANOVA	79

☞	A. BAZE TEORETICE.....	79
☞	B. CONSIDERAȚII PRACTICE.....	81
☞	V.1. ANALIZA DISPERSIONALĂ MONOFACTORIALĂ.....	83
☞	V.2. ANALIZA DISPERSIONALĂ BIFACTORIALĂ	90
	VI. CORELAȚIE ȘI REGRESIE	98
☞	COEFICIENTUL DE CORELAȚIE.....	99
☞	COEFICIENTUL DE DETERMINARE.....	106
☞	ECUAȚIA DE REGRESIE	107
	VII. ALEGEREA UNUI ANUMIT TEST STATISTIC	112
	VIII. STUDII EPIDEMIOLOGICE ȘI CLINICE	115
☞	VIII.1. INDICATORI STATISTICI UTILIZAȚI ÎN STUDIILE EPIDEMIOLOGICE	115
	<i>Rate și proporții.....</i>	<i>116</i>
	<i>Indicatori ai frecvenței bolii</i>	<i>117</i>
	<i>Indicatori ai asocierii factori de risc (expunere)-boală</i>	<i>119</i>
☞	VIII. 2. TIPURI DE STUDII EPIDEMIOLOGICE	134
	<i>STUDII DESCRIPTIVE</i>	<i>134</i>
	<i>C. Studiu descriptiv de tip ecologic</i>	<i>137</i>
	<i>STUDII ANALITICE.....</i>	<i>138</i>
	<i>A. Studii de tip Case-Control (Case-Control studies).....</i>	<i>139</i>
	<i>B. Studii de tip cohortă (Cohort studies)</i>	<i>142</i>
	<i>Selecția unui anumit tip de studiu (study design).....</i>	<i>146</i>
	<i>Asociere versus cauzalitate.....</i>	<i>147</i>
	<i>Criterii pentru stabilirea unei relații de tip cauză-efect</i>	<i>148</i>
	ANEXE - EXERCIȚII RECAPITULATIVE	150
	ANEXA 1 - STATISTICĂ DESCRIPTIVĂ, EXERCIȚII RECAPITULATIVE.....	151
☞	SCURTĂ RECAPITULARE.....	151
☞	INDICATORI STATISTICI.....	152
☞	INDICATORI STATISTICI PENTRU SERII DE DATE CU APARIȚII FRECVENTE ALE ACELEIAȘI VALORI.....	154
☞	FOLOSIREA CALCULATORULUI ÎN STATISTICA DESCRIPTIVĂ. APLICAȚII ȘI EXEMPLE. 155	
☞	STATISTICĂ DESCRIPTIVĂ – EXERCIȚII RECAPITULATIVE	160
	ANEXA 2 - TESTE STATISTICE DE SEMNIFICAȚIE, EXERCIȚII RECAPITULATIVE.....	163
☞	TESTUL STUDENT	163
☞	ETAPELE APLICĂRII TESTULUI STUDENT	163
☞	UTILIZAREA FUNCȚIEI TTEST A UTILITARULUI EXCEL	165
☞	TESTUL STUDENT – EXEMPLE ȘI EXERCIȚII RECAPITULATIVE	167
☞	TESTUL χ^2	174
	ETAPELE APLICĂRII TESTULUI χ^2	174

UTILIZAREA FUNCȚIEI CHITEST A PROGRAMULUI EXCEL	176
↳ TESTUL χ^2 – EXERCIIȚII RECAPITULATIVE.....	177
ANEXA 3 – CORELAȚII ȘI REGRESII, EXERCIIȚII RECAPITULATIVE.....	185
ANEXA 4 – EXEMPLE DE STUDII EPIDEMIOLOGICE. STUDIU DE TIP CASE-CONTROL.....	198
↳ STUDIUL INFLUENȚEI UNOR FACTORI DE RISC BIOLOGICI - INFECȚIA CU HELICOBACTER PYLORI – ÎN APARIȚIA UNOR LIMFOAME MALIGNNE (LIMFOAME NONHODGKIN –LNH) ÎN CAZUL POPULAȚIEI JUDEȚULUI MUREȘ [22].....	198
ANEXA 5 - VALORI CRITICE PENTRU TESTUL F CORESPUNZĂTOARE PRAGULUI DE SEMNIFICAȚIE $\alpha = 0,05$.....	208
ANEXA 6 - VALORILE CRITICE PENTRU TESTUL “T” STUDENT CORESPUNZĂTOARE DIFERITELOR PRAGURI DE SEMNIFICAȚIE α ȘI NUMĂRULUI “N ” AL GRADELOR DE LIBERTATE	210
ANEXA 7 - VALORILE CRITICE PENTRU TESTUL χ^2 CORESPUNZĂTOARE DIFERITELOR PRAGURI DE SEMNIFICAȚIE α ȘI NUMĂRULUI “N ” AL GRADELOR DE LIBERTATE	212
ANEXA 8 – SOLUȚIILE EXERCIIȚIILOR PROPUSE ÎN ANEXE	214
BIBLIOGRAFIE SELECTIVĂ	216

Capitolul I

I. Introducere

[12, 13]

Statistica matematică este știința care urmărește explicarea fenomenelor de masă printr-un număr relativ redus de observații. Ea folosește metode inductive de cercetare, plecând de la particular la general.

Desigur, concluziile rezultate în urma prelucrării statistice a datelor experimentale nu sunt legități absolut sigure, însă gradul de incertitudine poate fi calculat. Cunoscând gradul de incertitudine al concluziilor trase, metoda statistică de cercetare poate fi considerată o metodă matematică exactă.

Biostatistica este o ramură a statisticii, specializată în studiul fenomenelor biologice, inclusiv al celor medicale. Se ocupă de culegerea, centralizarea și gruparea datelor, precum și de prelucrarea și determinarea unor *parametri* sau *indicatori statistici* pentru descrierea fenomenelor biomedicale studiate, pe baza evidențierii unor regularități sau variabilități statistice. Totodată aplică și dezvoltă tehnici statistico-probabilistice pentru analiza datelor biomedicale.

Începuturile biostatisticii au fost determinate de nevoia obținerii unor informații cantitative dintre cele mai simple, formulate de regulă sub forma „câți bolnavi ? ” , „câți decedați ? ” , etc. Cu timpul s-a constatat însă că asemenea metode sunt insuficiente pentru caracterizarea fenomenelor, datorită existenței unor variații în răspunsurile care se obțin între diverse măsurători sau, cu alte cuvinte, datorită faptului că fenomenele biologice sunt caracterizate prin variabilitate. Chiar și în aceste condiții, prin

observarea unor 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 câștigat 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 complex și a calculului probabilităților în special.

Pentru a putea studia procesul biologic dorit, datele experimentale trebuie sistematizate prin grafice și tabele, întocmite prin luare în considerare fie a numărului total de date, fie a unui eșantion extras din acestea.

Din acest punct de vedere, statistica matematică operează cu două noțiuni de bază:

1. *Populația sau colectivitatea statistică;*
2. *Proba (eșantionul) extras din populația aflată în studiu.*



Populația statistică și probele statistice

O *populație statistică* poate fi definită prin totalitatea fenomenelor sau a obiectelor calitativ omogene având una sau mai multe caracteristici comune. De exemplu 100 de cobai cărora le-a fost injectat intraperitoneal o anumită substanță activă pentru a studia un anumit efect farmacologic al acesteia reprezintă un exemplu de *populație statistică*.

Trebuie făcută o deosebire între o populație statistică *finită*, *infinită* și *ipotetică*. Exemplul de mai sus reprezintă o *populație statistică finită*, în timp ce urmărirea frecvenței de apariție a unui anumit efect secundar pentru un medicament dat la nivelul unei clinici pe o durată nedefinită de timp (ce poate continua luni sau ani) reprezintă un exemplu de *populație statistică infinită*. Drept exemplu de *populație statistică ipotetică* se pot menționa parametrii farmacocinetici ai unui model farmacocinetic simulat pe calculator care va urma să fie verificat și experimental.

Pentru cunoașterea proprietăților unei populații statistice este necesar să se cunoască proprietățile elementelor din care aceasta este compusă. Practic este însă imposibil să se determine compoziția inițială a populației statistice studiate, deoarece ea fie ar fi distrusă, fie ar fi necesare un număr foarte mare de determinări, imposibil de efectuat. Din aceste motive se apreciază proprietățile caracteristice ale populației statistice numai pe baza unei părți finite din populație numită **eșantion (probă)**, care trebuie să îndeplinească o condiție *sine qua non*: trebuie ca ea să fie luată din populația statistică în așa fel încât fiecare element din populație să aibă aceeași probabilitate de a face parte din eșantion. Eșantioanele care satisfac acest deziderat se numesc **eșantioane reprezentative**.

Înainte de a trece în revistă diferiți *indicatori și teste statistice* trebuie menționată însă *variabilitatea* ce există în cadrul populațiilor statistice studiate și importanța evaluării acestora pentru obținerea unor rezultate corecte în urma experimentelor efectuate.

Variabilitate biologică

Așa cum am menționat, analiza statistică dă cercetătorului posibilitatea de a preciza *variabilitatea* existență în sânul unei colectivități. Măsura acestei variabilități dă indicații cu consecințe practice, în special în domeniul biologiei unde se spune că „variabilitatea este singura realitate, media fiind o ficțiune”. În fața proceselor biologice, atât de complexe, cum s-ar putea cunoaște valorile normalului și limitele lui de variație?

Fiind dat un anumit experiment, cu o împrăștiere mare a rezultatelor, cum se poate trage o concluzie justă în aprecierea fenomenului cercetat? Cum putem ști dacă am efectuat un număr suficient de experiențe pentru a atrage o concluzie valabilă ? Când putem spune, suficient de exact, că o diferență între două măsurători este semnificativă ?

Fără sprijinul statisticii matematice aceste întrebări nu și-ar găsi niciodată răspunsul. Mai mult, interpretarea simplistă a rezultatelor este neconcludentă, empirică, putând duce la erori grave de interpretare, ceea ce, în cazul domeniului medical poate pune vieți umane în pericol.

Cercetarea biologică se bazează pe rezultate obținute pe un număr limitat de observații din multiplele posibile, este deci o *cercetare de eșantion*. Se pune problema dacă putem generaliza observațiile obținute pe un număr limitat de cazuri, la întreaga colectivitate studiată, obținând astfel legi cu aplicare generală. Generalizarea este posibilă doar dacă ținem cont de variabilitatea cifrelor obținute, iar statistica matematică dă posibilitatea aprecierii acestei variabilități.

Astfel, dacă eșantionul este *reduc ca număr*, rezultatele obținute în cercetările biologice pot fi diferite, uneori chiar contrarii adevărului, ceea ce poartă numele de *fluctuație de eșantionaj*. Dacă eșantionul este *numeros*, rezultatul va fi evident mai apropiat de adevăr, media rezultatelor putând fi generalizată la întreaga populație. Un alt aspect deosebit de important este cel al *împrăștierei rezultatelor*: cu cât împrăștierea va fi mai mare, cu atât rezultatele se vor îndepărta mai mult de cele obținute când se examinează întreaga populație (deci de valoarea reală).

În concluzie, generalizarea - în scopul ajungerii la concluzii valabile - depinde, din punct de vedere statistic, de două caracteristici ale datelor analitice:

1. *Numărul observațiilor efectuate (n);*
2. *Împrăștierea (dispersia) acestora (σ).*

Statistica matematică dă posibilitatea aprecierii acestor caracteristici și de aici putem deduce că ea este cea care acordă valoarea unei anumite cercetări ce dorește desprinderea unor concluzii cu caracter generalizator.

Factori de eroare

Este cunoscut faptul că aceeași substanță activă, experimentată prin aceeași metodă, poate da rezultate diferite, nu numai în laboratoare diferite, ci chiar și în același laborator. De aceea cunoașterea de către cercetător a factorilor care determină această variabilitate, precum și a tipurilor de erori ce pot să apară este o necesitate.

În general variabilitatea rezultatelor unui experiment biomedical sau farmacologic este determinată de următorii factori:

I. Factori care țin de animalul de experiență:

- A. Factori interni:*** greutate, sex, vârstă, rasă, origine, sănătate.
- B. Factori externi:*** condiții de întreținere (alimentație, temperatură ambiantă), condiții sezoniere.
- C. Factori care țin de individualitatea animalului,*** proprii fiecărui individ și care caracterizează reactivitatea acestuia (amintim aici, de exemplu, mare variabilitate întâlnită la metabolizarea alcoolului etilic în cazul indivizilor speciei umane, în funcție de cantitatea de ADH disponibilă în cazul fiecărui individ).

II. Factori care țin de mediul geografic și climateric. Se pot obține rezultate diferite în diverse părți ale globului investigând același medicament și urmând chiar aceeași tehnică de lucru, organismul uman sau animal fiind o entitate biologică a cărei reactivitate individuală depinde de mediul înconjurător, de factorii micro- și macroclimatici înconjurători.

III. Factori care țin de tehnicile întrebuițate. Fiecare tehnică de lucru poate da un rezultat care poate fi chiar în discordanță cu cel obținut printr-o altă tehnică, de aceea cele două rezultate nu pot fi obiectul unei comparații realizate științific.

Data fiind multitudinea factorilor de eroare, cercetarea biologică ar fi insuficientă sau neconcludentă dacă nu s-ar ține seama de anumite norme științifice în experimentare. Aceste norme se referă, în primul rând la

înlăturarea factorilor care determină variabilitatea răspunsurilor biologice.

Câteva măsuri importante vor fi menționate mai jos:

- ⇒ selecția riguroasă a animalelor;
- ⇒ întocmirea loturilor dintr-un număr suficient de indivizi, siguri din punct de vedere biologic și/sau statistic (sănătoși, de vârstă apropiată, etc.);
- ⇒ asigurarea omogenității unui lot precum și între loturile luate într-o anumită experiență; prin măsurile mai sus menționate se înlătură factorii de eroare ce depind de animalul de experiență;
- ⇒ determinările comparative, prin folosirea standardelor sau a unor substanțe de referință, contribuie hotărâtor la înlăturarea factorilor de eroare care țin de tehnicile întrebuințate.

Existența variabilității biologice ne determină să admitem că între valoarea răspunsului biologic obținut de noi în urma unui experiment și valoarea reală poate exista o anumită diferență. Când datele obținute de către noi se abat de la valoarea reală putem spune că au fost comise *erori*, ce se datorează factorilor mai sus menționați sau altor factori necunoscuți. Prin termenul de *eroare* se înțelege diferența numerică dintre valoarea găsită de către experimentator și *valoarea reală (adevărată)* a unui parametru măsurat.

$$E = |M - A| \quad (I.1)$$

Unde E – eroarea absolută

M – valoarea măsurată

A – valoare adevărată

Această valoare “A” a unui sistem nu poate fi cunoscută exact. În cursul experimentului se obțin valori mai mult sau mai puțin apropiate de valoarea A; problema care se pune este însă care sunt valorile ce pot fi acceptate. Pentru că valoarea A nu poate fi cu certitudine cunoscută se urmărește ca o valoare acceptată să se găsească într-un anumit domeniu de valori, în sarcina statisticii matematice căzând stabilirea întinderii acestui domeniu și, deci, a validării rezultatelor unui experiment.

Factorii de eroare care pot fi înlăturați alcătuiesc așa-numitele *erori sistematice* și ele afectează *exactitatea* rezultatului. Factorii care țin de reactivitatea individuală, de exemplu, nu pot fi înlăturați, ei determinând ceea ce numim *erori aleatoare (întâmplătoare)*, care afectează atât *exactitatea* cât și *precizia* rezultatelor experimentale.

Cu toate că acești factori de eroare nu pot fi înlăturați în totalitate, variațiile pe care ei le provoacă în cadrul unui experiment dat pot fi apreciate și acestor variații li se adresează calculele de eroare. Deoarece aceste variații aleatoare se supun legilor de distribuție normală a frecvenței (despre care vom vorbi în detaliu ceva mai târziu), valoarea lor poate fi calculată. Odată acest lucru fiind realizat, precizia cercetărilor biologice poate fi confirmată științific.

Exactitatea – reprezintă apropierea valorii numerice determinate experimental de valoarea adevărată. Aceasta reprezintă de fapt *eroarea absolută*. Raportul

$$E_r \% = \frac{M - A}{M} \cdot 100 \text{ (I.2.)}$$

reprezintă *eroarea relativă*, exprimată uzual în procente.

Evident, cu cât rezultatul obținut se apropie mai mult de rezultatul real, cu atât determinarea este mai exactă.

Precizia unei determinări este dată de concordanța valorilor obținute în urma determinărilor efectuate. Se spune despre o metodă că este precisă când rezultatele determinărilor sunt reproductibile, adică sunt apropiate ca valoare în contextul unor experimente repetate. Concordanța între rezultate nu trebuie judecată numai prin prisma diferenței reale dintre ele, ci ținând cont și de mărimea, în valoare absolută, a acestora. De exemplu, să presupunem că, în cazul urmăririi perioadei de latență a inducerii somnului, pentru două hipnotice S_1 și S_2 , la șoareci, s-au obținut următoarele:

	<i>Șoarecele I</i>	<i>Șoarecele II</i>
<i>Substanța S_1</i>	<i>32 secunde</i>	<i>33 secunde</i>
<i>Substanța S_2</i>	<i>6 secunde</i>	<i>5 secunde</i>

Se observă că, în ambele cazuri, diferența între rezultate este de o secundă, însă concordanța între rezultate este mult mai bună în primul caz.

Deoarece valoarea reală nu poate fi cunoscută cu precizie se folosește în locul acesteia, în special în cazul distribuție normale de frecvență, media aritmetică a tuturor rezultatelor individuale:

$$\bar{x} = \frac{\sum_{i=1}^n X_i}{n} \quad (\text{I.3.}), \text{ unde } \quad \bar{x} \text{ media aritmetică a rezultatelor individuale}$$

$$n = \text{numărul determinărilor efectuate}$$

Pentru a putea aprecia *împrăștierea* rezultatelor unui experiment a fost definită **abaterea** (s):

$$s = x - \bar{x} \text{ (I.4.) unde } x - \text{rezultatul individual}$$

$$\bar{x} = \text{media aritmetică a rezultatelor individuale}$$

s este valoarea abaterii rezultatelor individuale de la valoarea medie, și arată precizia determinărilor.

În concluzie, când împrăștierea rezultatelor individuale față de medie este mică, iar media rezultatelor este apropiată de valoarea reală, eroarea determinării va fi mică, deci vom avea atât exactitate cât și precizie bună.

Erori mari survin atunci când rezultatele sunt mult dispersate față de valoarea medie (precizie slabă), iar media rezultatelor se îndepărtează mult față de valoarea reală (exactitate scăzută a metodei). Pot surveni și situații paradoxale, când, de exemplu, rezultatele individuale să fie mult dispersate față de medie (precizie redusă), dar media lor să fie totuși apropiată de valoarea reală (exactitatea metodei este bună).

Capitolul II

II. Distribuții de frecvență

[11-13]

Repartizarea datelor calitative și a celor cantitative (numerice) dintr-o populație statistică sau un eșantion se poate efectua după frecvența de apariție a caracteristicilor lor, obținându-se structura colectivității. De exemplu, o mulțime de date experimentale poate fi repartizată după “calitatea” efectelor observate: cu efect, fără efect sau cu efect gradat în funcție de doză.

Datele (măsurate pe întreaga populație statistică sau pe un eșantion) sunt de obicei organizate în așa-numitele *distribuții de frecvență*, pentru că o atare prezentare, reprezintă formatul cel mai convenabil de sinteză și prezentare a acestora.

În cazul distribuțiilor de frecvență se afectuează o descriere calitativă sau cantitativă a observațiilor (măsurătorilor) împreună cu numărul de apariții ale unui anumit rezultat al măsurătorii respective - *frecvența absolută*.

Se mai poate utiliza și *frecvența relativă* obținută prin împărțirea frecvenței absolute la numărul total al observațiilor.

De asemenea, pentru variabilele cantitative (numerice) continue este util să se construiască așa-numitele *intervale de variație*. Regulile generale de construire a acestor intervale sunt:

- numărul de intervale este bine să fie mai mic de 15
- limitele fiecărei clase să se potrivească cu gradul de acuratețe (precizie, exactitate) al măsurătorilor

- sunt de preferat intervalele de lungime egală, deoarece sunt mai convenabile, facilitând prelucrările ulterioare
- intervalele trebuie să fie mutual exclusive (capetele lor nu trebuie să se suprapună)

Se obține astfel o *distribuție de frecvență* a colectivității respective.

Se pot obține distribuții de frecvență *homograde* (cum este cazul diagramelor), cu o singură scară de comparație în sistemul cartezian, sau *distribuții heterograde*.

Pentru a fi mai expliciti, să luăm un exemplu:

Fie o serie de 33 de date numerice obținute experimental (de exemplu valori ale tensiunii arteriale sistolice): 180, 120, 110, 200, 140, 210, 200, 190, 150, 170, 140, 130, 150, 170, 160, 120, 160, 140, 160, 170, 180, 160, 150, 130, 160, 180, 190, 160, 170, 170, 150, 150, 130.

Cu această serie se poate alcătui o *diagramă*, așezând datele, în ordinea frecvenței, pe o singură scară a graficului cartezian (fig. 2.1.):

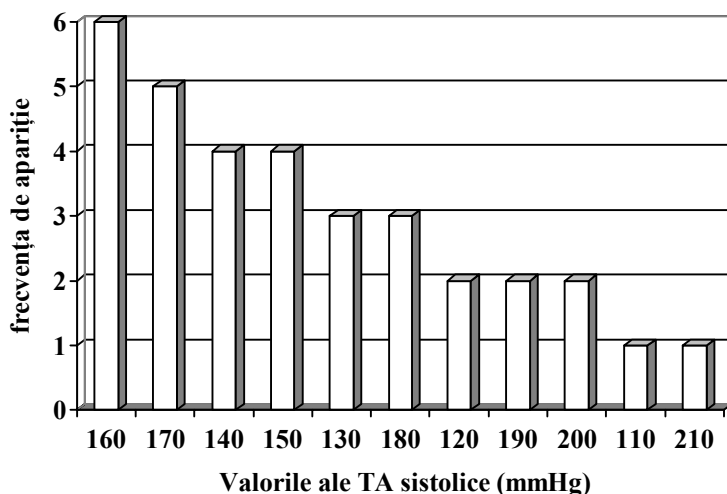


Fig. 2.1. Reprezentarea grafică a datelor sub formă de diagramă

În felul acesta se obține structura acestei colectivități și se poate constata frecvența de apariție a unor rezultate (de exemplu câte valori ale tensiunii arteriale 110 se găsesc în respectiva colectivitate statistică).

Căutând și ale modalități de caracterizare a colectivității, se poate stabili o distribuție de frecvență *heterogradă*, pe două scări ale sistemului cartezian, înșiruiind pe abscisă numerele, în ordine crescătoare sau descrescătoare și notând, în același timp, *intervalele de clasă*, iar pe ordonată punând frecvențele de apariție (relative sau absolute). Se obține astfel o *histogramă*. Prin unirea ordonatelor care trec prin mijlocul intervalelor de clasă se obține *poligonul de frecvență*.

Practic, pentru a reprezenta grafic corect datele colectate în cadrul unei histograme, trebuie parcurși următorii pași:

i. Pentru șirul valorilor măsurate se va determina *numărul de intervale de grupare (de clasă)* – M (eventual *lungimea intervalului de grupare* - d), conform formulei lui Sturges:

$$M = 1 + 3,22 * \log n \quad (II.1),$$

unde n = numărul măsurărilor efectuate

Valoarea numărului de intervale se va rotunji pentru a obține un număr întreg.

Pentru cazul nostru, numărul vom obține $M = 1 + 3,22 * \log 33 = 5,889$ adică vom avea 6 intervale.

Eventual, lungimea intervalului de grupare va fi calculată cu formula:

$$d = \frac{X_{\max} - X_{\min}}{1 + 3,22 * \log n} \quad (II.2),$$

unde $X_{\max}$, $X_{\min}$ reprezintă valoarea maximă și minimă măsurată.

Pentru cazul nostru $d = \frac{210 - 110}{1 + 3,22 * \log 33} = 16,97$. Dată fiind precizia

măsurărilor efectuate cu tensiometrele clasice, putem lua în considerare un o lungime a intervalului de grupare de circa 20 mmHg.

ii. În baza numărului calculat de intervale (M), respectiv a valorii lungimii intervalului de grupare (d), între valorile limită masurate ($X_{\min} - X_{\max}$), se vor stabili intervalele de grupare. Datele obținute vor fi centralizate într-un tabel.

iii. Se determină *frecvența absolută* (n_i), care reprezintă “numărul de apariții” a datelor corespunzătoare fiecărui interval de grupare în parte. Și aceste date se vor trece în tabelul sus-menționat.

iv. Se poate determina și *frecvența relativă* (f_i), care reprezintă raportul frecvenței absolute (n_i) la numărul total de măsurători (n)

v. Se construiește histograma, care reprezintă o diagramă în formă de “dreptunghiuri” având baza egală cu intervalul de grupare, iar înălțimea cu frecvența (absolută sau relativă)

vi. Se construiește poligonul de frecvență, care se obține prin unirea mijloacelor superioare ale histogramei prin segmente de dreaptă.

Tabelul cu intervale și frecvența absolută, precum și histograma rezultată în cazul celor 33 de valori ale tensiunii arteriale sistolice sunt prezentate mai jos:

<i>Interval de grupare</i>	<i>Frecvența absolută</i>
<110	1
111-130	5
131-150	8
151-170	11
171-190	5
>190	3

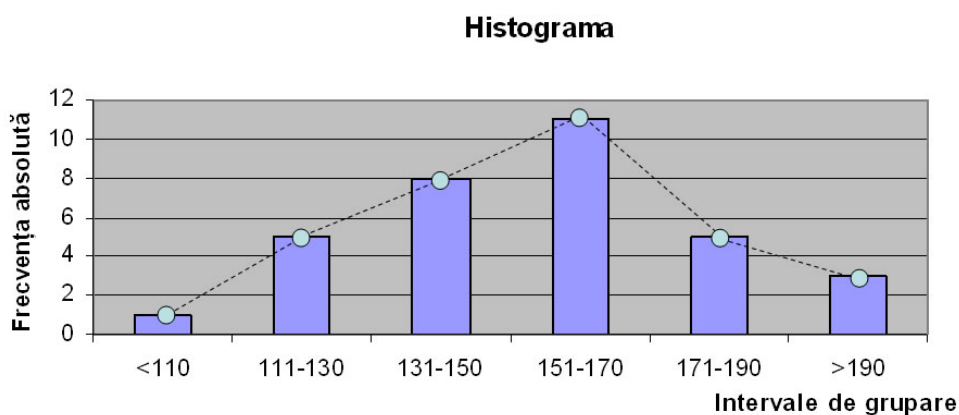


Fig. 2.2. Histograma

Distribuția de frecvență permite determinarea și a celorlalte caracteristici: *tendința centrală* (cu alte cuvinte, *media*), *mediana*, *forma distribuției*, *variabilitatea din interiorul ei*. În figura de mai sus se poate constata că forma distribuției se apropie destul de mult de distribuția normală Laplace-Gauss (distribuția în formă de clopot). În acest caz, valoarea medie reprezintă în cele mai bune condiții tendința centrală (pentru cazul studiat, media = 158,7).

Curba de distribuție normală

Van Vijnngaarden (1926) a arătat pentru prima dată că variația rezultatelor biologice se datorează sensibilității individuale a animalelor (care generează, astfel, erorile întâmplătoare) și că ele se supun legii de distribuție normală a frecvenței stabilită, în 1820, de Laplace și Gauss.

Curba de distribuție normală a frecvenței reprezintă frecvența cu care revine același rezultat în mai multe determinări succesive. Ea se poate obține așezând pe *abscisa* unui grafic diferențele, obținute în mai multe determinări, între media rezultatelor și rezultatele individuale, iar pe ordonată frecvențele de apariție a răspunsului pentru fiecare diferență. Graficul are forma unui clopot (fig. 2.3.):

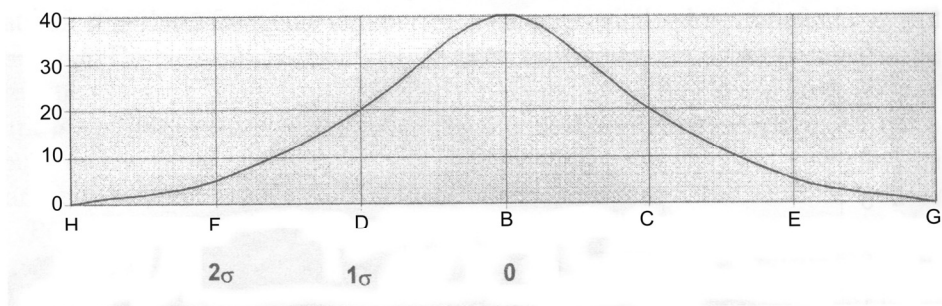


Fig. 2.3. Graficul de distribuție normală a frecvenței

Media rezultatelor individuale, care se repetă cel mai des (are cea mai mare frecvență de apariție), este punctul cel mai înalt al curbei. Valoarea medie este notată pe abscisă cu 0, deoarece diferența sa față de medie este, evident, zero. De o parte și de alta a punctului “culminant”, se desfășoară simetric frecvențele corespunzătoare diferențelor dintre media rezultatelor și rezultatele individuale, care se găsesc pe abscisă; cele negative (mai mici decât media) în partea stângă, cele pozitive (deci mai mari decât media) în partea dreaptă a valorii medii.

Distanța BD sau BC (0-1 sau 0+1) reprezintă convențional o unitate denumită *abatere standard* și notată cu σ (sigma). Perpendiculara pe valoarea medie este axul de simetrie al suprafeței acoperită de curbă. Perpendicularele în punctele de pe abscisă care corespund valorii medii plus abaterea standard și valorii medii minus abaterea standard, închid două treimi din suprafața acoperită de curbă (66%). Perpendicularele care corespund valorii medii plus sau minus 2σ închid circa 95% din suprafața acoperită de curbă. Suprafețele terminale ocupă numai 5 % din suprafața totală.

Pentru a demonstra că sensibilitatea animalelor de laborator față de o substanță activă se supune legilor de distribuție normală a frecvenței, Van Vijnngaarden a determinat doza minimă letală prin perfuzie lentă, cu aceeași soluție digitală, lucrând pe 573 de pisici. Efectuând calculele necesare, a obținut un grafic asemănător curbei în formă de clopot a lui Gauss (distribuția normală a frecvenței) și care se supune acelorași legi. S-a demonstrat astfel că variația sensibilității animalelor de experiență față de diferite substanțe medicamentoase studiate, se încadrează în teoria distribuției normale a frecvenței, stabilită de Gauss.

Cunoscând această lege, putem efectua experiențe și dozări biologice suficient de precise, din care să fie eliminate erorile determinate de reactivitatea individuală a animalelor de experiență.

Tot din această lege de distribuție normală a frecvenței reiese, însă, că valoarea științifică a unui singur rezultat obținut pe un animal sau a unor experimentări ce folosesc puține animale este mică, rezultate precise fiind cele obținute pe un număr mare de animale, prin stabilirea valorii medii a determinărilor și efectuarea unor prelucrări statistice ulterioare. Numărul mare de rezultate duce la obținerea unei curbe de frecvențe mai înaltă, micșorând, totodată, distanța dintre capetele curbei.

Caracteristicile distribuțiilor de frecvență

Orice serie de date cantitative se poate descrie prin două elemente caracteristice:

1. *indicatorii tendinței centrale*
2. *indicatori ai împrăștierii sau dispersiei datelor*

Dacă toate aceste elemente menționate sunt obținute în urma unui studiu populațional, ele poartă numele de *parametri statistici*, iar dacă sunt obținute în urma analizei unui eșantion se numesc *indicatori statistici*.

Din punctul de vedere al tendinței centrale, distribuțiile de frecvență se caracterizează prin *medie, mod (dominantă) și mediană*.

Media – reprezintă tendința centrală a unei distribuții (vom studia mai târziu modul ei de calcul).

Dominanta – reprezintă valoarea cea mai frecventă a unei distribuții, care se confundă, de fapt, cu vârful poligonului de frecvență.

Mediana – corespunde valorii care se găsește la punctul care împarte seria statisticii în două grupuri egale.

O altă caracteristică a distribuțiilor, care uneori este foarte folositoare, este *forma curbei de frecvență*. Ea poate fi:

1. simetrică

2. asimetrică: - *pozitivă*, caz în care „coada lungă” a distribuției este de partea valorilor pozitive.

- *negativă*, caz în care „coada lungă” a distribuției este de partea valorilor negative.

Două curbe cu aceeași medie, dominantă și aceeași mediană se pot deosebi după bază și înălțime: mai îngustă și mai înaltă sau mai largă și mai joasă. Întinderea bazei poate da o măsură a variabilității. Deschiderea este cu atât mai mare cu cât participarea factorilor întâmplători este mai mare (în figura 2.3) se pot observa diferențele între două curbe cu aceeași medie). Calculul precis la împrăștierea rezultatelor se face cu ajutorul *abaterii standard*.

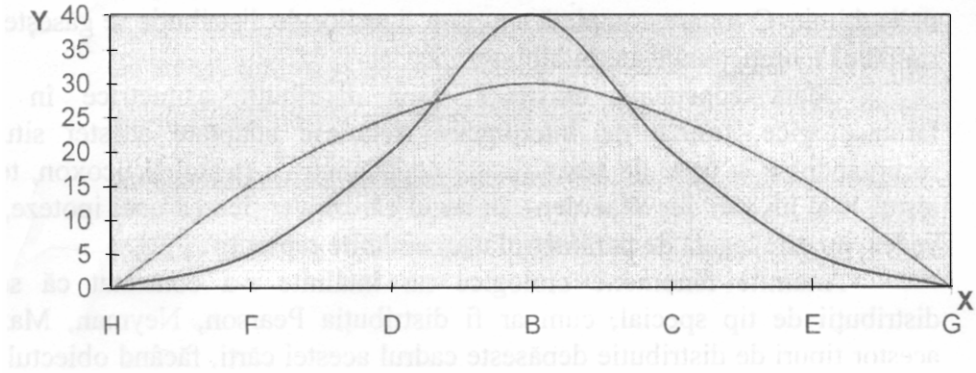


Fig. 2.3. Două curbe cu aceeași medie și dominantă, dar cu împrăștiuri diferite ale rezultatelor

În cazul distribuțiilor simetrice și unimodale există egalitate între media aritmetică, mod și mediană (este vorba despre curba normală de distribuție a frecvenței Laplace-Gauss).

↪ Distribuții anormale (non-Gaussiene)

Se cunosc, în afara distribuției normale unimodale, și distribuții purimodale sau asimetrice (fig. 2.4):

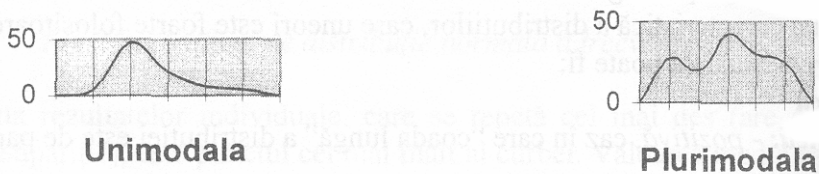


Fig. 2.4. Distribuții anormale

Pentru a înțelege mai bine importanța tipului de distribuție în cazul datelor provenite din cercetarea biomedicale trebuie spus că, de exemplu, existența unor distribuții anormale pot arăta o lipsă de omogenitate a afectelor farmacodinamice ale unui medicament. Unele distribuții pot lua forma literei U, unde importantă este valoarea minimă (de exemplu, în cazul acțiunii hipoglicemiante a unor substanțe active).

Există cazuri, însă, când fenomenele studiate se supun unei distribuții de tip special. Cele mai des întâlnite distribuții de acest tip au fost descrise de Bernouli (*distribuția binomială*) și de Poisson (*distribuția evenimentelor rare*).

Distribuția binomială prezintă interes mai ales în studiul fenomenelor eredității, iar *distribuția Poisson* în studiul unor efecte ale compușilor radioactivi și în radiochimie. O tratare completă a tuturor tipurilor de distribuție se găsește în tratatele de statistică teoretică indicate în bibliografie.

Odată constatată existența unor distribuții non-Gaussiene în cazul datelor provenite din cercetarea biomedicală, testele statistice de semnificație trebuiesc adaptate acestei situații, eventual recurgându-se la teste de semnificație neparametrice (testul Wilcoxon, testul U - Man Whitney etc), în cazul cărora verificarea unei ipoteze, așa cum vom vedea, nu este legată de parametrul unei anumite repartiții.

Anumite fenomene biologice rar întâlnite s-a constatat că se supun unor distribuții de tip special, cum ar fi distribuția Pearson, Neuman,

Maxwell. Studiul acestor tipuri de distribuție depășește cadrul acestei cărți, făcând obiectul unor manuale de specialitate.

În funcție de particularitățile distribuțiilor găsite, se pot alege procedeele matematice cele mai indicate pentru calculul statistic și interpretarea rezultatelor.

În cazul experimentării în domeniul farmacodinamic, rezultatele unei cercetări, odată reprezentate grafic dau distribuții empirice sau experimentale. Compararea acestor distribuții cu distribuțiile teoretice poate fi de un real folos pentru o interpretare justă a fenomenelor observate.

Trebuie menționat că majoritatea distribuțiilor obținute în urma analizei datelor rezultate din cercetarea biomedicală se supun legilor normale de repartiție a frecvenței, de aceea calculele și tehnicile de lucru prezentate în această carte se referă, în mod special, la această ipoteză .

Capitolul III

III. Statistică descriptivă. Considerații generale. Indicatori statistici

[11-13, 15,16, 21, 24]

Principalii indicatori care caracterizează o serie de date (un set de valori ale unui eșantion statistic) sunt fie *indicatori ai tendinței centrale*, fie *indicatori ce caracterizează împrăștierea datelor în jurul unei valori medii (indicatori ai dispersiei)*.

Deoarece acești indicatori “descriu” din punct de vedere statistic distribuția datelor studiate, permițând chiar unele comparații ale acestora cu distribuția normală, modalitățile de utilizare ale acestor indicatori statistici fac obiectul unei ramuri a statisticii denumită **statistică descriptivă**.

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

$$X_1, X_2, \dots, X_n.$$

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



A. Indicatori ai tendinței centrale

Media aritmetică - notată de regulă cu $\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$ (III.1),

Mediana - este acea valoare din șirul de date care împarte în două părți egale șirul ordonat de valori (atenție, șirul este ordonat crescător), situându-se la mijlocul seriei statistice. Dacă numărul de valori n este un număr impar, atunci mediana este valoarea

$$M_e = x_k \text{ (III.2), unde } k = \frac{n}{2} + 1.$$

Dacă n este par, deci avem un număr par de valori, mediana este definită ca fiind $M_e = \frac{x_k + x_{k+1}}{2}$ (III.3), unde $k = n/2$.

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

B. Indicatori ai dispersiei (împrăstierii) datelor în jurul valorii medii

Varianta notată s_x^2 este un indicator de împrăstiere a datelor. Formula de calcul este: $s_x^2 = \frac{\sum x_i^2 - (\bar{x})^2}{n-1}$ (III.4),.

Abaterea standard sau deviația standard reprezintă rădăcina pătrată din varianță (dispersie) : $s_x = \pm \sqrt{s_x^2}$ (III.5),

Coefficientul de variație se calculează ca un raport procentual între abaterea standard și valoarea medie a șirului de valori.

$$CV\% = \frac{s_x}{\bar{x}} \cdot 100 \text{ (III.6),}$$

Asimetria (skewness) caracterizează gradul de “asimetrie” a unei distribuții în jurul valorii medii, comparativ cu distribuția normală. Valori pozitive ale asimetriei indică o distribuție de frecvență ce prezintă o “coadă” (în engleză tail) asimetrică în zona valorilor “pozitive” ale distribuției (valori mai mari decât media). Similar, valori negative ale asimetriei indică o distribuție de frecvență ce prezintă o “coadă” (în engleză tail) asimetrică în zona valorilor “negative” ale distribuției (valori mai mici decât media).

$$Skewness = \frac{n}{(n-1) \cdot (n-2)} \cdot \sum \left(\frac{x_i - \bar{x}}{S} \right)^3 \quad (III.7), \text{ , unde } S = \text{abaterea}$$

standard.

Aplatizarea (kurtosis) caracterizează gradul de “aplatizare” a unei distribuții, comparativ cu distribuția normală. Valorile pozitive ale acestui indicator indică o distribuție cu un “vârf” mai înalt decât distribuția normală. Similar, valori negative ale kurtosisului indică o curbă relativ aplatizată, comparativ cu distribuția normală.

$$Kurtosis = \left\{ \frac{n(n+1)}{(n-1)(n-2)(n+3)} \cdot \sum \left(\frac{x_i - \bar{x}}{S} \right)^4 \right\} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

(III.8), unde S = abaterea standard.

Amplitudinea - este diferența dintre valoarea maximă și cea minimă

$$A = A_{max} - A_{min} \quad (III.9).$$

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

Atunci când avem foarte multe date se recomandă includerea lor în clase egale ca mărime, ceea ce ușurează 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 aceleiași valori

Dacă datele pe care le studiem conțin valori care se repetă des, se obișnuiește să se grupeze datele care au aceeași valoare . Numărul de

apariții ale unei valori anume se numește frecvența de apariție și se notează cu f_i .

Presupunem că în urma măsurărilor am obținut șirul 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

$$\bar{x} = \frac{\sum_{i=1,n} x_i \cdot f_i}{\sum_{i=1,n} f_i} = \frac{x_1 \cdot f_1 + x_2 \cdot f_2 + \dots + x_n \cdot f_n}{f_1 + f_2 + \dots + f_n} \quad (III.10)$$

Mediana – este x_k (III.11,) unde $k = \frac{\sum_{i=1,n} f_i + 1}{2}$

$$\textbf{Dispersia (varianta)} : s_x^2 = \frac{\sum_{i=1,n} (x_i - \bar{x})^2 \cdot f_i}{\sum_{i=1,n} f_i - 1} \quad (III.12)$$

Data fiind importanța lor, în cele ce urmează vor fi descriși pe larg unii dintre indicatorii statistici enumerați mai sus.

III.1. Medii

Valoarea medie definește cel mai bine tendința centrală a unei distribuții de frecvență. Totuși trebuie menționat că valoarea medie nivelează variațiile valorilor prin obținerea unei valori mijlocii, care dă impresia unei stabilități a fenomenelor, care nu este reală în biologie, de aceea pentru un studiu complet al unei compatibilități statistice, sunt necesare metode statistice care stabilesc variațiile rezultatelor obținute și care, pentru o bună interpretare, trebuie să însoțească valoarea medie.

Cea mai uzuală în statistică este media aritmetică care corespunde formulei de mai jos:

$$M_a = \bar{x}_a = \frac{1}{n} \sum_{i=1}^n x_i \quad (III.13)$$

Media aritmetică poate fi calculată în mai multe feluri:

Media aritmetică simplă

Calculată după formula de mai sus. Pentru exemplul din capitolul II, unde suma celor 33 de rezultate individuale era 5240, media aritmetică simplă este:

$$\bar{x}_a = \frac{5250}{33} = 159,09 \text{ mmHg}$$

Media aritmetică ponderată

Dacă, pentru exemplul din capitolul II, se iau în considerare frecvențele cu care vin numerele, se observă că ele au însemnătate inegală,

numerele 110 și 210 revin numai o singură dată, în timp ce 140 sau 150 revin de patru ori în șirul de date exprimatele studiat. În acest caz, se spune că valorile nu au pondere egală, iar numărul (frecvența) care arată de câte ori se repetă fiecare valoare va fi ponderea valorii respective. Se poate calcula media, ținând seama de aceste ponderi, după formula:

$$\bar{x}_a = \frac{\sum_{i=1}^n p_i x_i}{\sum_{i=1}^n p_i} \quad (III.14)$$

Se face, deci, suma produselor dintre fiecare valoare și ponderea sa și se împarte la suma ponderilor.

Există un procedeu matematic pentru a determina media care ușurează calculul atunci când avem de-a face cu serii statistice alcătuite din numere mari. De exemplu, dacă variabilele studiate sunt reprezentate de masa corporală a șoarecilor unei biobaze, iar frecvența lor fiind reprezentată de familii de șoareci, să calculăm greutatea medie a șoarecilor pe familie:

Procedeeul se realizează prin alegerea unei medii arbitrare, notată cu a (frecvența cea mai mare), în timp ce cu x notăm abaterile fiecărei valori de la originea arbitrară ($-1, +1$ etc). Frecvența o notăm cu f .

Formula de calcul a mediei în acest caz este:

$$\bar{x}_a = a \pm \frac{\sum f \cdot x}{\sum f} \quad (III.15)$$

Se face astfel produsul dintre fiecare valoare x și fiecare frecvență f și, deoarece se obțin numere pozitive și negative, se face suma lor algebrică, care se împarte la suma frecvențelor. În final, valoarea obținută se va scădea din a .

III. 2. Dispersia (varianța). Abaterea standard

Așa cum am mai menționat, *dispersia (varianța)* notată s_x^2 este un indicator de împrăștiere a datelor. Formula ei de calcul este:

$$S_x^2 = \frac{\sum x_i^2 - (\bar{x})^2}{n-1} \quad (III.16)$$

Unitatea de abatere individuală față de medie a fost denumită *abaterea standard* și a fost notată cu S (σ se utilizează numai în cazul curbei ideale). Ea reprezintă o măsură a preciziei determinărilor sau, cu alte cuvinte, o măsură a împrăstierii rezultatelor individuale față de medie.

Abaterea standard poate fi calculată după următoarea formulă:

$$S = \frac{\sum (x_i - \bar{x})}{n} \quad (3.17) \text{ unde } x_i - \bar{x} = \text{abaterea valorilor individuale}$$

Față de valoare medie (indiferent de semn) se notează cu d (diferența).

Deoarece cercetarea biologică se bazează pe eșantionaj, abaterea standard se calculează în acest caz după formula:

$$S = \frac{\sum (x_i - \bar{x})}{n-1} \quad (III.18)$$

Practic abaterea standard se calculează ca rădăcina pătrată din dispersie (varianță) :

$$S_x = \pm \sqrt{S_x^2} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} \quad (III.19)$$

Făcând pătratul diferenței, se evită posibilitatea ca aceste diferențe în plus sau în minus să se anuleze, obținând numere în valoare absolută. Aceasta obligă, însă să se extragă rădăcina pătrată pentru a obține rezultatul. Se calculează, prin urmare, pătratul fiecărei diferențe față de medie și se face suma acestor pătrate, care se împarte la numărul determinărilor minus 1. Rădăcina pătrată a acestei valori este abaterea standard S .

Cunoscând *abaterea standard*, adică răspândirea răspunsurilor individuale față de medie, se cunoaște *precizia* determinărilor. Totodată, deoarece pentru curba lui Gauss, dublul abaterii standard este reprezentat de suprafața închisă de perpendicularele care trec prin punctele BF și BE și care acoperă 95% din suprafața totală, va trebui să ținem cont de aceasta, luând $2S$ ca și interval de încredere (vezi figura 2.3.). Aceasta ne va certifica faptul că 95% din rezultatele noastre experimentale se vor încadra în limitele calculate și numai 5% din ele se vor găsi în afara acestor limite.

III.3. Eroarea standard

Este cunoscut faptul că determinările biologice sunt supuse influenței a două tipuri de erori: cele care influențează *precizia* determinării și cele care influențează *exactitatea* determinării. Pentru a afla *exactitatea* cu care s-au făcut o serie de determinări, trebuie să se calculeze *abaterea medie a valorilor medii obținute* sau, altfel spus, media erorilor ce se pot comite într-o determinare.

Această abatere a fost denumită *eroare standard*, notată cu E . Calcularea ei se face cu ajutorul formulei:

$$E^2 = \frac{\sum (x_i - \bar{x})^2}{n(n-1)} \quad (III.20)$$

Știind că, în cazul distribuției normale gaussiene, împrăștierea în jurul mediei colectivității a unei medii de șantion este $\sqrt{n}$ ori mai mică decât împrăștierea rezultatelor individuale, eroarea standard este dată și de formula:

$$E = \frac{S}{\sqrt{n}} \quad (III.21)$$

Ea reprezintă formula clasică a erorii standard.

Rezultatele experimentărilor biologice trebuie să fie însoțite întotdeauna de eroarea standard sau de abaterea standard, utilizându-se formulări de tipul $M \pm S$ sau $M \pm E$, pentru a permite o justă interpretare a lor.

III.4. Eroarea procent

După cum știm, majoritatea efectelor farmacodinamice se pot încadra în două categorii: efecte gradate și, respectiv, efecte cuantale. De multe ori acestea din urmă sunt reprezentate sub formă de procente. Atunci când eșantionul este mare, putem spune că procentele (pe) sunt distribuite normal în jurul mediei cu o abatere standard egală cu

$$S = \sqrt{\frac{p \cdot q}{n}} \quad (III.22) \quad \text{unde } p = \text{procentajul de răspuns pozitiv}$$

$q = \text{procentajul de acțiune negativ}$

$n = \text{numărul cazurilor}$

evident, $q = 100 - p$

Putem spune deci că abaterea standard a unui procent de acțiune calculat cu această formulă reprezintă limitele probabile, în plus sau în minus, ale procentajului de acțiune pentru o doză dată de substanță activă.

III.5. Coeficientul de variație

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

$$CV\% = \frac{sx}{\bar{x}} \cdot 100 \quad (III.23)$$

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) și de unitățile de măsură folosite. Se poate considera că un coeficient de variație sub 10% indică o dispersie mică (o împrăștiere) a datelor, adică seria este omogenă. Un coeficient între 10% și 30% indică dispersie mijlocie, iar peste 30% indică dispersie mare. *Dacă dispersia este mare, media nu este un indicator reprezentativ.*

III.6. Grade de libertate

Din cele discutate până acum am văzut cum, plecând de la un eșantion al unei colectivități, am înlocuit abaterea standard teoretică (σ) prin

abaterea standard de eșantionaj (S). De asemenea, ca factor de corecție s-a folosit pătratul diferențelor individuale (d^2) și s-a calculat S^2 . În aceeași ordine de idei, pentru a putea apropia pe S de s (abaterea teoretică) am diminuat numărul cazurilor din experiment cu o unitate, în locul efectivului total „ n ” punând „ $n-1$ ”.

Practic formula de calcul a abaterii standard a devenit

$$S = \sqrt{\frac{\sum d^2}{n-1}} \quad (III.24)$$

Spunem că „ $n-1$ ” este *numărul gradelor de libertate*.

Gradele de libertate reprezintă, în cazul determinărilor biologice *numărul mărimilor* (animale, determinări, observații) folosite în experimentarea respectivă, din care se scade o unitate.

Pornind de la această premiză, **gradele de libertate** reprezintă practic *numărul mărimilor independente* folosite în experimentarea respectivă.

Ținând seama de cele afirmate mai sus, în calculele de determinare a erorii va interveni un factor de corecție „ t ”, care depinde de numărul gradelor de libertate (**$tn-1$**). Cantitatea „ t ” se găsește în tabele (vezi Anexa 1), calculată pentru diferite probabilități, în funcție de numărul de grade de libertate folosit (în general vom lucra cu $p=0,05$). Valoarea lui „ t ” scade cu cât crește numărul observațiilor, deci cu cât este mai mare numărul gradelor de libertate.

În cazul determinărilor comparative martor/probă sau a mai multor doze (loturi) se scade din efectivul total câte o unitate pentru fiecare lot. (de

exemplu numărul gradelor de libertate pentru două loturi, în cazul mai sus menționat va fi (n_1+n_2-2) .

III. 7. Limite fiduciale (interval de încredere)

Am văzut până acum că principalii parametrii care descriu o populație statistică sunt media și abaterea standard. În practică parametrii unei populații se estimează pe baza determinărilor efectuate pe eșantioane luate din respectiva populație statistică. Evident parametrii probelor extrase nu sunt perfect identici cu cei ai populației studiate; există însă posibilitatea de a calcula intervalul în care se pot încadra acești parametrii, acordând acestui interval o anumită „încredere” (probabilitate), aleasă în funcție de exactitatea dorită (de obicei se alege un nivel de probabilitate de 95% sau 99%). Aceasta înseamnă că, dacă vom lua un număr mare de probe din aceeași populație, 95% respectiv 99% din probe vor avea parametrii care se încadrează în intervalul calculat și va exista riscul ca 5% respectiv 1% din proces să se găsească în afara intervalului calculat.

Limitele fiduciale, denumite și limite de eroare sau de securitate, reprezintă intervalul în care se poate prevedea că se găsește valoarea unei medii (atât în cazul efectelor gradate cât și a celor cuantale). Intervalul respectiv se mai numește și *interval de încredere*.

Limitele de eroare sunt, în general, proporționale cu valoarea mediei și pot fi convențional exprimate ca procente ale acestei medii. De regulă, în determinări biologice calculăm limitele de eroare la o probabilitate $p = 0,05$.

În calculul limitelor de eroare se folosește factorul de corecție „t”, despre care știm că depinde de numărul gradelor de libertate.

Pentru o interpretare corectă, rezultatele experimentărilor biologice trebuie exprimate după relația de mai jos:

$M \pm tS$ (III.25) unde M – media determinărilor

t – factorul de corecție pentru probabilitatea
dorită

S – abaterea standard

Dacă abaterea standard (S) este exprimată în procente limitele de eroare sunt $100 \pm tS$ la sută.

Dacă folosim calculul logaritmico pentru calcularea abaterii standard în anumite experimente farmacologice, limitele de eroare sunt date de *antilogaritmul* lui $2 \pm tS$.

Să presupunem că la testarea unui anumit analgezic, în urma experimentării prin testul plăcii încălzite, timpul de latență al reacției nociceptive a fost de 100 ± 20 secunde, la o probabilitate $p = 0,05$. Limitele fiduciale sunt deci cuprinse între 80-120 secunde. Aceasta înseamnă că în 95 de determinări din 100 rezultatul găsit va fi superior timpului de 80 secunde și inferior timpului de 120 secunde, oscilând în jurul valorii celei mai probabile (media $M=100$ secunde). Cu alte cuvinte, dacă se repetă determinarea în aceleași condiții, rezultatul se va găsi în 95% din cazuri între aceste limite și numai în 5% din cazuri valoarea experimentală va fi în afara acestor limite.

Capitolul IV

IV. Statistică inferențială. Interpretarea statistică a rezultatelor unui experiment

[1, 3, 5, 6, 10, 11-13, 15, 20, 23,24]

Odată obținute rezultatele experimentale, ele trebuie prelucrate, prezentate, și, mai ales sintetizate și interpretate, pentru a putea desprinde legalitatea urmărită. Trebuie acordată o importanță deosebită interpretării rezultatelor experimentale obținute, deoarece o interpretare prea simplistă sau, dimpotrivă, prea pretențioasă poate duce la o scădere a valorii cercetării efectuate.

Sprijinul hotărâtor în interpretarea corectă a rezultatelor și în afirmarea concluziilor îl aduce statistica matematică. Concluziile unui anumit experiment trebuie verificate, datele experimentale trebuie să fie reproductibile, concluziile trase trebuie să fie științific întemeiate, acest lucru nefiind posibil fără o prelucrare statistică adecvată a datelor.

Totalitatea testelor statistice folosite cu scopul de a demonstra existența sau inexistența unor legături sau diferențe statistic semnificative, respectiv științific corecte între variabilele unui anumit studiu sau experiment, fac obiectul unei ramuri a statisticii numite **statistică inferențială**.

Se evidențiază în principal patru categorii de teste/analize statistice folosite cu succes în descrierea/interpretarea rezultatelor unui experiment științific:

1. *Teste de valabilitate*
2. *Teste de concordanță (goodness of fit)*

3. *Teste de semnificație*

4. *Corelații și regresii*

IV.1. Teste de valabilitate

Noțiuni introductive

Definiție: Prin *teste de valabilitate* se înțeleg testele care permit aflarea valorii unei experimentări în funcție de probabilitățile apariției variației în natură.

Cu alte cuvinte, prin aceste teste putem preciza dacă un anumit rezultat experimental poate fi considerat ca fiind datorat variabilității biologice normale și care nu va îndepărta semnificativ rezultatul final al experimentului de “valoarea adevărată” sau este un *rezultat aberant* care se datorește unor factori accidental apăruiți în timpul experimentului (animale bolnave, tarate etc).

Rezultate aberante (outliers)

De foarte multe ori într-un lot de rezultate experimentale apar cazuri foarte îndepărtate față de celelalte. Problema care se pune este dacă aceste rezultate pot sau nu pot fi luate în considerare, având în vedere faptul că influențează semnificativ valoarea finală a mediei. Există tendința de a elimina din start aceste rezultate, care par întâmplătoare, deoarece se consideră că ele deviază media într-un sens care poate fi foarte departe de valoarea reală. Acest mod simplist de a raționa este total greșit.

Biostatistica pune la dispoziție o serie de criterii de apreciere a acestei situații și de eliminare a valorilor care se abat foarte mult de la medie (așa numitele *rezultate aberante - outliers în engleză*).

Metodele cele mai utilizate în practică sunt *testul Grubbs*, respectiv *criteriul de eliminare a lui Chauvenet*, ce folosesc modalități de calcul oarecum similare.

IV.1. 1. Testul Grubbs

De-a lungul timpului statisticienii au găsit mai multe căi de a detecta și elimina rezultatele aberante. Testul Grubbs este doar una dintre aceste metode, având însă avantajul că este ușor de aplicat.

Primul pas în aplicarea acestui test constă în cuantificarea “îndepărtării” rezultatului considerat “aberrant” de celelalte rezultate.

Practic, vom calcula raportul Z , constând în împărțirea diferenței dintre valoarea mediei și valoarea rezultatului aberrant la valoarea abaterii standard (notată cu S sau SD – standard deviation). Trebuie subliniat aici faptul că abaterea standard este calculată incluzând și valoarea rezultatului considerat aberrant !

$$Z = \frac{|\bar{X} - X_i|}{SD} \quad (IV.1), \text{ unde } \bar{X} - \text{media valorilor}$$

X_i – valoare rezultatului considerat
“aberrant”

SD - abaterea standard

Dacă valoarea astfel calculată lui Z este mare, rezultatul respectiv este destul de “îndepărat” de celelalte rezultate.

Este știut faptul că atunci când analizăm date experimentale, nu cunoaștem aprioric abaterea standard a populației statistice studiate. Dimpotrivă, abaterea standard este calculată chiar din datele obținute în urma experimentului. Din această cauză, prezența unui “rezultat aberant” va crește valoarea abaterii standard calculate.

Deoarece prezența unei valori aberante va crește atât numărătorul (valoare absolută a diferenței dintre medie și valoarea aberantă), cât și numitorul (abaterea standard a tuturor valorilor) în cazul formulei de mai sus, valoarea lui Z nu poate fi mai mare decât $\frac{(n-1)}{\sqrt{n}}$, unde n – numărul de valori experimentale. De exemplu, dacă $n=5$, Z nu poate fi mai mare decât 1,789, pentru orice set de date experimentale.

Ca și în cazul altor teste statistice și pentru testul Grubbs au fost calculate un set de *valori critice* pentru Z . Evident, valoarea critică va crește cu mărimea eșantionului (n), așa cum se poate vedea din tabelul următor:

Tabel IV.1. *Valoarile critice ale lui Z*

n	Valoarea critică a lui Z	n	Valoarea critică a lui Z
3	1.15	27	2.86
4	1.48	28	2.88
5	1.71	29	2.89
6	1.89	30	2.91
7	2.02	31	2.92
8	2.13	32	2.94

9	2.21	33	2.95
10	2.29	34	2.97
11	2.34	35	2.98
12	2.41	36	2.99
13	2.46	37	3
14	2.51	38	3.01
15	2.55	39	3.03
16	2.59	40	3.04
17	2.62	50	3.13
18	2.65	60	3.2
19	2.68	70	3.26
20	2.71	80	3.31
21	2.73	90	3.35
22	2.76	100	3.38
23	2.78	110	3.42
24	2.8	120	3.44
25	2.82	130	3.47
26	2.84	140	3.49

Dacă valoarea particulară calculată pentru Z în cazul rezultatului considerat “aberrant” este mai mare decât valoarea critică din tabelul de mai sus, atunci P este mai mic decât 0,05. Aceasta înseamnă că există mai puțin de 5% șanse ca, datorită întâmplării, să întâlniți o valoare “aberrantă” mai îndepărtată de celelalte valori (în orice direcție), câtă vreme datele provin dintr-un eșantion extras dintr-o populație gaussiană. În acest caz, dacă $P < 0,05$, rezultatul poate fi considerat “aberrant” și poate fi eliminat.

De remarcat faptul că testul Grubbs poate fi aplicat cu succes numai la valorile extreme (în ambele sensuri) întâlnite în cazul eșantionului studiat. Exemplu:

A fost măsurată greutatea a 15 indivizi adulți. Rezultatele măsurătorilor sunt cele din tabelul următor:

Tabel IV.2. Rezultatele măsurătorilor

Nr. crt.	Greutatea (kg)
1	58
2	60
3	80
4	77
5	83
6	75
7	82
8	79
9	50
10	35
11	70
12	160
13	80
14	65
15	55
Media =73,93 kg Abaterea standard = 27,59 n = 15	

Aplicând formula de calcul a lui Z pentru valoarea $n_{12}=160$ kg obținem $Z = \frac{|73,93 - 160|}{27,59} = \frac{86,07}{27,59} = 3,12$, o valoare mai mare decât valoarea critică pentru $n=15$, care este 2,55, deci $P<0,05$. În concluzie, pentru eșantionul studiat o valoarea greutatei corporale de 160 kg poate fi considerat rezultat aberant, putând fi luată în calcul eliminarea acestuia.

IV.1. 2. Criteriul de eliminare a lui Chauvenet

Se bazează pe considerentul că orice valoare a cărei probabilitate de apariție este mai mică decât o valoare limită care depinde de numărul „n” de rezultate, trebuie eliminată. (*Farmacopeea Română ediția a X-a*).

Pentru aplicarea acestui criteriu la eliminarea unor rezultate aberante trebuie urmați pașii de mai jos:

- se calculează abaterea standard (S sau SD) a șirului de valori, conform formulei menționate în capitolele anterioare;
- din tabelul de mai jos, se obține valoarea raportului x/s , în funcție de numărul „n” de rezultate;
- se înmulțește valoarea acestui raport cu valoarea abaterii standard (S), obținându-se astfel valoarea x , care reprezintă valoarea absolută maximă pe care o poate avea d_i (unde d_i – abaterea față de medie, $d_i = x_i - \bar{x}$), pentru ca valoarea experimentală respectivă să nu fie eliminată. Orice valoare căreia îi corespunde o abatere față de medie, în mărime absolută, mai mare decât x ($|d_i| > x$), trebuie eliminată.

Dacă printre valorile rămase după aplicarea de eliminare se consideră, că mai există o valoare ce ar trebui eliminată, se aplică criteriul încă o dată. În general, se repetă aplicarea criteriului de eliminare de câte ori este necesar.

Tabelul IV.3. Valoarea raportului x/s folosit pentru criteriul de eliminare

n	x/s	n	x/s	n	x/s
5	1,68	14	2,10	30	2,39

6	1,73	16	2,16	40	2,50
7	1,79	18	2,20	50	2,58
8	1,86	20	2,24	100	2,80
9	1,92	22	2,28	200	3,02
10	1,96	24	2,31	500	3,29
12	2,03	26	2,35		

Vom exemplifica aplicarea criteriului de eliminare în cazul unor valori (x_i) ale timpului de latență a instalării efectului hipnotic în cazul amobarbitalului:

Tabelul IV.4. Prima aplicare a criteriului de eliminare

Nr. crt.	x_i (secunde)	d_i	d_i^2	
1	16,1	3,6	12,96	$s = \sqrt{\frac{212,56}{9}} = 4,86$ $x/s = 1,96$ $x = 1,96 \times 4,86 = 9,53$
2	15,5	3,0	9,00	
3	13,4	0,9	0,81	
4	22,8	10,3	106,09	
5	12,1	-0,4	0,16	
6	11,3	-1,2	1,44	
7	11,6	-0,9	0,81	
8	6,3	-6,2	38,44	
9	8,8	-3,7	13,69	
10	7,1	-5,4	29,16	
$\sum x_i = 125,00 \quad \sum d_i^2 = 212,56$ $\bar{x} = 12,5 \text{ (media aritmetica a rezultatelor)}$				

După cum rezultă din tabel, diferența $d_i = 10,3$, corespundența valorii de 22,8 secunde depășește valoarea maximă admisă ($x = 9,53$); prin urmare, valoarea respectivă va trebui să fie eliminată din datele supuse prelucrării.

Tabelul IV.5. *A doua aplicare a criteriului de eliminare*

Nr. crt.	x_i (secunde)	d_i	d_i^2	
1	16,1	4,7	22,09	$s = \sqrt{\frac{94,7}{8}} = 3,4$ $x/s = 1,92$ $x = 1,92 \times 3,4 = 6,5$
2	15,5	4,1	16,81	
3	13,4	2,0	4,00	
4	12,1	0,7	0,49	
5	11,3	-0,1	0,01	
6	11,6	0,2	0,04	
7	6,3	-5,1	26,01	
8	8,8	-2,6	6,76	
9	7,1	-4,3	18,49	
$\sum x_i = 102,2 \qquad \sum d_i^2 = 94,70$ $\bar{x} = 11,4 \text{ (media aritmetica a rezultatelor)}$				

Aplicarea criteriului de eliminare a doua oară conduce la o valoare maximă admisă ($x = 6,5$) superioară oricărui d_i , deci nu va mai fi necesară eliminarea nici unei valori.

Efectuarea unei analize, folosind un eșantion adecvat, nu poate conduce însă, de cele mai multe ori, la determinarea mediei adevărate a populației statistice din care face parte acel eșantion. În schimb se pot găsi, cu o anumită probabilitate, limitele între care se află valoarea medie adevărată. În acest scop se calculează mai întâi abaterea standard a medie eșantionului (S_x), conform formulei de mai jos:

$$S\bar{x} = \frac{s}{\sqrt{n}} \quad (\text{IV.2})$$

În continuare, intervalul de încredere al mediei (J) se stabilește pentru o probabilitate de eroare dorită, de obicei 5% (altfel spus $p = 0,05$)

folosind valoarea „t” Student, a cărei valori corespunzătoare gradelor de libertate ale determinării, pentru $p = 0,05$, sunt date în tabelul din Anexa 1:

Se aplică formula:

$$J = \bar{x} \pm t S \bar{x} \text{ (IV.3)}$$

Gradele de libertate sunt reprezentate de numărul mărimilor independente ale determinării. În cazul de față, gradele de libertate se calculează scăzând din numărul total de valori (x_i) cifra 1 (numărul de loturi).

IV.2. Teste de concordanță (Goodness-Of-Fit)

Noțiuni introductive

În speță, mai ales în domeniul biostatisticii, suntem în primul rând interesați dacă distribuția de frecvență a populației din care sunt prelevate eșantioanele analizate de noi urmează sau nu modelul “ideal” al unei distribuții normale (distribuție Gaussiană).

Ce înțelegem prin distribuție Gaussiană ?

Așa cum am arătat și în capitolele anterioare, chiar dacă o multitudine de factori aleatori acționează în mod convergent, creând variabilitate, de cele mai multe ori distribuția de frecvență a datelor investigate urmează mai mult sau mai puțin o așa-numită distribuție în formă de “clopot”, numită distribuție normală sau distribuție Gaussiană.

Acest tip de distribuție are o serie de “proprietăți matematice” speciale ce stau la baza mării majorități a celor mai utilizate teste statistice (așa-numitele teste statistice parametrice – testul t , testul F , testul CHI^2 etc).

Chiar dacă - în general – datele colectate de noi în urma investigării unui eșantion extras din populația statistică studiată nu urmează acest model matematic “ideal”, de cele mai multe ori aceste date urmează o distribuție ce este “aproximativ” Gaussiană.

Distribuția Gaussiană (normală) joacă un rol central în analiza statistică datorită unei legități matematice cunoscute ca Teorema Limitei Centrale (Central Limit Theorem).

La modul intuitiv, putem afirma că această teoremă statuează că dacă o serie de variabile măsurate au o varianță “finită”, atunci distribuția de frecvență a acestor variabile poate fi aproximată ca fiind o distribuție normală (Gaussiană).

Cu alte cuvinte, conform acestei teoreme, putem afirma că dacă eșantioanele studiate sunt suficient de mari, distribuția de frecvență în jurul mediilor acestor eșantioane va urma modelul unei distribuții gaussiene, chiar dacă în sine populația statistică din care au fost prelevate aceste eșantioane nu este Gaussiană.

De vreme ce mai multe teste statistice (de exemplu testul t sau analiza de varianță - ANOVA) sunt teste ce cuantifică existența unor diferențe între medii, Teorema Limitei Centrale permite acestor teste să furnizeze rezultate pertinente, chiar dacă populațiile statistice din care provin eșantioanele studiate nu sunt Gaussiene.

Condiția necesară și suficientă este ca eșantioanele respective să fie suficient de mari ! Cât de mari anume ? Din nefericire, acest lucru depinde, în speță, de cât de diferită este distribuția populației studiate de distribuția Gaussiană.

La modul general, testele de concordanță sunt teste ce se permit luarea unei decizii asupra faptului că un eșantion provine sau nu dintr-o populație statistică ce se supune unui anumit tip de distribuție de frecvență.

În cazul biostatisticii, principala utilizare a unor astfel de teste constă în verificarea faptului că eșantioanele studiate sunt prelevate dintr-o populație statistică ce se supune sau nu legilor unei distribuții normale (Gaussiene).

Există mai multe astfel de teste de concordanță:

- Testul de concordanță Kolmogorov-Smirnov
- Testul de concordanță Anderson-Darling
- Testul de concordanță Chi-Square
- Testul de normalitate Shapiro-Wilk etc

Deoarece multe dintre programele de analiză statistică utilizează numai unul sau două dintre aceste teste (în general testul Kolmogorov-Smirnov) ne vom referi în continuare la acesta.

IV.2.1. Testul de concordanță Kolmogorov-Smirnov

Testul Kolmogorov-Smirnov (Chakravart, Laha and Roy, 1967) este folosit pentru a decide dacă un anumit eșantion provine dintr-o populație statistică ce se supune unui anumit tip de distribuție de frecvență.

Testul Kolmogorov-Smirnov (K-S) se bazează pe așa-numita funcție cumulativă de distribuție empirică (ECDF - The Empirical Cumulative Distribution Function).

Fiind dat un număr N de puncte $Y_1, Y_2, \dots, Y_N$, ECDF poate fi definită ca

$$E_N = n(i)/N \quad (\text{IV.4})$$

unde $n(i)$ este numărul de puncte mai mici decât Y_i , punctele $Y_{1..i}$ fiind ordonate în ordine crescătoare, de la cea mai mică la cea mai mare valoare.

Graficul de mai jos reprezintă funcția de distribuție empirică pentru o distribuție normală cumulativă în cazul a 100 de numere generate aleator. Testul Kolmogorov-Smirnov (K-S) se bazează pe estimarea maximului distanțelor dintre cele două curbe.

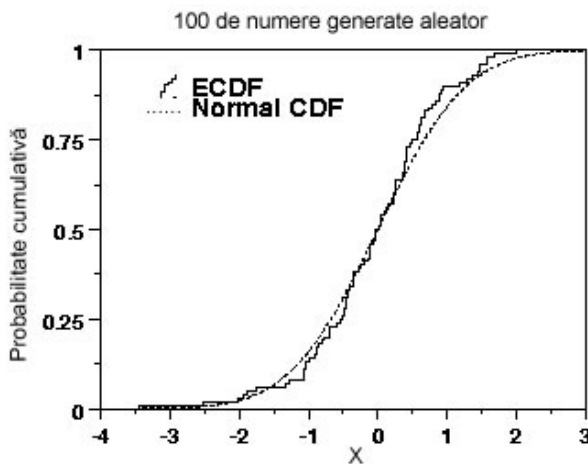
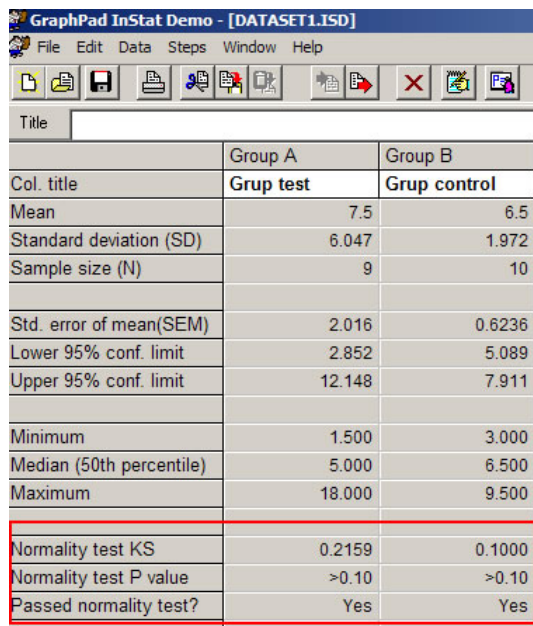


Figura IV.1. Funcția de distribuție empirică pentru o distribuție normală cumulativă în cazul a 100 de numere generate aleator

Aplicarea testului Kolmogorov-Smirnov, presupune următorii pași:

Ipoteza nulă, H_0	Datele studiate urmează distribuția specificată (în cazul nostru distribuția Gaussiană)
Ipoteza alternativă, H_1	Datele studiate NU urmează distribuția specificată (în cazul nostru distribuția Gaussiană)
Statistica testului	$D = \max_{1 \leq i \leq N} \left(F(Y_i) - \frac{i-1}{N}, \frac{i}{N} - F(Y_i) \right),$ unde F este funcția cumulativă de distribuție teoretică a distribuției testate (este necesar să fie o distribuție continuă, testul neputându-se aplica în cazul distribuțiilor “discrete” – cum este distribuția binomială sau distribuția Poisson), tipul distribuției testate trebuind să fie clar specificat.
Pragul de semnificație	α
Valori critice	Ipoteza H_0 este respinsă dacă D_{calculat} are o valoare mai mare decât o valoare critică obținută din table. Deoarece există o serie de variații ale tabelelor cu valori critice pentru acest test în datele din literatură, am preferat să nu furnizăm un astfel de tabel în anexele cărții, cu atât mai mult cu cât programele de analiză statistică ce efectuează acest test (cum este GraphPad InStat) furnizează valorile critice relevante la un anumit prag de semnificație.



	Group A	Group B
Col. title	Grup test	Grup control
Mean	7.5	6.5
Standard deviation (SD)	6.047	1.972
Sample size (N)	9	10
Std. error of mean(SEM)	2.016	0.6236
Lower 95% conf. limit	2.852	5.089
Upper 95% conf. limit	12.148	7.911
Minimum	1.500	3.000
Median (50th percentile)	5.000	6.500
Maximum	18.000	9.500
Normality test KS	0.2159	0.1000
Normality test P value	>0.10	>0.10
Passed normality test?	Yes	Yes

Figura IV.2. Testul Kolmogorov-Smirnov, aplicat în programul GraphPad InStat

IV. 3. Teste de semnificație

În general, după calcularea rezultatelor unor determinări biologice este necesar să se decidă dacă diferențele obținute de pildă între mediile sau medianele eșantioanelor analizate sunt datorate numai întâmplării (de exemplu variabilității biologice) sau sunt diferențe reale (cu alte cuvinte dacă cele două eșantioane fac parte din aceeași populație statistică sau aparțin unor populații statistice diferite).

Pentru a putea face aceasta este necesară înțelegerea corectă a unor noțiuni fundamentale din domeniul statisticii, cum ar fi **valoarea lui P**,

formularea ipotezei nule (H_0) sau existența unor dependențe între eșantioane (date pereche/nepereche).

Valoarea lui P.

În cazul unui experiment efectuat asupra a două eșantioane, o întrebare pertinentă pe care am putea să o punem este următoarea:

Dacă cele două populații statistice studiate chiar au aceeași medie (sau media nu diferă semnificativ), care este probabilitatea de a observa o aceeași diferență sau chiar una mai mare între mediile unor eșantioane de aceeași mărime într-un viitor experiment ?

Parametrul statistic ce răspunde la această întrebare este **valoarea lui P.**

“P” este de fapt o măsură a probabilității menționate mai sus, putând lua **valori între 0 și 1.**

Dacă valoarea P este mică (de exemplu $P < 0,05$) putem concluziona că este improbabil ca diferența observată între mediile celor două eșantioane să se datoreze șansei (eșantionării aleatorii). Cu alte cuvinte, putem trage concluzia că cele două populații studiate au într-adevăr medii diferite.

Ipoteza nulă (H_0).

Când specialiștii în statistică fac referire la valoarea lui P în cazul aplicării unor teste statistice, folosesc totodată și termenul de **ipoteză nulă (notată cu H_0).**

Ipoteza nulă statuează pur și simplu, aprioric, faptul că nu există nici o diferență între grupurile (eșantioanele) studiate.

Simultan se formulează și o **ipoteză alternativă (notată cu H_a sau H_1)**, ce statuează exact contrariul: că există o diferență între grupurile (eșantioanele) studiate.

Pornind de la acest fapt, putem defini valoarea lui P ca fiind probabilitatea de a observa (într-un viitor experiment) o aceeași diferență sau chiar una mai mare între două eșantioane aleatorii prelevate din cele două populații statistice studiate, decât am observat în acest moment, în cazul în care ipoteza nulă s-ar fi dovedit adevărată.

Este relativ ușor să interpretăm greșit valoarea lui P în contextul acceptării sau respingerii ipotezei nule. [instat].

Să luăm un exemplu: în urma aplicării unui test statistic pentru compararea mediilor a două eșantioane prelevate aleator din două populații statistice am obținut o valoare $P=0,03$.

În acest moment am fi tentați să tragem concluzia facilă că există 97% șanse ca diferența observată să reflecte o diferență reală între populațiile studiate, respectiv 3% șanse ca diferență să se datoreze întâmplării. O astfel de interpretarea este însă greșită !

Ceea ce putem afirma cu adevărat este faptul că în cazul unor experimente viitoare, în urma prelevării unor eșantioane aleatoare din populațiile respective vom obține o diferență mai mică decât cea constatată acum între populațiile studiate în 97 % din cazuri, respectiv o diferență mai mare în cel mult 3% din experimente.

În general în cazul testelor statistice, o valoare a lui $P>0,05$ duce la acceptarea ipotezei nule H_0 , respectiv o valoare a lui $P<0,05$

determină respingerea ipotezei nule și acceptarea ipotezei alternative H_a .

Dependența/independența eșantioanelor studiate (date pereche/nepereche – paired/unpaired data)

Anumite teste statistice despre care vom discuta în paginile următoare, țin cont de modul în care eșantioanele studiate sunt dependente sau nu unul de celălalt, existând fie versiuni speciale ale testului pentru date pereche/nepereche (cum este cazul testului t), fie testul este recomandat pentru date nepereche – testul Mann-Whitney, respectiv date pereche – testul Wilcoxon.

Una dintre principalele probleme rezidă din modul în care datele sunt grupate: avem de-a face cu un experiment cu date “pereche” sau „nepereche” (paired – unpaired în engleză) ?

Putem vorbi despre date pereche, atunci când experimentul are următorul “design”:

- Este măsurată o variabilă înainte și după o intervenție asupra aceluiași subiect
- Se efectuează un experiment de laborator de mai multe ori, de fiecare dată efectuând în paralel determinări asupra preparatului studiat, respectiv asupra controlului
- Subiecții experimentului au fost recrutați ca “perechi” pe baza unor criterii cum ar fi vârsta, rasa sau gravitatea bolii

- Sunt măsurați anumiți parametri în cazul gemenilor sau a unor perechi de genul părinte/copil etc

La modul general, putem considera datele ca fiind pereche (eșantioane dependente), dacă ne așteptăm ca o fiecare valoare dintr-un eșantion să fie corelată cu o valoare particulară din celălalt eșantion. Evident “potrivirea” NU trebuie să aibă loc pe baza variabilei studiate. De exemplu dacă vom compara tensiunea arterială între două grupuri, vom ordona eșantioanele, “potrivindu-le” pe baza vârstei pacienților, de exemplu, și nu pe baza valorilor tensiunii arteriale !

Statistică parametrică/neparametrică

În practică întâlnim două situații, pe care le vom aborda diferit:

1. *Dacă rezultatele se încadrează într-o distribuție normală și dacă cele două eșantioane nu există altă diferență semnificativă (diferențe de vârstă, sex, masă corporală etc.) în afară de tratamentul aplicat se consideră că abaterile standard ale celor două loturi nu diferă semnificativ. În acest caz se poate testa semnificația statistică a diferenței mediilor cu ajutorul mai multor teste, dintre care cel mai folosit este *testul „t” Student*.*
2. *Dacă rezultatele nu se încadrează într-o distribuție normală sau nu poate fi testată normalitatea distribuției datelor experimentale, datorită numărului mic de date colectate (existența unor eșantioane mici) este indicat să se aplice un test neparametric, cele mai des utilizate fiind *testul Wilcoxon*, respectiv *testul U Mann-Whitney*.*

IV.3.1. Teste de semnificație parametrice

Este cunoscut faptul că majoritatea efectelor farmacodinamice obținute prin administrarea substanțelor active se pot încadra în următoarele două categorii:

i. **Efecte gradate**, care variază în funcție de doză sau, altfel spus, unde relația dintre doză și efect este gradată. Aceste efecte nu se mai numesc *cantitative*.

ii. **Efecte unice (cuantale)** exprimate printr-un quantum (proces) sau efecte cu răspuns unic, de tipul „tot sau nimic”. Se mai numesc și efecte calitative.

Testele de semnificație se împart datorită acestui fapt în:

- A. Teste de semnificație pentru efecte gradate** (de exemplu testul „t”, testul „F”, testul „U”).
- B. Teste de semnificație pentru efecte cuantale** (de exemplu testul X^2).

A. Teste de semnificație pentru efecte gradate

Testele de semnificație aplicate între două medii, stabilesc valoare mediilor în funcție de posibilitățile de eroare.

Mai precis ele stabilesc care este probabilitatea ca diferența dintre două efecte (două medii) să fie reală sau, eventual, să se datoreze unei fluctuații de eșantionaj.

În fapt, se testează ipoteza nulă ce statuează faptul că nu există nici o diferență între cele două medii. Dacă această afirmație nu se verifică, înseamnă că diferența între cele două medii este semnificativă statistic, putându-se deci susține existența unui rezultat diferit de cel datorat întâmplării (cu alte cuvinte existența unui efect biologic al substanței cercetate).

Calcululele se efectuează, de obicei, la un prag de semnificație $p=0,05$, ce acoperă deci 95% din cazuri, după legile distribuției normale. Dacă se consideră necesar, se poate lucra și cu un prag mai mic de semnificație (de exemplu $p=0,01$, adică 99%), fapt care permite o “siguranță” mai bună din punct de vedere statistic.

4.1.1. Testul „F” Fischer - Snedecor

Testul propus de Snedecor, ia în considerare varianțele de eșantionaj.

Pentru a transforma varianțele în mărimi apte de a fi comparate, de exemplu în cazul efectului farmacologic a două forme farmaceutice conținând aceeași substanță activă, un instrument statistic eficient este testul „F”.

Formulele practice de calcul, obținute prin transformări algebrice, sunt următoarele:

$$S_1^2 = \frac{\sum x_1^2 - \frac{(\sum x_1)^2}{n_1}}{n_1 - 1} \quad (\text{IV.5}) \quad S_2^2 = \frac{\sum x_2^2 - \frac{(\sum x_2)^2}{n_2}}{n_2 - 1} \quad (\text{IV.6})$$

$$F = \frac{S_1^2}{S_2^2} \quad (\text{IV.7}), \text{ unde } S_1^2 > S_2^2$$

Exemplu: Vom încerca să aflăm, folosind testul „F”, dacă există o diferență semnificativă între varianța în jurul a două medii $\bar{x}_1$ și $\bar{x}_2$, provenite dintr-o determinare comparativă a două produse farmaceutice A și B. Rezultatele obținute sunt trecute în tabelul de mai jos:

Tabel IV.5. Rezultate obținute pentru două produse farmaceutice A și B

Efect A	x_1^2	Efecte B	x_2^2
6	36	15	225
4	16	4	16
3	9	10	100
7	49	10	100
6	36	5	25
4	16	11	121
		9	81
$n_1=6$		$n_2=7$	

$$\Sigma x_1 = 30$$

$$\Sigma x_2 = 64$$

$$\bar{x}_1 = 5,0$$

$$\bar{x}_2 = 9,1$$

$$\Sigma x_1^2 = 162$$

$$\Sigma x_2^2 = 668$$

$$S_1^2 = \frac{162 - (30)^2 / 6}{6 - 1} = \frac{162 - 150}{5} = 2,4$$

$$S_2^2 = \frac{668 - (64)^2 / 7}{7 - 1} = \frac{668 - 585}{6} = 13,8 \text{ deci } F = \frac{S_1^2}{S_2^2} = \frac{2,4}{13,8} = 0,174$$

Comparăm valoare obținută de noi cu cea din tabelul Anexei 5, ținând cont de gradele de libertate. Pentru $p = 0,05$ găsim valoare 4,95. Deoarece valoarea găsită de noi este mai mare decât cea teoretică, înseamnă că există o diferență statistic semnificativă între varianța în jurul celor două medii obținute experimental.

A.1.2. Testul „t” Student

A fost propusă de Gosset în cazul în care eșantioanele sunt mici. El ține cont, în calculul diferenței semnificative dintre medii, de măsura variabilității și de ponderea observațiilor în funcție de numărul acestora (grade de libertate).

Există 3 variante ale testului testul „t” implementate în diferite versiuni ale unor programe de analiză statistică:

- i. *Testul „t” pentru date “pereche” (eșantioane dependente)*
- ii. *Testul „t” pentru date “nepereche” (eșantioane independente) ce au varianță egală (homoscedatic)*
- iii. *Testul „t” pentru date “nepereche” (eșantioane independente) ce au varianță inegală (heteroscedatic)*

Pentru eșantioane mici există diferențe între testul „t” și celelalte teste de semnificație parametrice, dar începând de la $n > 15$ cifrele testului „t” se apropie de 2, ceea ce arată o coincidență cu distribuția normală.

Formulele de calcul a diferenței semnificative, în cazul testului „t” sunt următoarele:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_d} \cdot \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}} \quad (\text{IV.8}), \quad \text{unde } \bar{x}_1, \bar{x}_2 = \text{media rezultatelor}$$

eșantionului 1, respectiv 2

n_1, n_2 = numărul de animale din eșantionul 1, respectiv 2.

s_d = eroarea standard a diferenței, care se calculează conform formulei:

$$s_d = \sqrt{\frac{\sum d_1^2 + \sum d_2^2}{n_1 + n_2 - 2}} \quad (\text{IV.9})$$

unde:

$$\sum d_1^2, \sum d_2^2 = \sum (x_i - \bar{x})^2 \quad \text{în eșantioanele 1, respectiv 2.}$$

x_i - valorile individuale în eșantioanele

1 și 2.

Dacă înlocuim în formula (IV.8) valoarea erorii standard conform formulei (IV.9), obținem:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sum d_1^2 + \sum d_2^2}{n_1 + n_2 - 2}}} \cdot \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}} \quad (\text{IV.10})$$

Se consideră o diferență semnificativă, cu o probabilitate de eroare de 5% ($p=0,05$) dacă „t” calculat este superior celui din tabelul A, pentru gradele de libertate corespunzătoare.

În cazul în care cele două eșantioane sunt egale numeric ($n_1=n_2$), putem reprezenta acest număr egal de cazuri prin n ($n=n_1=n_2$) și formula (IV.10) devine:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sum d_1^2 + \sum d_2^2}{n(n-1)}}} \quad (\text{IV.11})$$

Același test se poate folosi și în cazul în care condițiile experimentale permit administrarea concomitentă a ambelor tratamente la același animal. În această situație putem admite că răspunsurile obținute la același animal sunt rezultatul exclusiv al diferențelor între acțiunea substanțelor testate, restul condițiilor fiind identice. Aceasta ne permite să scădem unul din altul cele două rezultate, obținute la același animal și să testăm semnificația diferențelor (d_i) astfel calculate (metoda poartă numele de *metoda cuplurilor*); aplică formula:

$$t = \frac{\bar{d}}{\sqrt{\frac{\sum (d_i - \bar{d})^2}{n(n-1)}}} \quad (\text{IV.12})$$

unde: d_i = valorile individuale ale diferențelor fiecărui cuplu

$$\bar{d} = \frac{\sum_{i=1}^n d_i}{n} \quad (\text{IV.13})$$

Pentru o mai bună înțelegere vom lua un exemplu. Să presupunem că în tabelul de mai jos se găsesc rezultatele obținute **la același** animal după administrarea standardului (s) și a probei (p), precum și diferența dintre aceste rezultate ($d_i = p-s$):

Tabelul IV.6. *Rezultate experimentale*

Nr.crt.	s	p	di	$d_i - \bar{d}$	$(d_i - \bar{d})^2$
1.	24	35	11	2	4
2.	20	10	-10	-19	361
3.	18	36	18	9	81
4.	45	50	5	-4	16
5.	60	74	14	5	25
6.	72	65	-7	-16	256
7.	65	70	5	-4	16
8.	54	90	36	27	729
$\sum d_i = 72$		$\sum (d_i - \bar{d})^2 = 1488$		$\bar{d}_i = 9$	

$$t = \frac{9}{\sqrt{\frac{1488}{7 \cdot 8}}} = 1.74$$

Aplicînd formula (4.2.5.) se obține un $t = 1,74$, mai mic decât valoarea „ t ” = 2,37 care este dată în tabele pentru 7 grade de libertate și o probabilitate de eroare de 5% ($p = 0,05$). Aceasta arată că efectul probei aflate în studiu nu diferă semnificativ față de standard.

B. Teste de semnificație pentru efecte cuantale

B.1.1. Testul χ^2 (Chi²)

În cazul determinărilor cuantale comparative, pentru a calcula diferența semnificativă între două activități (probe) exprimate în procente, sau pentru a stabili dacă există o anumită concordanță sau discordanță între frecvențele așteptate (teoretice) și cele observate (experimentale, empirice) sau, alte cuvinte legătura existentă sau inexistentă între o repartiție teoretică și o repartiție experimentală se folosește indicele χ^2 , propus pentru prima dată de Helmer și Pearson.

Testul χ^2 , spre deosebire de alte teste aplicate în cazul răspunsurilor biologice cuantale, ia în considerare și alți factori decât abaterea standard a procentelor, și anume numărul cazurilor, gradele de libertate, frecvențele teoretice și frecvențele experimentale.

Legătura funcțională este definită de concordanța sau neconcordanța dintre ipoteza de lucru (efecte teoretice) și rezultatele experimentale (empirice), gradul de legătură putându-se măsura prin stabilirea frecvenței asociației în comparație cu numărul cazurilor examinate, lucru care se poate exprima matematic prin raportul asociației Muster:

$$R_m = \frac{\text{Nr. cazurilor de asociere}}{\text{Nr. indivizilor examinați}} \quad (\text{IV.14})$$

Legătura funcțională dintre rezultatele teoretice (așteptate) și rezultatele experimentale ar putea fi aflată din însumarea diferențelor între

frecvențele teoretice și cele experimentale. Relația care exprimă matematic acest lucru este următoarea:

$$\sum d_i = \sum (f_{\text{teor.}} - f_{\text{exp.}}) \quad (\text{IV.15})$$

Frecvența teoretică totală poate fi egală cu frecvența empirică totală, iar diferențele pozitive se pot compensa cu cele negative, de aceea în calcul vom folosi pătratele acestor diferențe, ceea ce duce la relația:

$$\chi^2 = \frac{(f_{\text{teor.}} - f_{\text{exp.}})^2}{f_{\text{teor.}}} \quad (\text{IV.16})$$

În cazul concordanței perfecte între teorie și observație $\chi^2 = 0$.

Practic formula de calcul a lui χ^2 este cea de mai jos:

$$\chi^2 = \frac{e_e - e_t}{e_t} \quad (\text{IV.17}) \quad \text{unde} \quad e_e = \text{efectul observat experimental}$$

$$e_t = \text{efectul teoretic (așteptat)}$$

Testul χ^2 se aplică:

- frecvențelor absolute (numere, efecte de diverse categorii)
- frecvențelor relative (procentaje)

Trebuie menționat însă faptul că nu putem folosi acest test decât dacă efectele studiate au frecvențe de apariție mai mari de 10.

În determinările biologice testul χ^2 poate fi folosit ca test de semnificație în cazul răspunsurilor unice, cu ajutorul lui putându-se aprecia dacă există o diferență semnificativă între două distribuții (una teoretică și alta experimentală).

Pentru a compara o repartiție observată față de o repartiție teoretică a unui caracter calitativ cu N clase, aplicăm formula de mai sus și căutăm probabilitatea corespunzătoare la N-1.

Testul este cu atât mai semnificativ cu cât valoarea P găsită este mai mică și χ^2 calculat mai mare.

Să luăm un exemplu:

Administrând o doză egală cu DL_{50} la 40 de șoareci, se obține un efect de 30 de animale moarte și 10 animale supraviețuitoare. Am obținut deci un procent de mortalitate de 75% față de 50% cât era de așteptat. Dorim să aflăm dacă acest rezultat experimental diferă semnificativ de cel teoretic (50% mortalitate) sau se datorește doar unei fluctuații de eșantionaj. Vom aplica formula de mai sus. În cazul nostru $e_t = 20$. În urma experimentului au murit 30 de șoareci, deci $e_e = 30$. Rezultatele experimentului se trec, de obicei într-un tabel de forma:

Tabel IV.6. Rezultatele experimentului

	Morți	Supraviețuitori	Total	%
Teoretic (e_t)	20	20	40	50
Experimental (e_e)	30	10	40	75
$e_e - e_t$	10	-10	-	-

Introducând datele în formula de calcul obținem:

$$\chi^2 = \frac{(30 - 20)^2}{20} + \frac{(10 - 20)^2}{20} = \frac{10^2}{20} + \frac{(-10)^2}{20} = 5 + 5 = 10,0$$

În cadrul experimentului există două posibilități de evoluție, animale moarte sau supraviețuitoare, deci $N=2$, iar $N-1=1$. deci numărul de grade de libertate pentru care vom căuta în tabelul din Anexa 7 va fi egal cu 1. Vom constata că valoarea lui χ^2 obținută de noi corespunde, pentru $N=1$, unei valori $p=0,001$, deci rezultatul obținut în urma experimentului diferă semnificativ de cel estimat teoretic.

În cazul nostru putem interpreta rezultatul obținut ca fiind datorat unei toxicități crescute a produsului față de cea așteptată teoretic (am putea presupune, de exemplu, o descompunere a substanței active cu formarea unor produși cu toxicitate crescută: urmează să stabilim prin cercetări ulterioare care este adevărata cauză a creșterii toxicității compusului, creștere stabilită științific cu ajutorul testului χ^2).

În cazul comparației a două procentaje, formula de calcul se bazează pe coeficientul de asociație Q a lui Yule. Să o aplicăm în cazul a două produse A și B, cu câte două variabile a, respectiv b. Cifrele romane arată frecvențele absolute ale acestor variabile:

$$Q = \frac{I \cdot III - II \cdot IV}{I \cdot III + II \cdot IV} \quad (IV.18)$$

Datele pot fi grupate într-un tabel sinoptic ca cel de mai jos:

Tabel IV.7. *Tabel sinoptic*

	a	b
A	I	II
B	IV	III

Dacă luăm un număr M de cazuri (de exemplu un experiment ce folosește M animale), formula de mai sus poate fi scrisă:

$$\chi^2 = \frac{[(I \cdot III - II \cdot IV) - M/2]^2 \cdot M}{(I + II)(III + IV)(I + IV)(II + III)} \quad (\text{IV.19})$$

Să încercăm determinarea cu ajutorul testului χ^2 , a activității unui produs în comparație cu un produs martor, urmărindu-se supraviețuirea animalelor. Să presupunem că, în urma experimentării, am obținut rezultatele de mai jos:

Tabel IV.8. *Tabel sintetic cu rezultatele experimentului*

	<i>Supraviețuitori</i>	<i>Morți</i>	<i>Total</i>	<i>% supraviețuitori</i>
<i>Compus de cercetat</i>	I=25	II = 14	I + II = 39	64%
<i>Compus martor</i>	IV = 21	III = 22	IV + III = 43	49%
<i>Total</i>	I + IV = 46	II + III = 36	M = 82	-

Practic va trebui să determinăm dacă procentul de supraviețuitori de 64%, găsit în cazul compusului cercetat, diferă semnificativ statistic de procentul de supraviețuitori găsit în cazul compusului martor (49%), practic dacă compusul studiat este mai puțin toxic decât martorul. Introducem datele din tabel în formula lui Yule.

$$\chi^2 = \frac{[(25 \cdot 22 - 14 \cdot 21) - 82/2]^2 \cdot 82}{39 \cdot 43 \cdot 46 \cdot 36}$$

Deoarece avem din nou numai două posibilități, animale moarte sau supraviețuitoare, vom avea $N=2$, respectiv $N-1=1$ grad de libertate. Din tabelul Anexei 7 putem constata că statisticul lui CHI^2 corespunde unei valori $p=0,20$, statistic nesemnificativă. Putem afirma deci că produsul studiat nu este mai puțin toxic decât martorul, diferența dintre procentajele obținute datorându-se fluctuațiilor de eșantionaj.

IV.3.2. Teste de semnificație neparametrice. Statistica ordinii.

Dacă rezultatele experimentelor nu se încadrează într-o distribuție normală sau volumul eșantioanelor extrase din populația statistică este mic, este indicată recurgerea la un *test de semnificație neparametric*. Astfel de teste fac obiectul unei ramuri a statisticii numită și *statistica ordinii*, la care studiază sistemele de valori observate ale variabilelor aleatoare, din punctul de vedere al relațiilor de ordine. Un mare avantaj al acestor metode îl constituie, cum am mai spus, faptul că rezultatele ce se obțin nu depind de natura repartiției variabilei aleatoare studiate. Ele se numesc *neparametrice*, deoarece verificarea unei ipoteze nu este legată de parametrul unei anumite repartiții.

Testul Wilcoxon

Este unul dintre cele mai utilizate teste de semnificație neparametrice, extrem de util mai ales în cazul eșantioanelor dependente (date pereche - paired data).

Aplicarea lui la studiul a două eșantioane, pentru a verifica dacă acestea diferă semnificativ sau nu (și, în ultimă instanță dacă provin sau nu din aceeași populație statistică), presupune parcurgerea următorilor pași:

1. Se așează valorile n ($n = n_1 + n_2$) în ordine crescătoare, *făcând abstracție de eșantioanele din care provin*. Se atribuie apoi fiecărei valori un rang de ordine crescătoare începând cu 1. Dacă există valori egale, acestora li se atribuie ranguri egale cu media aritmetică a rangurilor pe care le-ar fi avut dacă aceste valori ar fi fost distincte.
2. Se formează un tabel în care se specifică, în ordine crescătoare, valorile obținute la fiecare eșantion și se reține, notând cu S , una din cele două sume.
3. Folosind tabelul (*testul Wilcoxon*, $p = 0,95$) de mai jos, se procedează astfel:

Dacă S este situat în afara intervalului din tabel, care se găsește la intersecția coloanei (n_1) și a liniei (n_2) se poate afirma că, la pragul de semnificație $\alpha=0,05$ (sau altfel spus $p = 0,05$), cele două eșantioane diferă. În caz contrar este justificat să se afirme că cele două eșantioane nu diferă semnificativ statistic.

Tabel IV.9. *Valori critice, testul l Wilcoxon ($\alpha = 0,05$)*

n_1								
		4	5	6	7	8	9	10
	4	11-26	17-33	24-42	32-52	41-63	51-75	62-88
	5	12-28	19-36	26-46	34-57	44-68	54-81	66-94

n_2	6	13-31	20-40	28-50	36-62	46-74	57-87	69-
	7	14-34	21-44	29-55	39-66	49-79	60-93	101
	8	15-37	23-47	31-59	41-71	51-85	63-99	72-
	9	16-40	24-51	33-63	43-76	54-90	66-	108
	10	17-43	26-54	35-67	45-81	56-96	105	75-
	11	18-46	27-58	37-71	47-86	59-	69-	115
	12	19-49	28-62	38-76	49-91	101	111	79-
	13	20-52	30-65	40-80	52-95	62-	72-	121
	14	21-55	31-69	42-84	54-	106	117	82-
					100	64-	75-	128
						112	123	86-
						67-	78-	134
						117	129	89-
							81-	141
							135	92-

Exemplu. Se ia un lot martor format din $n_1=10$ șoareci și un lot tratat, format din $n_2=9$ șoareci. Ca analgezic se folosește metamizol sodic (5mg/kg.corp), iar ca stimul chimic se folosește acid acetic 0,6% (1 ml / 10 g masă corporală). Se înregistrează numărul de contorsiuni, rezultatele fiind trecute în tabelul de mai jos:

Tabelul IV.10. Rezultatele experimentale. Testul Wilcoxon

<i>Lot martor</i>	<i>Lot tratat</i>	<i>Ranguri lot tratată</i>	<i>Ranguri lot martor</i>
-	20	1	-
-	21	2	-
22	-	-	3
27	27	4,5	4,5
-	29	6,5	-
-*	29	6,5	-
31	31	8,5	8,5
34	34	10,5	10,5
-	35	12	-
36	-	-	13,5
36	-	-	13,5
37	-	-	15
-	47	16	-
51	-	-	17
54	-	-	18
55	-	-	19
		$S_1 = 67,5$	$S_2 = 122,5$

În tabelul de mai sus, pentru $n_1=10$ și $n_2=9$, corespunde intervalul 79-121. Deoarece $S_2=122,5$ este situat în afara acestui interval, se poate afirma că $p<0,05$, deci cele două eșantioane studiate diferă semnificativ statistic.

În concluzie, acțiunea analgezică a metamizolului s-a manifestat la doza de 5 mg/kg.corp.

Ca o observație trebuie menționat că valorile fracționare apărute în coloanele rangurilor (de exemplu 4,5; 6,5; 8,5; 10,5; 13,5), rezultă din media rangurilor pe care le-ar fi avut valorile în primele două coloane, dacă ele ar fi fost distincte (de exemplu $4,5 = (4+5)/2$).

Aplicarea testului Wilcoxon este utilă în special în cazul existenței a două eșantioane dependente (date pereche), în contextul în care s-a demonstrat faptul că distribuția de frecvență nu este Gaussiană.

Testul Mann-Whitney U

Testul Mann-Whitney U (MWU) este cel mai utilizat test de semnificație neparametric pentru compararea (medianelor) a două *eșantioane independente* ce conțin măsurători ale unor date experimentale, ranguri sau scoruri. De pildă, în cazul existenței unor ranguri sau scoruri, testul MWU verifică dacă scorurile respective sunt aleator distribuite între cele două eșantioane.

Modul de calcul în cazul testului Mann-Whitney U.

Să considerăm două eșantioane independente A și B, fiecare având mărimea eșantionului $n_1, n_2=8$, pentru care studiul acordă o serie de scoruri (ranguri) de la 1 la 14 (variabile ordinale).

Tabel IV.11. *Rezultatele experimentale*

Nr.crt.	Scor pentru A	Scor pentru B
1	1	1
2	3	2
3	5	4

4	5	7
5	6	9
6	8	11
7	9	12
8	10	14

În tabelul de mai jos avem deja listați subiecții din ambele eșantioane împreună, în ordinea scorului acordat:

Tabel IV.12. *Rezultate experimentale, subiecții din ambele eșantioane, în ordinea scorului acordat*

Scorul acordat	1	1	2	3	4	5	5	6	7	8	9	9	10	11	12	14
Eșantionul	A	B	B	A	B	A	A	A	B	A	A	B	A	B	B	B

Vom calcula următoarele mărimi:

U_a = numărul de “A” (membrii ai eșantionului A) ce preced în tabelul cu scoruri primul B (membru al eșantionului B) + numărul de “A” de la începutul tabelului ce preced în tabelul cu scoruri cel de-al doilea B + ... + numărul de “A” de la începutul tabelului ce preced în tabelul cu scoruri ultimul B

Pentru datele din tabelul de mai sus,

$$U_a = 1 + 1 + 2 + 5 + 7 + 8 + 8 + 8 = 40$$

Similar putem calcula U_b . Pentru cazul luat în calcul

$$U_b = 0 + 2 + 3 + 3 + 3 + 4 + 4 + 5 = 24$$

De remarcat faptul că U_b putea fi calculat și cu formula $U_b = (n_1 * n_2)$

- U_a .

$$U_b = (8 * 8) - 40 = 24$$

Vom considera statisticul testului U ca fiind cea mai mică dintre valorile lui U_a sau U_b . În cazul nostru deci, $U = 24$. În tabelul cu probabilități

pentru testul Mann-Whitney U vom găsi pentru $n_1 = 8$, $n_2 = 8$, respectiv $U=24$ un P de 0,494, considerat ne semnificativ statistic.

Evident, majoritatea programelor de analiză statistică ne vor furniza direct valoarea lui U, respectiv valoarea lui P.

Capitolul V

V. Analiza dispersională (analiza de varianță) - ANOVA

[12, 13]

A. Baze teoretice

După cum se știe procesele biologice se pot afla, la un moment dat, sub influența mai multor factori, cu acțiune concomitentă. Pentru a pune în evidență în ce măsură unul sau mai mulți factori (sau chiar o combinație a acestora) influențează în mod esențial asupra unei caracteristici rezultative se utilizează *analiza dispersională*.

Analiza dispersională, cunoscută și sub denumirea de *analiză de varianță* (**Anova**), a fost introdusă de statisticianul R.A. Fisher. Prin această metodă se verifică măsura în care valorile reale ale unei caracteristici se abat de la valorile teoretice, calculate, de regulă, sub forma unor mărimi medii sau ecuații de regresie, precum și măsura în care aceste variații sunt dependente sau nu de factorul de grupare.

Pe baza interpretării logice a variației celor două sau mai multe variabile luate în studiu se pot stabili relații de tipul cauză – efect. Uneori prin analiza dispersională trebuie să fie verificată dependența variabilei rezultative (z) de factorul (factorii) de grupare, ea putând fi considerată, în acest caz, ca o metodă auxiliară, utilizată înainte și după aplicarea metodelor corelației și regresiei statistice. Dacă, însă, trebuie verificată independența variabilei rezultative de o variabilă de sistematizare a datelor, atunci analiza

dispersională poate fi considerată ca o *metodă independentă*, ce duce la concluzii de sine stătătoare.

Analiza dispersională are la bază *metoda grupării*. Prin aceasta se separă influența asupra caracteristicii rezultative a factorilor înregistrați ca *esențiali* (determinanți) de influența factorilor *întâmplători* (accidentali).

În funcție de numărul factorilor (unu, doi sau mai mulți) care influențează asupra variației caracteristicii rezultative, avem *modele de analiză dispersională unifactorială, bifactorială sau multifactorială*.

Modelul de analiză dispersională are la bază *ipoteza* că mediile condiționate de factorul de grupare $\bar{y}_i$, reprezintă *valorile tipice care se formează la nivelul fiecărei grupe*, în timp ce media generală $\bar{y}$ este *valoarea tipică pentru întreaga colectivitate statistică*. Măsura în care valorile individuale se abat de la aceste valori tipice reprezintă rezultatul modului de asociere a factorilor care determină variația caracteristicii y .

Se știe că dispersia teoretică (generală) σ_0^2 se poate estima cu ajutorul funcției de selecție:

$$\frac{1}{n-1} \sum (y_{ij} - \bar{y})^2 = S^2 \quad (\text{V.1.})$$

s^2 fiind, în acest caz, un estimator nedeplasat al dispersiei teoretice

Ideea de bază a analizei dispersionale constă în împărțirea acestei sume de pătrate într-un anumit număr de componente, fiecare componentă corespunzând unei surse reale sau ipotetice de variație a mediilor.

Ipoteza nulă (ipoteza de zero), pe care urmează să o testăm în cadrul analizei dispersionale, este legată de egalitatea mediilor:

$$H_0: m_1 = m_2 \dots = m_i \dots = m_r$$

Cu alternativa: H_I : cel puțin două medii diferă între ele.

Mediile teoretice m_i se estimează cu ajutorul mediilor de grupă empirice sau de selecție simbolizate $\bar{y}_i$, adică:

$$H_0 : \bar{y}_1 = \bar{y}_2 = \dots = \bar{y}_i = \dots = \bar{y}_r$$

Notă. Testul sau criteriul egalității celor r medii sau selecții are la bază presupunerea că dispersiile de selecție $s_1^2, s_2^2, \dots, s_r^2$, sunt omogene, adică sunt *estimații ale uneia și aceleași dispersii generale*. De aceea, ori de câte ori există vreun dubiu în legătură cu omogenitatea celor r dispersii, se trece la verificarea egalității lor folosind de pildă testul χ^2 .



B.Considerații practice

Dacă până acum abordarea teoretică a analizei dispersionale poate părea dificil de înțeles, în cele ce urmează vom încerca abordarea acesteia de pe baze practice. Cu alte cuvinte vom vedea unde și când aplicăm analiza dispersională monofactorială sau bifactorială.

După cum se cunoaște, răspunsul biologic obținut în urma unui experiment poate fi influențat de mai mulți parametri care acționează simultan (doza administrată, administrarea simultană a substanței active și a unor antagoniști, modificarea concentrației substanței simulante în cazul testului stimulului chimic etc), fiecare din acești parametri având, însă, o influență specifică asupra rezultatului urmărit. Scopul analizei dispersionale este separarea și testarea efectelor cauzate de variația parametrilor

respectivi și eliminarea din câmpul de observații a parametrilor a căror variație nu este semnificativă pentru rezultatul urmărit.

Principiul matematic al analizei dispersionale se bazează pe *gruparea datelor* observate după unul sau mai multe criterii și *scoaterea în evidență a efectelor obținute* în funcție de influența particulară a criteriilor după care au fost grupate observațiile.

Efectele odată identificate, *testarea* are loc prin *compararea dispersiilor cauzate de factorii variabili, cu dispersia cauzată de factorii întâmplători* care acționează asupra procesului studiat.

Numărul de criterii după care se grupează datele depinde de numărul parametrilor cuprinși în analiză. Pentru a fi mai expliciti să luăm următorul exemplu:

Să presupunem că avem în studiu șase substanțe cu efect analgezic cărora trebuie să le demonstrăm acest efect folosind testul plăcii încălzite. Practic vom determina timpul de reacție medicamentos – TRM – (timpul de latență al reacției nociceptive la animalul tratat – lingerea labei). Considerând cele șase substanțe medicamentoase drept variabile independente și timpul de latență a reacției nociceptive drept variabilă dependentă, analiza dispersională ne permite testarea influenței fiecărei substanțe active studiate asupra timpului de latență a reacției nociceptive. Metoda de analiză dispersională cu un singur parametru variabil se numește *analiză dispersională monofactorială*.

Există însă posibilitatea de a dori să studiem influența celor șase substanțe cu efect analgezic asupra timpului de latență a reacției nociceptive în condițiile administrării concomitente și a unor substanțe medicamentoase

ce potențează acțiunea analgezicelor respective. În acest caz, analiza dispersională urmărește testarea influenței simultane a două variabile independente. Metoda cu ajutorul căreia putem testa influența a doi parametri variabili se numește *analiză dispersională bifactorială*. În mod asemănător există analiză dispersională cu trei factori, cu patru factori etc. (*analiză dispersională multifactorială*).

În cele urmează vom aborda analiza dispersională monofactorială și bifactorială, aceste două metode fiind suficiente pentru înțelegerea raționamentului care ne permite eliminarea din câmpul experimental a parametrilor a căror variație nu influențează semnificativ rezultatele urmărite printr-o lucrare de cercetare. (De exemplu, aceste metode ne permit să demonstrăm faptul că o anumită substanță potențează sau nu acțiunea unor analgezice, demonstrându-și astfel valoarea extarordinară în cercetarea farmacologică).



V.1. Analiza dispersională monofactorială

Cu ajutorul analizei dispersonale monofactoriale (cu un singur parametru variabil) se testează egalitatea valorilor medii care variază sub influența unei singure variabile independente. Să presupunem că trebuie testată egalitatea valorilor medii calculate dintr-un număr de n experiențe cu m ($M_1, M_2, \dots, M_m$) substanțe având efect analgezic, criteriul de etstare fiind timpul de latență a apariției reacției nociceptive – timpul scurs până în momentul în care animalul își linge laba.

Vom aranja mai întâi datele obținute sub forma unui tabel (tabelul V.I.), în care x_{ij} ($i=1, 2, \dots, m$ și $j=1, 2, \dots, n$) reprezintă timpul de latență a apariției reacției nociceptive. De exemplu x_{32} înseamnă timpul înregistrat (în secunde) în urma experienței nr. 2 cu substanța activă nr. M_3 .

Tabelul V.1. Gruparea rezultatelor pentru analiza dispersională monofactorială

Nr. Exp.	Subst. activă	M_1	M_2	M_3	...	M_m
Experimentul 1		x_{11}	x_{21}	x_{31}	...	x_{m1}
Experimentul 2		x_{12}	x_{22}	x_{32}	...	x_{m2}
Experimentul 3		x_{13}	x_{23}	x_{33}	...	x_{m3}
....			...	...	...	...
Experimentul n		x_{1n}	x_{2n}	x_{3n}	...	x_{mn}
Total		x_1	x_2	x_3		x_m

În tabelul V.I. *diferențele* observate între datele numerice înscrise în coloane se datorează faptului că s-au folosit *diferite substanțe cu efect analgezic* (diferența între coloane), iar *diferențele între rânduri* sunt cauzate de *reproductibilitatea condițiilor experimentale*. După cum știm, dispersia valorilor individuale față de media aritmetică a lor este dată de formula:

$$S^2 = \frac{\sum_{ij} x_{ij}^2 - \frac{1}{n} \left(\sum_{ij} x_{ij} \right)^2}{n - 1} \quad (\text{V.2})$$

După cum se observă din tabel, în cazul de față dispersia este cauzată de un singur parametru, și anume folosirea unei anumite substanțe active, la care se adaugă, evident, și reproductibilitatea condițiilor experimentale. Datorită proprietății sale aditive, dispersia totală se compune din dispersia cauzată de diferența între coloane plus dispersia totală în componentele sale. Pentru simplificare, vom introduce următoarele notații pentru sumele auxiliare:

1. Suma pătratelor tuturor observațiilor individuale:

$$S_1 = \sum_{ij} x_{ij}^2 \quad (V.3)$$

2. Suma pătratelor sumei coloanelor împărțită la numărul de observații pe coloane:

$$S_2 = \frac{\sum_{i=1}^m x_i^2}{n} \quad (V.4)$$

3. Pătratul sumei tuturor observațiilor împărțit la numărul total de observații:

$$S_3 = \frac{\left(\sum_y x_y \right)^2}{mn} \quad (V.5)$$

Pentru analiza dispersiilor și testarea egalității între valorile medii observate, calculele vor trebui centralizate într-un tabel de forma celui de mai jos:

Tabelul V.2. *Analiza dispersională monofactorială*

<i>Tipul variației</i>	<i>Suma pătratelor</i>	<i>Numărul gradelor de libertate</i>	<i>Estimația dispersiei</i>	<i>F</i>
Între coloane (între grupe)	$S_2 - S_3$	$m - 1$	$S_1^2 = \frac{S_2 - S_3}{m - 1}$	$\frac{S_1^2}{S_2^2}$
Între rânduri (în interiorul grupelor)	$S_1 - S_2$	$m(n - 1)$	$S_2^2 = \frac{S_1 - S_2}{m(n - 1)}$	
Total	$(S_2 - S_3) + (S_1 - S_2) = S_1 - S_3$	$mn - 1$	-	-

Testul de semnificație (verificarea ipotezei) trebuie să se refere la raportul dintre variația între grupe (sistematică) și variația în interiorul grupei (reziduală). Pentru a verifica dacă factorul de grupare este semnificativ, se folosește testul F, dat de relația:

$$F = \frac{S_1^2}{S_2^2} \quad (\text{V.6}) \quad \text{unde} \quad S_1^2 = \frac{S_2 - S_3}{m - 1} \quad (\text{V.7}) - \text{dispersia corectată între grupe (sistematică)}.$$

$$S_2^2 = \frac{S_1 - S_2}{m(n-1)} \quad (\text{V.8}) \quad - \text{dispersia corectată din interiorul grupelor}$$

(rămasă sau reziduală)

Valoarea F rezultată din calcul se compară cu valoarea F găsită în Anexa 5 pentru pragul de semnificație ales. Întâlnim următoarele situații:

1. Dacă F calculat este mai mic decât F_α găsit în tabel, la gradele de libertate folosite pentru calcularea dispersiilor S_1^2 și S_2^2 , se acceptă ipoteza de zero adică ipoteza conform căreia *parametrul variabil nu influențează asupra variabilei dependente*, diferențele observate datorându-se erorilor experimentale.

2. Dacă F calculat este mai mare decât F_α găsit în tabel, ipoteza de zero se respinge și se trage concluzia că *parametrul variabil are influență asupra rezultatului urmărit*, diferențele observate între medii fiind reale.

Exemplu. Fiind date 6 substanțe ($M_1, M_2, \dots, M_6$), presupuse ca având efect analgezic, s-au efectuat un număr de 8 experiențe injectându-se intraperitoneal câte unui șoarece substanța de cercetat, după care se efectuează testul plăcii încălzite, conform metodologiei descrise în bibliografie [].

Intensitatea efectului se exprimă prin prelungirea timpului de reacție, adică diferența între TRM-TRI, exprimată în secunde (pentru semnificația acestor timpi, vezi testul plăcii încălzite).

Tabelul V.3. Rezultate experimentale

Nr. Exp.	Subst. activă	M_1	M_2	M_3	M_4	M_5	M_6
Experimentul 1		25,1	22,8	25,5	24,5	25,5	24,7
Experimentul 2		27,0	23,8	27,9	25,2	28,7	27,1
Experimentul 3		29,6	27,1	28,8	27,7	26,2	26,0
Experimentul 4		26,6	22,7	26,9	26,9	25,7	26,2
Experimentul 5		25,2	22,8	25,4	27,1	27,2	25,7
Experimentul 6		28,3	27,4	30,0	30,6	27,9	29,2
Experimentul 7		24,7	22,2	29,6	26,4	25,6	28,0
Experimentul 8		25,1	25,1	23,5	26,6	28,5	24,0
Total		211,6	193,9	217, 6	215, 0	215,3	211,3
Media		26,5	24,1	27,2	26,9	26,9	26,4

Pentru a ne putea orienta în vederea alegerii substanței cu cel mai puternic efect analgezic va trebui să verificăm dacă între rezultatele obținute există o diferență semnificativă sau diferențele sunt cauzate de erori experimentale. Vom calcula deci sumele auxiliare:

$$S_1 = 25,1^2 + 27,0^2 + \dots + 29,2^2 + 28,0^2 + 24,4^2 = 33511,11$$

$$S_2 = \frac{211,6^2 + 193,9^2 + 217,6^2 + 215,3^2 + 211,3^2}{8} = 33368,53$$

$$S_3 = \frac{(211,6 + 193,9 + 217,6 + 215,0 + 215,3 + 211,3)^2}{48} = \frac{1264,7^2}{48} = 33322,20$$

$$S_2 - S_3 = 46,33$$

$$S_1 - S_2 = 142,58$$

$$S_1 - S_3 = 188,91$$

$$m-1=5$$

$$m(n-1)=42$$

$$mn-1=47$$

$$S_1^2 = \frac{46,33}{5} = 9,27$$

$$S_2^2 = \frac{142,58}{42} = 3,16$$

$$F = \frac{9,27}{3,16} = 2,93$$

Cu aceste date vom completa tabelul V.4. pentru analiza dispersională, în conformitate cu modelul arătat în tabelul V.2.

Tabelul V.4. Analiza dispersională a rezultatelor experimentale

<i>Sursa dispersiei</i>	<i>Suma pătratelor</i>	<i>Gradele de libertate</i>	<i>Dispersia</i>	<i>F</i>	<i>F_{0,05}</i>
<i>Substanțele cercetate (variație între grupe)</i>	46,33	5	9,27	2,93	2,44
<i>Erorile experimentale (variație în interiorul grupei)</i>	142,58	42	3,16	-	-
Total	188,91	47	-	-	-

Concluzii. Deoarece $F > F_{0,05}$ ($2,93 > 2,44$), unde $F_{0,05}$ se găsește din tabel, la pragul de semnificație $\alpha = 0,05$, $v = 5$ și $v = 42$ grade de libertate (în tabel se caută pe orizontal numărul de grade de libertate al dispersiei de la numărul, iar pe vertical numărul de grade de libertate al dispersiei de la numitor), ipoteza nulă se respinge și putem trage concluzia că cele șase substanțe diferă

semnificativ din punct de vedere al efectului analgezic. Firește cercetarea nu se oprește aici, urmând să aibă loc noi determinări experimentale, menite să stabilească acele substanțe care pot fi utilizate în terapia, datorită efectului suficient de puternic.



V.2. Analiza dispersională bifactorială

În cazul în care se urmărește influența simulată a doi parametrii asupra unui rezultat, diferențele observate între rânduri (vezi tabelul V.1) se vor considera ca fiind cauzate de variația unuia dintre parametri, iar diferența dintre coloane – de variația celui de-al doilea parametru luat în considerare. Fiecare cifră din tabel reprezintă o observație care corespunde uneia dintre combinațiile posibile ale variabilelor independente.

Având doi parametrii variabili trebuie testate două ipoteze de zero, dintre care una se referă la diferența valorilor medii între rânduri, iar a doua la diferența valorilor medii între coloane. Pentru testarea primei ipoteze trebuie calculată diferența sumei pătratelor între rânduri, iar pentru testarea celei de-a doua ipoteze trebuie calculată diferența sumei pătratelor între coloane.

Valorile obținute experimental vor trebui grupate, și în acest caz, sub forma unui tabel, de felul celui de mai jos.

Să presupunem că, în cazul exemplului folosit la analiza dispersională monofactorială, dorim, de data aceasta, să observăm efectul analgezic al diferitelor substanțe studiate, în prezența unor substanțe active ce le potențează efectul. Primal parametru variabil va fi, în acest caz, natura substanței active, cel de-al doilea parametru variabil fiind natura substanței

cu efect de potențare a analgeziei ($P_1, P_2, \dots P_n$). Deci, în acest caz dispersia totală se compune din dispersia datorată diferențelor între coloane, plus dispersia datorată diferențelor între rânduri, rămânând și o dispersie residual, datorată erorilor experimentale. Scopul nostru este să cunoaștem separat fiecare dintre aceste abateri medii pătratice.

Tabelul V.5. Gruparea observațiilor pe criteriul factorilor de influență pentru analiza dispresională bifactorială

<i>Rândul</i>	<i>Coloana</i>					<i>Total</i>
	M_1	M_2	M_3	...	M_m	
x_1	x_{11}	x_{21}	x_{31}	...	x_{m1}	x_{i1}
x_2	x_{12}	x_{22}	x_{32}	...	x_{m2}	x_{i2}
x_3	x_{13}	x_{23}	x_{33}	...	x_{m3}	x_{i3}
...		...	...	...	...	...
x_n	x_{1n}	x_{2n}	x_{3n}	...	x_{mn}	x_{in}
<i>Total</i>	x_{11}	x_{12}	x_{13}		x_{im}	x_{ij}

Pentru simplificarea calculelor vom introduce notații similare celor folosite la analiza dispersională monofactorială:

Suma pătratelor tuturor observațiilor individuale:

$$S_1 = \sum_{ij} x_{ij}^2 \quad (\text{V.9})$$

Suma pătratelor coloanelor, împărțită la numărul observațiilor înscrise în coloană:

$$S_2 = \frac{\sum_{i=1}^m x_i^2}{n} \quad (\text{V.10})$$

Suma pătratelor sumei rândurilor împărțită la numărul observațiilor înscrise pe rânduri:

$$S_3 = \frac{\sum_{j=1}^n x_j^2}{m} \quad (\text{V.11})$$

Pătratul sumei tuturor observațiilor împărțit la numărul total de observații:

$$S_3 = \frac{\left(\sum_y x_y \right)^2}{mn} \quad (\text{V.12})$$

Suma pătratelor reziduale:

$$S_r = S_1 + S_4 - S_2 - S_3 \quad (\text{V.13})$$

Pentru calcularea dispersiilor cauzate de parametrii considerați mai trebuie să cunoaștem numărul de grade de libertate pentru fiecare abatere medie pătratică parțială. Deoarece S_2 s-a obținut din suma coloanelor, va avea $m-1$ grade de libertate, iar S_3 , obținută din suma rândurilor, va avea $n-1$ grade de libertate. În sfârșit S_r , în a cărei formulă de calcul au intervenit atât diferențele între rânduri, dispersia cauzată de factorii aleatori (întâmplători) va avea $(m-1)(n-1)$ grade de libertate. Procedul de calcul al dispersiilor, precum și testarea diferențelor între valorile medii ale coloanelor și, respectiv, rândurilor, sunt prezentate în tabelul V.2.2 :

Tabelul V.6. Analiza dispersională bifactorială

<i>Tipul variației (sursa dispersiei)</i>	<i>Suma pătratelor</i>	<i>Numărul gradelor de libertate</i>	<i>Estimația dispersiei</i>	<i>F</i>
Între coloane (primul parametru variabil)	$S_2 - S_4$	$m - 1$	$S_1^2 = \frac{S_2 - S_4}{m - 1}$	$\frac{S_1^2}{S_r^2}$
Între rânduri (al doilea parametru variabil)	$S_3 - S_4$	$m(n - 1)$	$S_2^2 = \frac{S_3 - S_4}{m(n - 1)}$	$\frac{S_2^2}{S_r^2}$
Rezidual S_r	$S_1 + S_4 - S_2 - S_3$	$(m - 1)(n - 1)$	$S_r^2 = \frac{S_r}{(m - 1)(n - 1)}$	
Total	$(S_2 - S_4) + (S_3 - S_4) = S_1 - S_4$	$mn - 1$	-	-

Cu ajutorul dispersiilor calculate conform schemei de mai sus, se face testul F, care va arăta dacă diferențele observate între coloane, respectiv între rânduri, sunt reale sau sunt cauzate de factori întâmplători.

Pentru aceasta vom calcula rapoartele $F = \frac{S_1^2}{S_r^2}$, respectiv $F = \frac{S_2^2}{S_r^2}$.

Dacă valorile rezultate din calcul sunt mai mici decât cele găsite în tabelul

distribuției F , la aceleași grade de libertate pentru care s-au calculat dispersiile S_1^2 , S_2^2 și respectiv S_r^2 , vom accepta ipoteza de zero și vom concluziona că parametrul considerat nu influențează semnificativ valorile medii, diferențele fiind cauzate de factori întâmplători. În caz contrar, însă, vom respinge ipoteza de zero și vom trage concluzia că variația parametrilor studiați are o influență semnificativă asupra variabilei dependente.

Exemplu. Să luăm experimentul de la analiza dispersională monofactorială, considerând că am lucrat cu patru substanțe cu efect analgezic (M_1 , ..., M_4), în prezența unei substanțe P , ce potențează efectul analgezic al acestora, substanță administrată în trei doze diferite (D_1 , D_2 , D_3). Rezultatele experimentului (în secunde) se regăsesc în tabelul de mai jos:

Tabelul V.7. Analiza dispersională bifactorială. Rezultate experimentale

Doza administrată pentru efect de potențare	Substanța de cercetat				Total $\sum x_j$	Medi a $\bar{x}_j$
	M_1	M_2	M_3	M_4		
D_1	25	28	22	24	99	24,7
D_2	27	29	23	23	102	25,5
D_3	30	32	26	29	117	29,2
Total	$\sum x_i = 82$	89	71	76	$\sum x_{ij} = 318$	
	$\bar{x}_i = 27,3$	29,7	23,7	25,3		

Din tabel se observă că *valorile medii ale coloanelor* sunt influențate de natura *substanței analgezice*, în timp ce *variația valorilor rândurilor* este cauzată de *variația dozei* substanței cu efect de potențare a analgeziei. Pentru calcularea dispersiilor valorilor individuale în jurul acestor medii, vom calcula, mai întâi, sumele auxiliare:

$$S_1 = 25^2 + 27^2 + \dots + 23^2 + 29^2 = 8538$$

$$S_2 = \frac{82^2 + 89^2 + 71^2 + 76^2}{3} = 8487,3$$

$$S_3 = \frac{99^2 + 102^2 + 117^2}{4} = 8473,5$$

$$S_4 = \frac{318^2}{12} = 8427,0$$

$$S_r = 8583 + 8427 - 8487,3 - 8473,5 = 4,2$$

Folosind schema de calcul din tabelul V.6, vom găsi dispersia cauzată de fiecare parametru variabil în parte, inclusiv dispersia reziduală, cauzată de factori întâmplători. Toți acești parametri statistici vor servi pentru testarea egalității valorilor medii, pentru care vom calcula în prealabil diferențele:

$$S_2 - S_4 = 60,3 \text{ cu } m-1 = 3 \text{ grade de libertate}$$

$$S_3 - S_4 = 46,53 \text{ cu } m-1 = 2 \text{ grade de libertate}$$

$$S_r = 4,2 \text{ cu } (m-1)(n-1) = 6 \text{ grade de libertate}$$

De aici se obțin *dispersiile*:

$$S_1^2 = \frac{60,3}{3} = 20,1; S_2^2 = \frac{46,5}{2} = 23,3; S_r^2 = \frac{4,6}{6} = 0,7$$

și *testul F*:

$$F_{\text{col}} \frac{S_1^2}{S_r^2} = \frac{20,1}{0,7}; F_{\text{rând}} = \frac{S_2^2}{S_r^2} = \frac{23,3}{0,7} = 33,3$$

În Anexa 5 găsim pentru coloane $F_{0,05} = 4,76$, iar pentru rânduri $F_{0,05} = 5,14$.

Odată calculate toate aceste elemente, putem completa tabelul pentru analiza dispersională bifactorială:

Tabel V.8. Rezultatele analizei Anova, bifactorială

<i>Felul variației</i> (sursa dispersiei)	<i>Suma</i> <i>pătratelor</i>	<i>Numărul</i> <i>gradelor</i> <i>de</i> <i>libertate</i>	<i>Estimația</i> <i>dispersiei</i>	<i>F</i>	<i>F_{0,05}</i>
Între analgezice	60,3	3	20,1	28,7	4,76
Între dozele de s.a. cu efect de potențare	46,5	2	23,3	33,3	5,14
Reziduală	4,2	6	0,7	-	-
Total	111,0	11	-	-	-

Concluzii. Deoarece $F_{\text{col}} > F_{0,05}$ găsit în Anexa 5 pentru gradele de libertate ale dispersiilor respective, vom respinge ipoteza conform căreia valorile medii ale coloanelor sunt egale și vom trage concluzia că substanțele active studiate diferă semnificativ din punct de vedere al efectului analgezic, la pragul de semnificație ales ($\alpha = 0,05$).

Similar, deoarece $F_{ln} > F_{0,05}$ vom respinge ipoteza conform căreia mediile liniilor sunt egale și vom trage concluzia că doze diferite de substanță P, influențează semnificativ efectul analgezic al compuşilor studiați.

Odată făcute aceste constatări, vor trebui continuate experiențele, pentru a vedea care dintre substanțe are cel mai puternic efect analgezic și, respectiv, care este doza optimă de substanță P, pentru potențarea efectului analgezic.

Mai trebuie studiați faptul că dispersia mică datorată factorilor întâmplători indică faptul că experiențele s-au desfășurat în condiții satisfăcătoare, reproductibile.

Până acum, la analiza dispersiilor cu doi parametri valabili am presupus că efectul acestora este aditiv. Dacă această condiție nu este satisfăcută, analiza dispersională se complică, deoarece eventuala interacțiune între parametrii variabili se manifestă prin creșterea dispersiei reziduale, deoarece aceasta va cuprinde în ea și dispersia cauzată de efectul de interacțiune a parametrilor luați în calcul.

Capitolul VI

VI. Corelație și regresie

[11-13, 15, 17]

Termenul **corelație** este folosit pentru a sublinia existența unei anumite forme de asociere între două variabile studiate. De exemplu, în domeniul medical putem spune că am observat o “corelație” între zilele cu ceață și declanșarea crizelor de astm.

Pe de altă parte în domeniul biostatisticii, termenul de corelație este folosit pentru a reliefa existența unei asocieri între două variabile cantitative. În mod obișnuit, suntem tentați să presupunem că această asociere este “lineară”, în sensul că una dintre variabile (să o notăm cu y) crește sau descrește într-o anumită măsură, “proporțional” cu creșterea sau descreșterea celeilalte variabile studiate (notată cu x).

Variabila “ y ” va fi considerată “**variabila dependentă**”, ce prezintă un anumit grad de asociere față de variabila “ x ”, “**variabila independentă**”.

În astfel de circumstanțe este adesea folosit termenul de **regresie** (liniară), termen ce implică estimarea celei mai potrivite linii drepte care să reliefeze asocierea, așa cum veți vedea în următoarele pagini.

Coeficientul de corelație

Gradul de asociere al variabilelor menționate anterior este măsurat cu ajutorul coeficientului de corelație, propus de Pearson și Bravais și care este o măsură a asocierii “liniare” a celor două variabile. Dacă însă de asocierea dintre variabile nu este liniară ci poate fi exprimată doar cu ajutorul unor curbe, aceasta înseamnă că sunt necesare alte măsurători ale corelației, folosind metode mai complexe, ce depășesc scopul acestei cărți.

Coeficientul de corelație este o măsură a asocierii între două variabile (variabila independentă și cea dependentă) ce poate lua valori cuprinse între $-1 \dots 0 \dots +1$.

Coeficientul de corelație „r” este un număr calculat direct din datele observate și poate varia între -1 și $+1$. Formulele de calcul ale coeficientului de corelație „r” diferă ușor, în funcție de notațiile folosite de diverși autori.

Dacă x_i sunt valorile măsurate ale variabilei X (variabila independentă) și y_i sunt valorile măsurate ale variabilei Y (variabila dependentă), atunci coeficientul de corelație se calculează astfel:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{[\sum (x - \bar{x})^2] [\sum (y - \bar{y})^2]}} \quad (\text{VI.1})$$

Ce se mai poate scrie și

$$r = \frac{\sum xy - n\bar{x}\bar{y}}{(n-1)SD(x)SD(y)} \quad (\text{VI.2})$$

unde n = numărul perechilor de date, iar $SD(x)$, $SD(y)$ – abaterile standard în cazul celor două variabile.

Putem întâlni următoarele situații:

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 și numărul de fire de păr din 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 și 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 și 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 și +1 sau între -1 și 0 , atunci valoarea lui r ne dă tăria dependenței celor două variabile.

Aceste situații sunt prezentate în figura de mai jos:

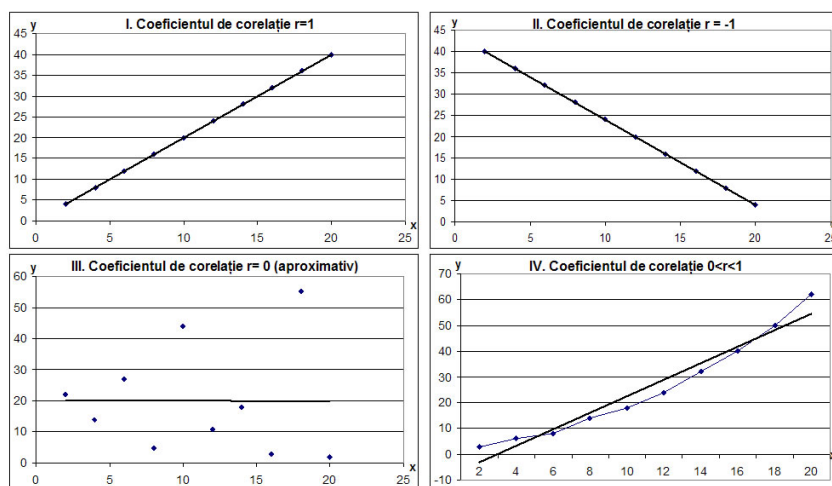


Figura VI.1. Valoarea coeficientului de corelație și semnificația lui

Dacă dorim să realizăm neapărat o clasificare a intensității asocierii (corelației) între variabila independentă și cea dependentă, putem considera, în valori absolute, următoarele intervale

$|0 < r < 0,19|$ - asociere foarte slabă

$|0,20 < r < 0,39|$ - asociere slabă

$|0,40 < r < 0,59|$ - asociere moderată

$|0,60 < r < 0,79|$ - asociere puternică

$|0,80 < r < 1|$ - asociere foarte puternică

Trebuie ținut însă seama de faptul că aceste limite de mai sus sunt oarecum arbitrare, astfel că trebuie să ținem seama și de contextul în care am desfășurat experimentele, respectiv în care am făcut măsurătorile.

Reprezentarea grafică a datelor în cazul analizei corelației și regresiei

În momentul în care un cercetător a colectat două serii de observații (măsurători) și dorește să vadă dacă există o asocierie între ele, primul lucru care trebuie făcut este reprezentarea lor grafică, sub forma unei așa-numite “diagramme de împrăștiere” (scatter diagram - diagramă de împrăștiere a rezultatelor, într-o traducere aproximativă a termenului).

Majoritatea programelor de calcul tabelar (cum este MS Excel) oferă posibilitatea realizării unei astfel de diagrame, similară celei din figura de mai jos:

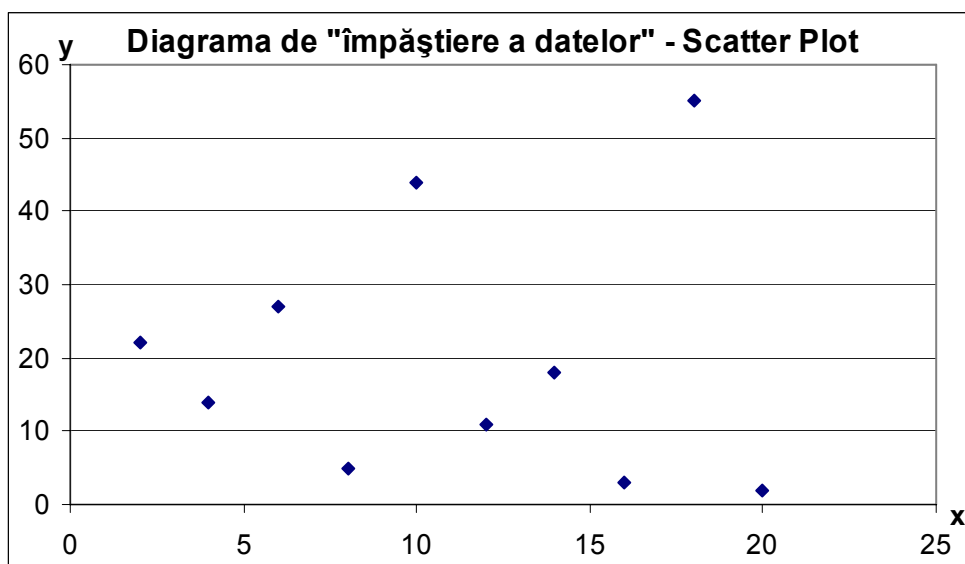


Figura VI.2. Reprezentarea datelor sub formă de diagramă (scatter plot)

Acest tip de diagramă folosește cele două axe de coordonate pentru a reprezenta cele două seturi de măsurători: pe **axa X** se află **măsurătorile legate de variabila independentă**, iar pe **axa Y** măsurătorile efectuate în cazul **variabilei dependente**.

De exemplu, dacă un set de observații sunt rezultatul unor măsurători experimentale, iar cealaltă variabilă ar fi timpul la care s-a făcut măsurarea, în acest contt timpul ar fi variabila independentă (reprezentată pe axa X), iar rezultatele măsurătorilo ar fi variabila dependentă (reprezentată pe axa Y).

Notă: *alegerea variabilelor dependente și independente trebuie făcută cu precauție, deoarece putem să greșim ușor datorită unor factori de confuzie (de exemplu o a treia variabilă care le poate influența pe amândouă). Astfel, este potrivit să presupunem că înălțimea unui lot de copii (variabila dependentă) este corelată pozitiv ($r > 0$) cu vârsta acestora (variabila independentă). Pe de altă parte, am putea constata existența unei corelații negative ($r < 0$) între numărul de cazuri de infarct miocardic (variabila “dependentă”) și consumul de înghețată (variabila “independentă”), când, de fapt, ambele variabile sunt influențate de o a treia, temperatura mediului înconjurător, fără a avea o legătură directă una cu cealaltă. Numărul de cazuri de infarct miocardic este corelat negativ, iar consumul de înghețată corelat pozitiv cu creșterea temperaturii mediului înconjurător.*

Exemplu de calcul:

Să presupunem că am măsurat în cazul a 15 copii înălțime acestora (în cm – variabila independentă), respectiv spațiul pulmonar anatomic mort (în ml – variabila dependentă). “Spațiul pulmonar anatomic mort” - pulmonary anatomical dead space, poate fi definit ca fiind volumul traectului respirator (nas, gură, trahee etc) până la nivelul alveolelor pulmonare, reprezentând acea fracțiune de aer inspirat ce nu este disponibilă schimbului de gaze la nivelul capilareor din plămân.

Rezultatele măsurătorilor se găsesc în tabelul următor:

Tabelul VI.1. Rezultatele măsurătorilor

Nr.crt. subiect	Înălțimea (cm) – variabila independentă	Spațiul pulmonar mort anatomic – variabila dependentă
1	110	44
2	116	31
3	124	43
4	129	45
5	131	56
6	138	79
7	142	57
8	150	56
9	153	58
10	155	92
11	156	78
12	159	64
13	164	88
14	168	112
15	174	101
Statistică descriptivă (n=15)	$\bar{x} = 144,60$ $SD_x = 19,37$	$\bar{y} = 66,93$ $SD_y = 23,65$

Reprezentarea grafică a datelor într-o diagramă de tip “scatter plot” este prezentată în figura de mai jos:

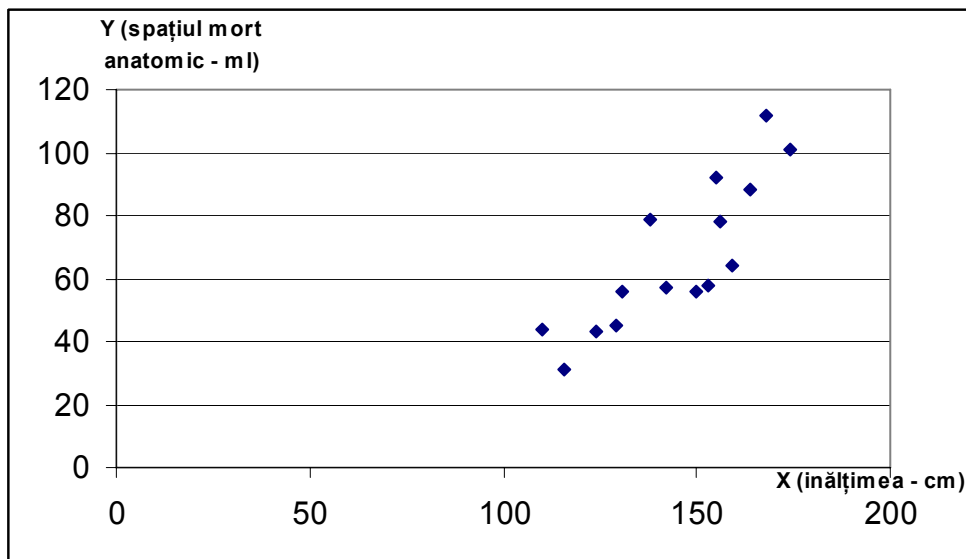


Figura VI.3. Reprezentarea grafică a datelor experimentului

Următorul pas este analiza graficului pentru a decide dacă împrăștierea datelor sugerează existența unei corelații.

Calculăm apoi coeficientul de corelație, conform celei de-a doua formule (VI.2).

Obținem astfel

$$r = \frac{150605 - (15 * 144,60 * 66,93)}{14 * 19,37 * 23,65} = \frac{5426,6}{6412,06} = 0,846$$

Este evident faptul că în acest moment dispunem de o multitudine de software ce calculează automat acest coeficient de corelație, precum și intervalul de încredere (95%) asociat acestui coeficient.

Pentru cazul nostru acest, utilizarea programului GraphPad InStat a returnat următoarele valori:

Number of points: 15

Correlation coefficient (r) = 0,8463

Confidence interval (95%): 0,5896 to 0,9477

Coeficientul de corelație pare să indice o corelație pozitivă puternică între mărimea spațiului mort anatomic și înălțimea copiilor.

Dar în interpretarea corelației este important să ne amintim că existența unei corelații între două variabile nu implică în mod necesar cauzalitatea, aceasta se poate datora unor cauze comune. Prin urmare trebuie avut grijă la interpretarea acestor coeficienți de corelație.



Coeficientul de determinare

O parte a variațiilor valorilor măsurate în cazul variabile dependente (exprimate cu ajutorul varianței, mărime calculată în cadrul analizei statistice descriptive) se pot datora într-adevăr existenței unei (co)relații cu variabila independentă, pe când o altă parte se datorează unor cauze nedeterminate (adesea aleatorii).

De aceea avem nevoie de o mărime care să cuantifice cât din această varianță a variabilei dependente se datorează influenței variabilei independente.

Această mărime se numește **coeficient de determinare** și este egal cu r^2 .

Pentru exemplul studiat anterior, $r^2 = 0,716$, astfel că putem afirma faptul că aproximativ 72% din variația existentă între volumul spațiului mort anatomic la lotul de copii studiat se datorează variațiilor înălțimii acestora.

Practic **coeficientul de determinare r^2** este extrem de util deoarece este o măsură a procentului variației ce poate fi “explicată” din totalul variației observate

Coeficientul de determinare poate avea valori cuprinse între 0 și 1 ($0 < r^2 < 1$).



Ecuția de regresie

Așa cum am văzut, coeficientul de corelație descrie intensitatea (tăria) asocierii între două variabile. Astfel, dacă două variabile sunt corelate, aceasta înseamnă că o modificare de o anumită mărime a valorii variabile independente va determina o modificare și în valoarea înregistrată la măsurarea celeilalte variabile.

Pentru exemplul de mai sus, putem spune că o valoare mai mare a înălțimii copiilor este asociată cu o creștere, de o anumită factură, a spațiului mort anatomic.

Dacă notăm cu Y variabila dependentă și cu X variabila independentă, putem afirma în consecință că relația poate fi descrisă ca o **regresie a lui Y în funcție de X** .

Această relație poate fi reprezentată de o ecuație numită **ecuație de regresie**.

În acest context termenul de **regresie** semnifică faptul că o anumită valoare a variabilei Y este o “funcție” de X , cu alte cuvinte se modifică odată cu modificarea valorii lui X , conform unei anumite ecuații mai mult sau mai puțin complexe.

Cea mai simplă astfel de ecuație este ecuația dreptei ($y = \beta x + \alpha$), iar regresia care folosește această ecuație poartă numele de **regresie liniară**.

Ecuatia de regresie ne arată cât de mult se schimbă valoarea variabilei Y în raport cu o anumită schimbare a variabile X și poate fi folosită pentru a trasa o așa-numită **linie de regresie**, în interiorul unei *diagramme scatter-plot*, iar cel mai simplu caz este cazul în care această linie este o **linie dreaptă**, caz în care se folosește termenul de **regresie liniară**.

Direcția de “înclinare” a acestei linii de regresie depinde de faptul că avem de-a face cu o corelație pozitivă sau negativă. Astfel dacă cele două seturi de observații (x și y) cresc împreună (corelație pozitivă), linia de regresie va fi ascendentă de la stânga spre dreapta. Dacă valorile variabilei X cresc, iar valorile corespunzătoare ale variabilei Y descresc, înclinarea liniei de regresie va fi descendentă de la stânga spre dreapta.

Din nefericire, în cazul regresiei liniare, de vreme ce avem de-a face cu o *dreaptă de regresie*, este foarte probabil ca ea să treacă prin relativ puține puncte reprezentate de noi în diagramă.

Fie ecuația dreptei de forma:

$$y = \beta x + \alpha \quad (\text{VI.3})$$

În momentul în care cunoaștem ecuația de regresie, pentru a putea trasa corect dreapta de regresie trebuie să ținem cont de cei doi coeficienți ai ecuației dreptei pentru a o putea trasa corect.

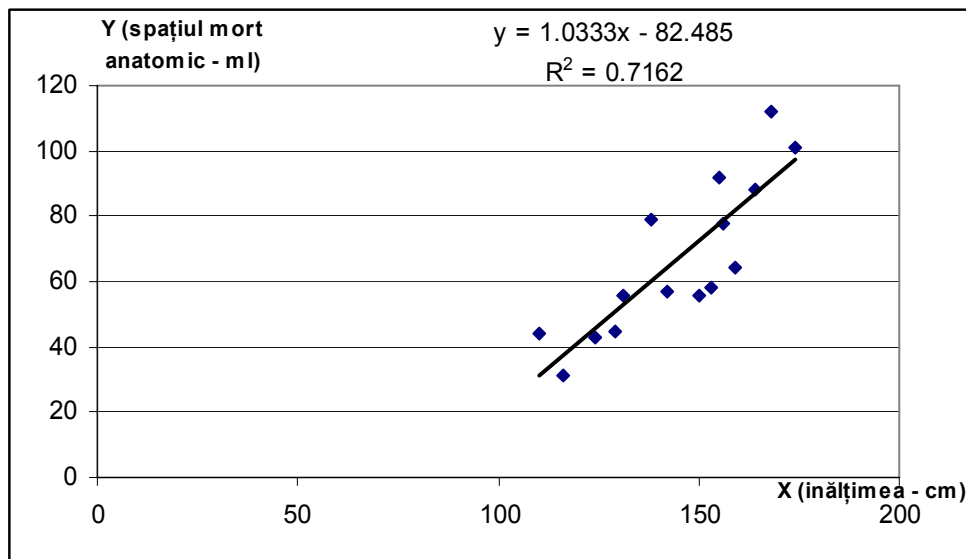


Figura VI.4. Dreapta de regresie, ecuația ei și coeficientul de determinare

Primul este *interceptul*, adică punctul în care dreapta de regresie va intersecta axa 0Y și este dat de valoarea lui α . Pentru exemplul de mai sus, dreapta de regresie ar intercepta axa 0Y în dreptul valorii $Y = -82,485$.

Cel de-al doilea coeficient este β , și poartă numele de *pantă* a dreptei de regresie. Acest ultim parametru mai poartă numele și de coeficient de regresie și poate fi asimilat ca fiind mărimea modificării înregistrate în cazul valorii variabilei Y în urma modificării cu o unitate a valorii variabilei X).

Semnul pantei ecuației de regresie liniară ne arată clar dacă avem de-a face cu o corelație pozitivă sau negativă între cele două variabile, X și Y.

În prezent programele de analiză statistică trasează automat dreapta de regresie, furnizând totodată și ecuația dreptei de regresie, respectiv coeficientul de determinare.

În trecut, trasarea corectă a dreptei de regresie se făcea cu ajutorul metodei celor mai mici pătrate - least squares estimate (dreapta se trasa astfel încât suma pătratelor distanțelor de la punctele reprezentate în diagramă la dreapta de regresie să fie minimă).

Ținând cont de cele afirmate mai sus, **coeficientul de determinare r^2** este extrem de util deoarece:

- este o măsură a procentului variației ce poate fi “explicată” din totalul variației observate
- este o măsură a procentului în care varianța (fluctuația) unei variabile (dependente) poate fi estimată (prezisă) din evoluția unei alte variabile (variabila independentă)
- este o măsură ce ne permite să determinăm cât de siguri putem fi în momentul în care facem “predicții” pentru un anumit model sau pentru date reprezentate într-o diagramă de tip scatter-plot
- coeficientul de determinare, ce poate lua valori cuprinse între 0 și 1 ($0 < r^2 < 1$) ne dă, în cazul regresiei liniare, o măsură a asocierii liniare dintre variabilele X și Y.
- coeficientul de determinare reprezintă procentul de date care este cel mai apropiat de dreapta de regresie. De exemplu dacă avem un coeficient de corelație $r = 0,922$ din care rezultă un coeficient de

determinare $r^2 = 0,850$, aceasta înseamnă că 85% din totalul variației lui Y poate fi explicat printr-o relație liniară între X și Y, relație descrisă de ecuația de regresie. Restul de 15% din variație va rămâne neexplicată.

- Coeficientul de determinare este, de asemenea, o măsură a gradului de exactitate (fidelitate) cu care o anumită linie de regresie reprezintă datele studiate. Astfel, dacă linia de regresie trece prin absolut toate punctele reprezentate în diagrama scatter, coeficientul de determinare va fi 1 și va putea explica întreaga variație. Cu cât linia de regresie este mai “îndepărtată” de puncte, cu atât coeficientul de va fi mai mic și un procent mai mare al variației nu va putea fi explicată.

În finalul acestui capitol, trebuie menționat faptul că existența unei corelații între două fenomene nu arată neapărat o legătură de la cauză la efect între ele.

Existența unei corelații este numai informativă, rămânând ca cercetătorul să stabilească legătura cauzală în funcție de cunoștințele sale în domeniul respectiv.

Capitolul VII

VII. Alegerea unui anumit test statistic

[16, 23]

De-a lungul capitolelor anterioare am discutat despre diverse teste statistice, respectiv despre modul lor de aplicare.

Poate unul dintre cele mai dificile lucruri este alegerea celui mai potrivit test pentru analiza datelor rezultate în urma unui anumit experiment sau studiu.

Pentru a alege cel mai potrivit test statistic pentru datele noastre, trebuie să ne punem două întrebări:

- i. *Ce fel de date am colectat?* Măsurători (variabile numerice, ranguri, scoruri sau date binomiale – bolnav/sănătos, expus/neexpus la un factor de risc ?
- ii. *Care este scopul analizei statistice pe care urmează să o efectuăm?* Trebuie să comparăm două eșantioane (dependente sau independente) între ele sau mai multe eșantioane. Există de asemenea posibilitatea de a compara un singur eșantion cu o serie de “valori ipotetice”

În tabelul de mai jos veți găsi câteva recomandări legate de alegerea unui anumit test în funcție de aceste considerente [Intuitive].

*Notă. Cu caractere normale sunt menționate testele statistice ce au fost descrise în această carte, în timp ce cu **bold** sunt testele statistice care nu au fost prezentate:*

Tabel VII.1. Alegerea unui anumit test statistic

Scopul analizei statistice	Măsurători (eșantioane prelevate dintr-o populație statistică Gaussiană)	Ranguri, scoruri sau măsurători (eșantioane prelevate dintr-o populație statistică non-Gaussiană)	Date binomiale
			(Două valori posibile)
Descrierea unui singur grup	Media, Abatere standard	Mediană	Proporții
Compararea unui singur eșantion cu o serie de valori "ipotetice"	Test <i>t</i> pentru un singur eșantion	Test Wilcoxon	Testul χ^2
Compararea a două eșantioane independente (date nepereche)	Test <i>t</i> pentru eșantioane independente	Testul Mann-Whitney U	Testul χ^2 , în special dacă eșantioanele sunt mari
			Testul Fisher pentru tabele de contingență
Compararea a două eșantioane dependente (date pereche)	Test <i>t</i> pentru eșantioane independente	Testul Wilcoxon	Testul McNemar
Compararea a trei sau mai multe eșantioane (grupuri) independente (unmatched groups)	Analiză ANOVA monofactorială	Testul Kruskal-Wallis	Testul χ^2
Compararea a trei sau mai multe eșantioane (grupuri) dependente (matched groups)	Analiză ANOVA multifactorială	Testul Friedman	Testul Cochran Q
Cuantificarea asocierii între două variabile	Coeficientul de corelație Pearson (<i>r</i>)	Coeficient de corelație Spearman	RR (Riscul relativ) OR (rata de șansă – Odds Ratio)
Predicția unei valori prin intermediul altei variabile măsurate	Regresie liniară	Regresie neparametrică	Regresie logistică simplă
	Regresie neliniară		

<i>Predicția unei valori prin intermediul mai multor variabile măsurate sau a mai multor variabile binomiale</i>	Regresie liniară multiplă	-	Regresie logistică multiplă
	sau		
	Regresie neliniară multiplă		

Capitolul VIII

VIII. Studii epidemiologice și clinice

[4, 12, 14, 15]

VIII.1. Indicatori statistici utilizați în studiile epidemiologice

Consider că este deosebit de importantă prezentarea schematică a clasificării tipurilor de indicatori statistici utilizați în epidemiologia clinică, precum și a tipurilor de studii utilizate în acest context, deoarece cunoașterea acestor indicatori și tipuri de studii permite justificarea protocolului statistic utilizat într-un anumit context.

În studiile epidemiologice sunt folosiți, în general, 4 tipuri de indicatori:

- I. Diverse tipuri de **raporturi, proporții și rate**
 1. Raporturi
 2. Proporții
 3. Rate
- II. **Indicatori ai frecvenței bolii** – utilizați pentru a da o măsură a frecvenței bolii sau a numărului de cazuri dintr-o populație
 1. **Prevalența**
 - i. Prevalența la un moment dat
 - ii. Prevalența de-a lungul unei perioade de timp
 - iii. Prevalența de-a lungul vieții
 2. **Incidența**
 - i. Incidența cumulativă
 - ii. Rata incidenței (sau densitatea incidenței)

III. Indicatori ai asocierii factori de risc (expunere)-boală

1. Riscul relativ (RR)
2. Rata de șansă (OR –Odds ratio)
3. Rata standardizată a mortalității
4. Rata proporțională a mortalității

IV. Indicatori ai diferenței:

1. Riscul atribuit/populație

Rate și proporții

1. Raportul - este folosit pentru a compara două cantități (de exemplu **numărul de femei/ numărul de bărbați** dintr-o localitate = $900/802 = 1,122$)

2. Proporția – un tip special de raport, în care numărătorul este inclus și în numitor, rezultând de fapt un procentaj (pentru exemplul anterior procentul femei în localitatea respectivă este **numărul de femei/numărul total de locuitori** (femei + bărbați) din localitatea respectivă = $900/1702 = 52,88 \%$)

3. Rata – un tip special de proporție ce include anumite specificații legate de timp; este unul dintre cei mai utilizați indicatori în epidemiologie deoarece poate exprima clar probabilitatea sau riscul unei boli sau a altor evenimente în cadrul populației studiate pentru a anumită perioadă de timp.

Formula de calcul:

$$\frac{\text{numărul de evenimente în perioada specificată de timp}}{\text{populația expusă la risc în perioada specificată}} * k \quad (\text{VIII.1})$$

unde k – folosit pentru unități de populație (cum ar fi 1 000 sau 100 000)

Exemplu: dacă într-un oraș de 120 000 de locuitori într-un an au fost diagnosticate 11 cazuri de leucemie acută (LA) vom avea o rată a îmbolnăvirilor de LA egală cu

$$\frac{11}{120000} * 100000 = 9,166 \text{ la } 100000 \text{ locuitori}$$

Indicatori ai frecvenței bolii

1. Prevalența – proporția de indivizi dintr-o anumită populație ce sunt bolnavi de o anumită boală la un anumit moment în timp (nu contează faptul că sunt bolnavi diagnosticați recent sau în trecut)

- permite estimarea probabilității sau riscului ca cineva din populația respectivă să devină bolnav la un anumit moment dat.

- dă o imagine a severității problemei respective

- este un indicator folositor pentru planificarea serviciilor de sănătate (personal, aparatură etc)

Formula de calcul:

$$\frac{\text{numărul de cazuri existente dintr - o anumită boală}}{\text{Totalul populației expuse la risc}} \text{ la un anumit moment în timp}$$

(VIII.2)

-poate fi și ea exprimată sub forma: la 1000 sau 100000 locuitori etc

Exemplu: dacă într-un oraș de 120 000 de locuitori la un moment dat există 70 de cazuri de leucemie acută (LA) vom avea o prevalență de

$$\frac{70}{120000} = 0,0005833 \text{ sau } 58,33 \text{ de cazuri la } 100000 \text{ locuitori}$$

Tipuri de prevalență:

- *Prevalența la un moment dat* – numărul de cazuri ce există într-o populație la un moment dat

- *Prevalența de-a lungul unei perioade de timp* - numărul de cazuri ce există într-o populație de-a lungul unei perioade specificate de timp

- *Prevalența de-a lungul vieții* – procentul dintr-o populație existentă ce a prezentat un istoric al unei anumite boli într-un moment al vieții

2. Incidența – reprezintă un indicator (o măsură) a cazurilor noi dintr-o anumită boală ce au apărut în populația studiată de-a lungul unei perioade de timp

- este un indicator al probabilității care există ca o persoană neafectată să dezvolte respectiva boală

Practic este folosită **incidența cumulativă – proporția de indivizi ce s-au îmbolnăvit într-o perioadă specificată de timp.**

Perioada de timp poate fi un an calendaristic, 6 luni, 3 ani, 5 ani etc.

Formula de calcul:

$$IC = \frac{\text{numarul de cazuri NOI dintr-o anumită boală de-a lungul unei perioade de timp}}{\text{Totalul populației expuse la risc}}$$

(VIII.3)

-poate fi și ea exprimată sub forma: la 1000 sau 100000 locuitori etc

Exemplu: dacă într-un oraș de 120 000 de locuitori la un moment dat există 30 de cazuri noi de leucemie acută (LA) în timp de un an, vom avea o incidență de

$$IC = \frac{30 \text{ de cazuri } \mathbf{noi} \text{ de LA}}{120000} = 0,00025 \quad \text{sau } 25 \text{ de cazuri la } 100000 \text{ locuitori}$$

3. Rata de incidență (densitatea incidenței) – un indicator mai puțin folosit, datorită costurilor relativ mari legate de urmărirea cazurilor de-a lungul unei perioade de timp.

- este o *măsură a incidenței capabilă să dea informații chiar dacă respectivele cazuri sunt urmărite pe perioade de timp inegale* (datorită decesului sau migrației unor persoane)
- în formula de calcul a indicatorului, similară celei a incidenței, numitorul este o sumă a așa numitului **persoana-timp la risc**

Indicatori ai asocierii factori de risc (expunere)-boală

- sunt calcule a căror rezultate sunt menite să dea o măsură a frecvenței bolii în funcție de existența altor factori care ar putea influența într-un anumit fel incidența bolii în sânul populației studiate

Extrem de importante în calculul acestor indicatori sunt așa numitele **tabele de contingență 2x2** (sau simplu tabele 2x2).

Acestea sunt folosite pentru:

- a ilustra frecvența bolii, respectiv a expunerii
- calculul asocierii dintre expunerea la factorul (de risc) studiat și boală

Pentru a putea fi folosite în determinarea unor indicatori epidemiologici tabelele de contingență de tip 2x2 trebuie să grupeze corect indivizii din populația studiată, ținând cont de expunere (liniile tabelului),

respectiv de starea de boală (coloanele tabelului), cum se poate vedea în exemplul ce urmează.

Tabel VIII.1. *Tabel de contingență de tip 2x2*

		Boală		Total
		Da	Nu	
Expunere	Da	<i>a</i>	<i>b</i>	$a + b$
	Nu	<i>c</i>	<i>d</i>	$c + d$
Total		$a + c$	$b + d$	$a + b + c + d$

unde:

a = numărul indivizilor din cadrul populației studiate **expuși** (care au suferit o expunere la factorul de risc studiat) și în același timp **bolnavi** (evident, cei suferinzi din cauza bolii studiate)

b = numărul indivizilor din cadrul populației studiate **expuși** dar **sănătoși**

c = numărul indivizilor **neexpuși** (care nu au suferit o expunere la factorul de risc studiat) și în același timp **bolnavi**

d = numărul indivizilor **neexpuși** și **sănătoși**

a + b = **numărul total al indivizilor** (bolnavi + sănătoși) din cadrul populației studiate **expuși** la factorul de risc

c + d = numărul total indivizilor (bolnavi + sănătoși) din cadrul populației studiate **neexpuși** la factorul de risc

a + c = numărul total al indivizilor bolnavi din cadrul populației studiate

b + d = numărul total al indivizilor sănătoși din cadrul populației studiate

a + b + c + d = Mărimea totală a populației studiate (eșantionului), practic suma tuturor celor patru celule ale tabelului

Există 2 indicatori epidemiologici foarte importanți pentru măsurarea unei asocieri expunere (factor de risc)-boală:

- **Riscul relativ** (Relative Risk - RR)
- **Rata de șansă** (Odds Ratio –OR)

Ambii indicatori sunt calculați *folosind gruparea datelor în tabele de contingență de tip 2x2*.

Riscul relativ (Relative Risk - RR)

- servește la măsurarea magnitudinii asocierii între incidența bolii studiate și factorul de risc cercetat
- reprezintă în fapt un raport dintre rata incidenței persoanelor expuse și rata incidenței în rândul celor neexpuși la factorul de risc
- folositor în cercetările legate de etiologia unor boli
- sinonime: Rata de risc (Risk ratio)

Formula de calcul:

$$RR = \frac{\text{Incidența bolii în rândul subiecților expuși la factorul de risc}}{\text{Incidența bolii în rândul subiecților neexpuși la factorul de risc}} \quad (\text{VIII.4})$$

sau ținând cont de modul de grupare a datelor într-un tabel de contingență de tip 2x2:

$$RR = \frac{a/(a+b)}{c/(c+d)} \text{ (VIII.5)}$$

Interpretarea valorilor riscului relativ:

- Riscul relativ poate avea, teoretic vorbind, valori între 0 și infinit
- $RR = 1$: Nu există nici un fel de asociere între expunere (factorul de risc) și boală; practic incidența (rata incidenței) este identică între cele două grupuri (expuși, respectiv neexpuși la factorul de risc)

$RR > 1$ = Există o asociere pozitivă între expunere (factorul de risc) și boală

$RR < 1$ = Există o asociere negativă între expunere (factorul de risc) și boală sau, în unele cazuri putem spune că expunerea la factorul respectiv are chiar un “efect protector” față de boală

Rata de șansă (Odds Ratio –OR)

- Singura modalitate de măsurarea a asocierii expunere (factor de risc)-boală disponibilă în cazul studiilor de tip Case-Control

Dă o măsură a riscului relativ (are valori apropiate de acesta) când:

- valoarea incidenței este mică ($< 5\%$ din populația totală) și
- grupul de control (martor) este reprezentativ pentru întreaga populație (inclusiv din punctul de vedere al expunerii la factorul de risc)

Trebuie definită în doi pași:

- **Șansă (ODDS)** – Raportul dintre probabilitatea ca un eveniment să se întâmple și probabilitatea ca respectivul eveniment să nu aibă loc
- **Rata de șansă (Odds Ratio –OR)** – Raportul dintre rata (șansa) expunerii în rândul celor bolnavi și rata (șansa) expunerii în rândul celor sănătoși

Formula de calcul:

- ținând cont de modul de grupare a datelor într-un tabel de contingență de tip 2x2, formula de calcul este:

$$OR = \frac{a/c}{b/d} = \frac{a * d}{b * c} \quad (\text{VIII.6})$$

Interpretarea valorilor ratei de șansă (OR):

- Rata de șansă (OR) poate avea, teoretic vorbind, valori între 0 și infinit

OR = 1 : Nu există nici un fel de asociere între expunere (factorul de risc) și boală;

OR > 1 : Expunerea reprezintă un factor de risc în cazul bolii respective (crește riscul apariției bolii)

OR < 1 : Expunerea reprezintă un factor „protector” în cazul bolii respective (scade riscul apariției bolii)

Interpretarea rezultatelor analizei statistice. Interpretarea valorii lui RR (OR) și P în cazul studiilor ce utilizează tabele de contingență de tip 2x2

Interpretarea existenței unei asocieri între factorul de risc și apariția bolii, în cazul unui studiu retrospectiv de tip cohortă, necesită o analiză atentă, ținând cont de următorii indicatori **(13, 145)** :

- 1. valoarea lui P (cel mai important parametru)**
- 2. valoarea minimă și maximă a intervalului de încredere, calculat pentru o probabilitate de 95 % (CI 95 %)**
- 3. valoarea ratei de șansă (Odds ratio-OR)**

Într-un astfel de studiu ne interesează atât valoarea lui P (cel mai important parametru cât și intervalul de încredere pentru un anumită rată de șansă (OR) sau risc relativ (RR).

Valoarea lui P ne răspunde la următoarea întrebare: Dacă, de exemplu, nu există nici o asocierie între variabilele care definesc coloanele și variabilele care definesc liniile unui tabel de contingență (practic în cazul studiului nostru între expunere și boală), care este șansa (probabilitatea) ca luând la întâmplare indivizi (așa numita *eșantionare aleatorie - random sampling*) dintr-o populație să descoperim o asocierie (mai) semnificativă decât cea din studiul în cauză (**145**)? Cu alte cuvinte, dacă într-adevăr nu există deloc nici o asociație între linii și coloane care este probabilitatea ca o eșantionare aleatorie să ducă la obținerea unei valori a OR (foarte) diferită de 1,0 precum și diferită de cea obținută în studiul în cauză ?

Pentru a explica mai bine cele afirmate mai sus, să luăm un exemplu:

Să presupunem că pentru un studiu al influenței radiațiilor ionizante asupra incidenței leucemiilor acute (LA) am obținut o valoare a **RR = 0,923** și o valoare a lui **P = 0,991**. Conform celor afirmate mai sus, aceasta valoare a lui P ne spune că printr-o eșantionare aleatorie, la modul teoretic, avem **99,1 %** șanse de a obține un RR diferit atât de valoarea obținută (0,923) de noi cât și de 1,0. De aceea putem afirma că ceea ce am obținut noi în studiu este *statistic nesemnificativ* dacă $P = 0,991$. Invers, dacă din analiza statistică ar rezulta o valoare a **RR = 4,204** și o valoare a lui **P < 0,0001**, aceasta ar însemna că avem mai puțin de 0,0001 % șanse de a obține, prin eșantionare aleatorie (alegând deci la întâmplare subiecți) un OR mult diferit atât de

valoarea obținută (4,204) de noi cât și de 1,0. De aceea putem afirma ca ceea ce am obținut noi în studiu este *statistic semnificativ* dacă **$P < 0,0001$** .

Din nefericire, termenul "*semnificativ statistic*" nu este de multe ori același lucru cu "*important din punct de vedere științific*".

Înainte de a interpreta din punct de vedere științific o anumită valoare a lui P sau un anumit interval de încredere trebuie să ne gândim atent la ordinul de mărime a ratei de risc (RR) sau ratei de șansă (OR). Cât de mare trebuie să fie această valoare a lui OR pentru a o considera importantă din punct de vedere științific (chiar dacă am obținut un rezultat considerat statistic semnificativ)? Cât de mică trebuie să fie acea valoare pentru a o putea considera "normală" și a putea trage concluzia că expunerea studiată nu poate fi considerată factor de risc în boala respectivă ?

Modul în care interpretăm rezultatele unui studiu depinde foarte mult de valoarea lui P: mică (**$P < 0,05$**) sau mare (**$P > 0,05$**).

A. Dacă valoarea lui P este mică ($P < 0,05$) = rezultat semnificativ statistic:

În acest caz este improbabil că asocierea expunere-boală observată să fie întâmplătoare. Putem deci respinge ideea că asocierea este o coincidență datorată eșantionării și putem concluziona că populația studiată are un RR sau OR diferit de 1,0. Asocierea este deci statistic semnificativă (dacă **$P < 0,05$**). Dar este acest rezultat important și din punct de vedere științific?

Intervalul de încredere (CI 95 %) este cel care ne poate da un răspuns.

Orice analiză statistică include întotdeauna, în proporție mai mică sau mai mare, o serie de erori datorate eșantionării (așa numitele effects of random sampling), astfel că valoarea *RR* sau *OR* calculată din datele studiului probabil nu este identică cu valoarea “adevărată” a *RR* sau *OR*. Dar, din nefericire, nu există posibilitatea determinării exacte a acestei valori “adevărate”. În schimb analiza statistică ne dă o măsură a acestei valori prin intermediul intervalului de încredere la 95 %. *Practic putem fi 95 % siguri că acest interval conține valoarea “adevărată” a RR sau OR.*

Pentru a putea interpreta rezultatele analizei statistice într-un context științific pertinent, trebuie să privim la ambele capete ale intervalului de încredere și să ne întrebăm dacă acestea reprezintă valori ce au importanță din punct de vedere științific sau sunt valori obișnuite, comune, fără importanță științifică.

Tabelul VIII.2 sistematizează variantele posibile:

Tabel VIII. 2. Interpretarea rezultatelor analizei statistice

Valoarea minimă a CI 95 %	Valoarea maximă CI 95 %	Interpretare
Comună (relativ apropiată de 1)	Comună (relativ apropiată de 1)	Chiar dacă <i>RR</i> sau <i>OR</i> nu este 1,0 , asocierea expunere-boală este mică și relativ lipsită de interes științific , chiar dacă este statistic semnificativă
Comună (relativ apropiată de 1)	Importantă (mult mai mare decât 1)	Deoarece CI 95 % variază între o valoare a <i>OR</i> comună din punct de vedere biologic (științific) și o valoare semnificativă, nu putem trage o concluzie certă cu privire la importanța științifică a rezultatului. <i>La modul ideal, o nouă serie de date ar trebui analizate pentru a putea avea o concluzie certă,</i>

		<i>atât din punct de vedere statistic cât și științific.</i>
Importantă (mult mai mică decât 1)	Comună (relativ apropiată de 1)	Deoarece CI 95 % variază între o valoare semnificativă și o valoare a OR comună din punct de vedere biologic (științific), nu putem trage o concluzie certă cu privire la importanța științifică a rezultatului. Apare o problemă legată însă de designul studiului nostru, expunerea pare să influențeze în sens negativ incidența bolii, putând fi considerată chiar ca un “factor de protecție” (13). <i>O nouă serie de date ar trebui analizate pentru a putea avea o concluzie certă, atât din punct de vedere statistic cât și științific, iar dacă rezultatele sunt similare trebuie verificat modul în care am conceput designul studiului și gruparea datelor în tabelele de contingență.</i>
Importantă	Importantă	De vreme ce până și valoarea minimă a CI 95% reprezintă o asociere expunere-boală suficient de mare pentru a fi considerată biologic importantă, se poate trage concluzia indubitabilă că există o asociere expunere-boală semnificativă atât din punct de vedere statistic, cât și științific.

B. Dacă valoare lui P este mare ($P > 0,05$) = rezultat ne semnificativ statistic:

Dacă valoarea lui P este mare, nu avem nici un motiv plauzibil să concluzionăm că RR sau OR diferă de 1,0. Aceasta **nu înseamnă însă că**

putem spune că RR sau OR sunt egale cu 1,0. Pur și simplu nu avem dovezi consistente că valorile RR sau OR diferă de 1,0.

Care este oare valoarea “adevărată” a riscului relativ sau a ratei de șansă ?

Cum datele noastre includ o serie de erori datorate eșantionării (effects of random sampling) valoarea RR sau OR calculată din datele studiului probabil nu este identică cu valoarea “adevărată” a RR sau OR. Neexistând posibilitatea determinării exacte a acestei valori “adevărate”, analiza statistică ne dă o măsură a acestei valori prin intermediul intervalului de încredere la 95 %. Practic putem fi 95 % siguri că acest interval conține valoarea “adevărată” a RR sau OR.

Când însă $P > 0,05$, **intervalul de încredere este mult lărgit**: el include așa numita ipoteză nulă (null hypothesis – OR sau RR = 1,0) și se întinde pe o gamă de valori care ar putea sugera **de la existența unei asocieri negative** expunere-boală (valoarea minimă a CI 95 % este RR <1,0 sau OR <1,0) până la **existența unei asocieri pozitive** expunere-boală (valoarea maximă a CI 95 % este RR >1,0 sau OR >1,0)

Pentru a putea interpreta rezultatele analizei statistice într-un context științific pertinent, trebuie să privim la ambele capete ale intervalului de încredere, situațiile ce pot fi întâlnite fiind prezentate schematic în tabelul IX:

Tabel VIII.3. Interpretarea rezultatelor analizei statistice

Valoarea minimă a CI 95 %	Valoarea maximă CI 95 %	Interpretare
Comună	Comună	Nu există o asociere semnificativă

(relativ apropiată de 1)	(relativ apropiată de 1)	statistic între expunere-boală. Cu alte cuvinte nu există o nici asociere între liniile și coloanele tabelului de contingență sau dacă există ea este cu certitudine foarte mică și lipsită de interes. Un rezultat ce poate fi considerat semnificativ și din punct de vedere științific.
Comună (relativ apropiată de 1)	Diferită de 1 (mai mare decât 1)	Nu se poate trage o concluzie foarte clară. Deși analiza statistică demonstrează inexistența unei asocieri statistice semnificative , expunerea la factorul de risc duce la apariția unui interval de încredere ce se întinde de la o valoare minimă specifică unei slabe asocieri negative expunere-boală, până la o valoare maximă specifică unei asocieri pozitive puternice expunere-boală; Cât de apropiată de 0,05 este valoarea lui P ? <i>Pentru a putea trage concluzii clare este necesară repetarea experimentului folosind mai mulți subiecți.</i>
Diferită de 1 (mai mică decât 1)	Comună (relativ apropiată de 1)	Nu se poate trage o concluzie foarte clară. Deși analiza statistică demonstrează inexistența unei asocieri statistice semnificative , expunerea la factorul de risc duce la apariția unui interval de încredere ce se întinde de la o valoare minimă specifică unei puternice asocieri negative expunere-boală, până la o valoare maximă specifică unei slabe asocieri pozitive

		expunere-boală; Cât de apropiată de 0,05 este valoarea lui P? Dacă P are o valoare relativ apropiată de 0,05 – atenție la designul studiului ! <i>Pentru a putea trage concluzii clare este necesară repetarea experimentului folosind mai mulți subiecți.</i>
--	--	--

Modul în care trebuie interpretate rezultatele analizei statistice în cazul studiilor ce utilizează tabele de contingență de tip 2x2 este cel de mai sus, clar prezentat într-un număr mare de surse bibliografice:

– **valoarea lui P** ca prim parametru, din care rezultă **semnificația statistică** a analizei

– **valorile CI 95 %** pentru **explicitarea sensului și magnitudinii asocierii**, din care rezultă **semnificația științifică a analizei**

Cu toate acestea, o serie de articole apărute în ultimii ani în prestigioase reviste internaționale (7, 93, 153) pe teme similare prezentului studiu, prezintă valorile OR sau RR fără a preciza valoarea lui P, ci numai intervalul de încredere la 95 % (CI 95 %).

Este recomandabil să luăm în considerare din punct de vedere al concluziilor certe doar valorile cu **$P < 0,05$** , dar se poate face o și o cuantificare a situațiilor în care valoarea lui P este apropiată de 0,05.

Aceasta deoarece pot fi întâlnite patru situații distincte:

I. În cazul existenței unei asocieri semnificative statistic ($P < 0,05$):

Cazul 1:

- valoarea OR (RR) > 1
- valoarea minimă a CI 95 % > 1 ; valoarea maximă a CI 95 % > 1 (de obicei mult mai mare decât 1)
- evident, $P < 0,05$

Interpretare: În acest caz putem spune că există o asociere **POZITIVĂ, SEMNIFICATIVĂ STATISTIC**, între factorul de risc și boala studiată.

Cazul 2:

- valoarea OR (RR) < 1
- valoarea minimă a CI 95 % < 1 (de obicei mult mai mic decât 1); valoarea maximă a CI 95 % < 1
- $P < 0,05$

Interpretare: În acest caz putem spune că există o asociere **NEGATIVĂ, SEMNIFICATIVĂ STATISTIC**, între factorul de risc și boala studiată. Cu alte cuvinte, în funcție de expunerea studiată, unele surse bibliografice susțin că, în acest caz, putem afirma că expunerea nu numai că nu influențează apariția bolii, ci are chiar o acțiune “protectoare”. (13)

II. În cazul **inexistenței unei asocieri semnificative statistic ($P > 0,05$)**:

Cazul 3:

- $P > 0,05$, dar valoarea lui P este mică (în aceasta lucrare am considerat acest caz pentru $P < 0,1$)
- valoarea OR (RR) > 1
- valoarea minimă a CI 95 % < 1 ; valoarea maximă a CI 95 % > 1

Interpretare: În acest caz putem spune că **poate fi observată o asociere POZITIVĂ, NESEMNICATIVĂ STATISTIC**, între factorul de risc și boala studiată.

Cazul 4:

- **$P > 0,05$, dar valoarea lui P este mică** (în aceasta lucrare am considerat acest caz pentru **$P < 0,1$**)
- **valoarea OR (RR) < 1**
- **valoarea minimă a CI 95 % < 1 ; valoarea maximă a CI 95 % > 1 (de obicei apropiată de 1)**

Interpretare: În acest caz putem spune că **poate fi observată o asociere NEGATIVĂ, NESEMNICATIVĂ STATISTIC**, între factorul de risc și boala studiată.

Există și o a cincea situație, cea în care **OR (RR) $\cong 1$** , ceea ce înseamnă între factorul de risc (expunere) și boală **nu poate fi făcută nici un fel de asociere**, factorul de risc (expunerea) nu are nici un fel de influență în cazul studiat (în această lucrare am considerat ca validă o astfel de situație pentru $0,95 < OR < 1,05$).

Practic, după cum se poate vedea, valoarea OR, respectiv CI 95 %, ne furnizează informații despre **sensul influenței factorului de risc (expunerii) asupra bolii** (în mod evident o valoare a OR egală cu 1, însemnând faptul că factorul respectiv nu influențează în nici un fel boala), în timp ce **valoarea lui P** ne dă informații despre **semnificația statistică** a acestei influențe (**$P < 0,05$ – asocierea descoperită este semnificativă statistic**).

Cele prezentate mai sus sunt sistematizate în tabelul VIII.4:

Tabel VIII.4. Interpretarea rezultatelor statistice privind asocierea factor de risc-boală

ASOCIERE factor de risc - boală STATISTIC SEMNIFICATIVĂ				
OR (RR)	Val. minimă CI 95 %	Val. maximă CI 95 %	P	Interpretare
> 1	> 1	>> 1	< 0,05	Există o asociere POZITIVĂ, SEMNIFICATIVĂ STATISTIC între factorul de risc și boala studiată
Asociere pozitivă			Statistic semnificativ	
OR (RR)	Val. minimă CI 95 %	Val. maximă CI 95 %	P	Interpretare
< 1	<<1	< 1	< 0,05	Există o asociere NEGATIVĂ, SEMNIFICATIVĂ STATISTIC între factorul de risc și boala studiată
Asociere negativă			Statistic semnificativ	
ASOCIERE factor de risc - boală STATISTIC NESEMNIFICATIVĂ				
OR (RR)	Val. minimă CI 95 %	Val. maximă CI 95 %	P	Interpretare
> 1	< 1	> 1	> 0,05	Poate fi observată o asociere POZITIVĂ, NESEMNIFICATIVĂ STATISTIC între factorul de risc și boala studiată
Asociere pozitivă			Statistic ne semnificativ	
OR (RR)	Val. minimă CI 95 %	Val. maximă CI 95 %	P	Interpretare
< 1	< 1	> 1	> 0,05	Poate fi observată o asociere NEGATIVĂ, NESEMNIFICATIVĂ STATISTIC, între factorul de risc și boala studiată
Asociere negativă			Statistic ne semnificativ	
OR (RR) ≅ 1 : între factorul de risc (expunere) și boală nu poate fi făcută nici un fel de asociere(în cazul studiului nostru am considerat această situație pentru 0,950 <OR (RR) < 1,050)				

Riscul atribuit (RA)

- numărul de cazuri din populația expusă ce ar putea fi „eliminate” dacă expunerea (la factorul de risc) ar putea fi îndepărtată. (239)

Formula de calcul

$$\text{RA} = \frac{\text{Incidența în rândul populației expuse} - \text{Incidența în rândul populației neexpuse}}{\text{(la factorul de risc)}}$$

VIII. 2. Tipuri de studii epidemiologice

Este cunoscut faptul că deosebim 2 tipuri mari de studii epidemiologice (239)

I. **Descriptive** (persoane, locuri și timp)

» Servesc la **generarea** de ipoteze de lucru

II. **Analitice** (cauzale)

» Servesc la **testarea** ipotezelor de lucru

STUDII DESCRIPTIVE

CARACTERISTICI GENERALE:

- **Sunt, în general, studii de tip corelațional**
 - » Examinează caracteristici ale întregii populații
 - Exemple: Studiarea unor relații între vânzările de țigări pe teritoriul unei țări și mortalitatea datorată BPOC în acea țară
 - » Reprezintă *primul pas* în studierea unei *relații de tip factor de risc-boală*

» **Avantaje**

- Metodă rapidă și ieftină, poate fi utilizată ca un prim pas în efectuarea unui studiu mai complex

» **Limitări**

- Nu poate furniza informații certe cu privire la existența sau nonexistența unor relații de tip cauză-efect
- Nu poate oferi cercetătorului o modalitate de control detaliat al potențialelor surse de eroare

SUBTIPURI ALE STUDIILOR DESCRIPTIVE:

A. Descriere de caz sau serii de cazuri (Case Reports and Case Series)

- » Descriu patologia unui singur bolnav sau a unui grup de bolnavi
- » Cea mai utilizată formă de studiu publicată în revistele medicale
- » **Avantaje**
 - Poate duce la formularea de noi ipoteze de lucru
 - Constituie un element major de legătură între medicina clinică și epidemiologie
- » **Limitări**
 - Nu poate fi folosit pentru verificarea unor ipoteze de lucru

B. Studiu descriptiv de tip urmărire a prevalenței (Cross-Sectional or prevalence Survey)

- » Expunerea la factorii de risc și starea de boală sunt simultan cuantificate în cadrul unei populații
- » Furnizează informații legate de frecvența și caracteristicile unei anumite boli
- » **Avantaje**
 - Foarte folositor în domeniul sănătății publice
 - Poate furniza informații utile legate de prevalența (incidența) unei boli în anumite grupuri de populație (expuse la anumiți factori de risc specifici profesiei de exemplu)
- » **Limitări**
 - De cele mai multe ori nu poate să dea informații legate de faptul dacă expunerea la factorii de risc a precedat boala sau a apărut în timpul ei
 - Datorită folosirii prevalenței ca indicator, nu se face o deosebire între cazurile noi și cele deja existente la momentul începerii studiului
 - Nu poate fi utilizat pentru studierea factorilor etiologici

C. Studiu descriptiv de tip ecologic

- » Studii care în care unitățile de analiza sunt mai degrabă populații întregi sau grupe de populații decât subiecți individuali (cazuri)
- » **Avantaje**
 - Este extrem de util în generarea de ipoteze de lucru
- » **Limitări**
 - Fiecare individ aparținând unei populații este caracterizat de media calculată pentru întreaga populație studiată
 - Pot apărea erori datorate faptului că o asociere observată la nivelul întregii populații studiate nu reprezintă în mod obligatoriu asocierea existentă la nivel individual
 - Aceste două limitări pot determina apariția unor erori ce constituie așa numitul “Ecologic Fallacy”

ASPECTE CE TREBUIE LUATE ÎN CONSIDERARE ÎN CAZUL EMITERII DE IPOTEZE ȘI A STABILIRII UNUI ANUMIT DESIGN PENTRU STUDIU:

● **Ce boală va fi studiată și cum?**

- » **Bine fundamentată teoretic:** boală, respectiv studiu bine fundamentat teoretic, pe bază unor surse bibliografice numeroase
- » **Studiată empiric,** pe baza criteriilor de diagnostic

- **Ce tip de expunere va fi studiată ?**
 - » **Bine fundamentată teoretic:** există numeroase surse bibliografice și studii similare
 - » **Studiată empiric,** pe baza criteriilor și metodelor de examinare alese de către cercetător
- **Care este timpul de inducție ?** (Cu alte cuvinte care este intervalul mediu de timp dintre expunerea la factorul de risc și efect – apariția bolii ?)
- **Ce alți factori pot influența această asociere ?**
 - » **Factori bine studiați (fundamentați teoretic)**
 - » **Empiric,** pe baza criteriilor și metodelor de examinare alese de către cercetător

STUDII ANALITICE

CARACTERISTICI GENERALE:

- **Permit efectuarea unei comparații explicite între expunere (la factorii de risc) și boală**
- **Utilizează gruparea populației studiate în tabele de contingență realizate în așa fel încât să se poată observa clar (și determina statistic mai apoi) dacă riscul este diferit între populația expusă și cea neexpusă**
- **Sunt singurele studii ce pot fi utilizate pentru TESTAREA DE IPOTEZE**

SUBTIPURI ALE STUDIILOR ANALITICE:

Studiilor analitice sunt grupate în două subtipuri mari:

- Studii analitice de tip **observațional** – urmează cursul natural al evenimentelor
- Studii analitice de tip **intervențional** – cercetătorul determină expunerea (de exemplu un anumit tratament) și urmărește subiecții (de exemplu așa numitele studii clinice - clinical trials)

În ceea ce urmează mă voi referi numai la prima categorie, studii analitice de tip **observațional**, *acesta fiind tipul de studii folosit de mine în cadrul acestei lucrări.*

STUDII ANALITICE OBSERVAȚIONALE – SUBTIPURI:

A. Studii de tip Case-Control (Case-Control studies)

- Gruparea în tabele de contingență se face în funcție de prezența sau absența bolii:
 - » Persoane care prezintă boala
 - » Grup de comparație
- Practic un astfel de studiu compară expunerea în rândul persoanelor bolnave (grup - cazuri) cu expunerea în rândul persoanelor sănătoase (grup - control)

Este studiul epidemiologic cel mai des utilizat, în ciuda potențialelor erori, ce pot apărea relativ frecvent fără o grupare atentă în tabele de contingență potrivite scopului propus

Avantaje:

- » Tipul de studiu cel mai potrivit în cercetări legate de boli cu perioada lungă de incubare
- » Eficient din punct de vedere al timpului necesar și al costurilor aferente
- » Potrivit pentru studiul unor boli rare

Poate fi utilizat pentru studierea simultană a mai multor factori de risc, în cadrul aceleași boli

Dezavantaje:

- » Lipsă de eficiență în cazul studierii unor expuneri la factori de risc rar întâlniți (cum ar fi studierea efectelor unor accidente industriale)
- » Nu permite calculul direct al ratei de incidență a bolii
- » Uneori poate fi extrem de greu stabilirea unor relații temporale între expunere și apariția bolii

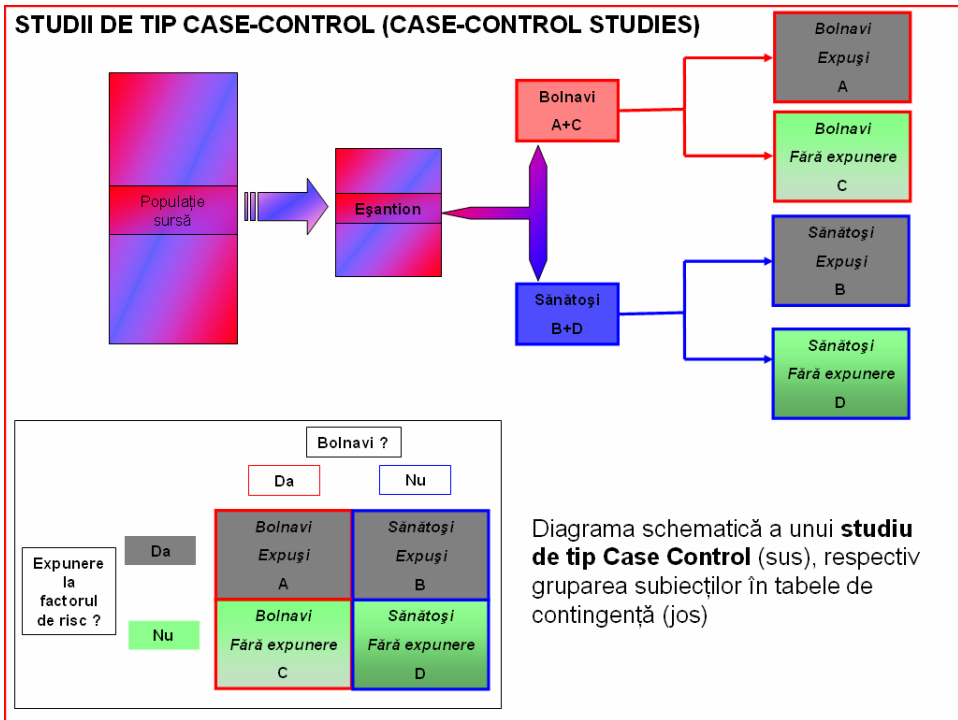


Figura VIII.1. Diagrama schematică a unui studiu de tip Case-Control

Exemplu: - Studiul asocierii dintre apariția limfoamelor nonhodgkin și infecția cu virusul Epstein-Bar sau cu *Helicobacter pylori* (factori biologici de risc)

ASPECTE CE TREBUIE LUATE ÎN CONSIDERARE ÎN CAZUL UNUI STUDIU DE TIP CASE-CONTROL:

- **definirea și selecția cazurilor**
- **selecția cazurilor pe baza incidenței sau a prevalenței** (cu alte cuvinte cazurile studiate erau diagnosticate înainte de începerea studiului sau sunt cazuri noi diagnosticate recent sau chiar la momentul studiului)
- **validitatea studiu versus generalizarea studiului** (Poate fi studiul considerat valid? – *probleme legate de designul studiului*; Pot fi concluziile lui extinse la totalul populației ? – *există alte studii similare ce au ajuns la concluzii asemănătoare?*)
- **sursa loturilor control:** clinici, eșantioane din populație în general, serii speciale de loturi martor etc.
- **raportul cazuri/control** (martor)
- **felul cum se face constatarea stării de boală, respectiv a expunerii la factorul de risc**

B. Studii de tip cohortă (Cohort studies)

- Subiecții studiului sunt clasificați pe baza expunerii la un anumit factor de risc
- Populația este urmărită pentru a determina prezența (apariția) bolii în cadrul acesteia ;
- În funcție de modul de urmărire a prezenței bolii în cadrul populației deosebim:
 - » Studii de tip cohortă **prospective**, respectiv **retrospective**

- » **Studiul de tip prospectiv** are ca punct de pornire timpul prezent, cercetătorul investigând populația (eșantionul) prospectiv (în viitor); **avantajul** major al unui astfel de studiu constă în faptul că cercetătorul poate colecta orice informație crede el de cuviință că are fi utilă.
- » În cazul studiului retrospectiv, cercetătorul **identifică** eșantionul (populația studiată – cohorta) și **studiază** acea populație de-a lungul unui anumit interval de timp **precedent** momentului în care a decis să înceapă studiul; **dezavantajul** unei astfel de abordări constă în faptul că cercetătorul are la dispoziție doar informații colectate în trecut, fără a mai putea avea acces la eventuale informații utile, dacă acestea nu au fost deja colectate și arhivate în vreun fel.

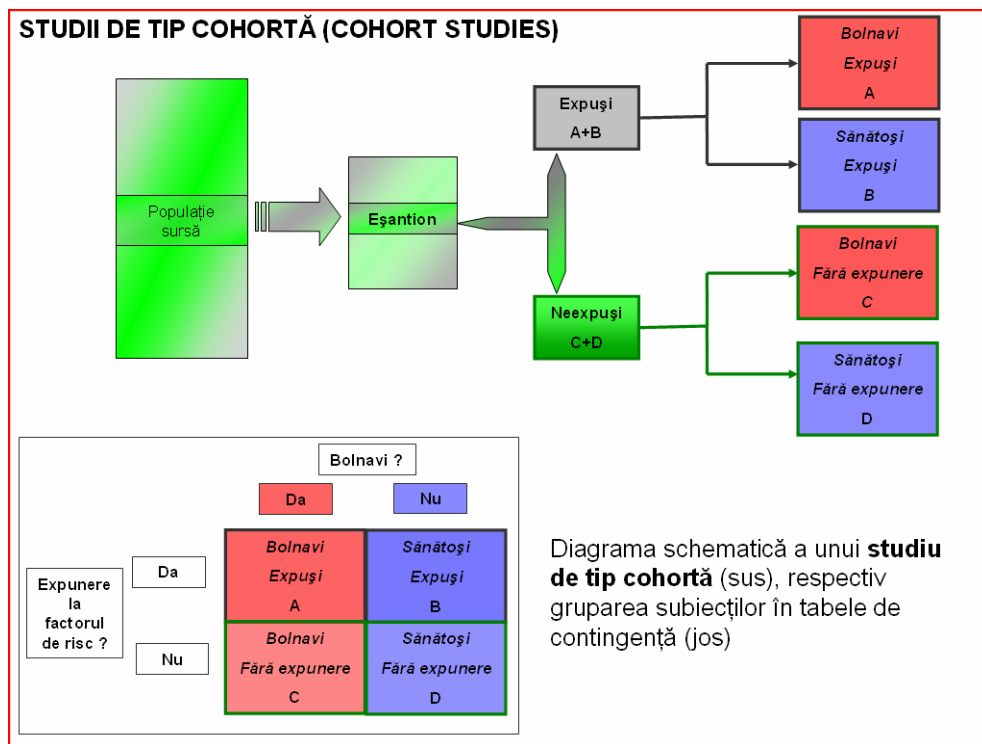


Figura VIII.2. Diagrama schematică a unui studiu de tip Cohortă

Avantaje:

- » Astfel de studii pot măsura incidența bolii și deci și cuantifica riscul
- » Expunerea precede apariția bolii
- » Poate servi la studiul unei mari categorii de boli
- » Este foarte eficient în cazul unor **expuneri la factori de risc rar întâlniți**, cum este cazul expunerilor datorate unor **accidente industriale**, aceasta deoarece *investigatorul poate să aleagă eșantionul avînd ca principal criteriu expunerea la factorul de risc*

Dezavantaje:

- » Necesită un număr mare mare de subiecți
- » Relativ lipsit de eficiență în cazul unor boli rare
- » Perioadă lungă de urmărire a subiecților studiului
- » Subiecții pot suferi schimbări în starea de sănătate de-a lungul perioadei de studiu, fără ca investigatorul să aibă cunoștință de acest lucru
- » Necesită, în general, costuri mari pentru realizare

Exemplu: Expuneri neobișnuite (accidente industriale), cum este și cazul accidentului nuclear de la Cernobîl (1986), **care ne-a permis efectuarea, unui studiu retrospectiv de tip cohortă** în scopul *studierii rolului radiațiilor ionizante ca factor de risc în apariția unor leucemii și limfoame*

**ASPECTE CE TREBUIE LUATE ÎN CONSIDERARE ÎN CAZUL
UNUI STUDIU DE TIP COHORTĂ:**

- **modul de selecție (grupare) a populației expuse**
- **modul de selecție (grupare) a populației neexpuse** (grup de comparație sau martor)
- **datele legate de sursa de expunere la factorul de risc**
- **datele legate de variabilele de ieșire** (apariția bolii, indicatorii folosiți, eventualele surse de eroare)

Selecția unui anumit tip de studiu (study design)

Trebuie să țină cont de următorii factori și criterii:

❖ Criteriul gradului de cunoaștere științifică a fenomenului studiat (bolii)

» Studiu ce **aduce informații noi** sau studiază o **boală nouă** (vezi Severe Acute Respiratory Syndrome – SARS)

» Studiu ce **adaugă** informații relevante despre o boală

» Studiu ce **confirmă** informații deja existente despre o boală

❖ Criteriul ipotezei de lucru

» Studiu ce servește la **generarea** de ipoteze

» Studiu ce servește la **testarea** unor ipoteze de lucru

❖ Criteriul epidemiologic / statistic

» Studiu ce servește la verificarea unor ipoteze privitoare la **boală** (boala – subiect principal de studiu)

» Studiu ce servește la verificarea unor ipoteze privitoare la **expunere** (expunerea la factorul de risc- subiect principal de studiu)

» Mărimea eșantionului și reprezentativitatea lui

❖ Criteriul resurselor disponibile (de obicei factorul limitant)

» Costuri

» Timp

Asociere versus cauzalitate

"The presence of an association ... in no way implies that the observed relationship is one of cause and effect." (Hennekens and Buring)

Principală idee care trebuie să răzbată în urma unei analize statistice efectuate în cadrul unui studiu epidemiologic, este faptul că *asocierea expunere (factor de risc)-boală* se referă strict la **existența unei dependențe statistice între două variabile**.

Pentru a putea considera, eventual, această **asociere** ca o **relație cauză-efect**, există o serie de criterii de interpretare a unui studiu epidemiologic a căror parcurgere este o condiție sine qua non.

Aceste criterii minimale sunt prezentate în cele ce urmează:

- Există o asociere expunere (factor de risc)-boală, validă din punct de vedere statistic?

DACĂ DA ...

- Se poate datora această asociere șansei sau nu...?

- Se poate datora această asociere unor erori ale studiului, legate de bias...?

- Se poate datora această asociere unor erori ale studiului, legate de anumite confuzii?

Criterii pentru stabilirea unei relații de tip cauză-efect

Pentru a răspunde la întrebarea: *Poate fi judecată o asociere statistică ca o relație de tip cauză-efect ?* – trebuie luate în considerare următoarele criterii:

1. Magnitudinea asocierii:

- i. Dacă a fost relevată existența unei asocieri statistic semnificative, cât de mare este magnitudinea acesteia ?

2. Credibilitatea biologică a ipotezei de lucru:

- i. Există un mecanism deja cunoscut care să explice științific asocierea ?

3. Consistența rezultatelor studiului vis-à-vis de alte studii similare:

- i. Există alte studii care să confirme concluziile noastre?

4. Secvența de timp:

- i. Pentru cazurile studiate, precede expunerea apariția bolii cu o perioadă de timp consistentă (a se vedea timpul de inducție al bolii) pentru un mecanism biologic credibil?

5. Relația “doză”- răspuns:

- i. Putem vorbi de un gradient al riscului ce poate fi relaționat cu nivelul expunerii?

6. Poate fi considerată asocierea ca fiind specifică?

- i. Câți factori cauzează boala?
- ii. Câte boli pot fi cauzate de factorul respectiv?

Anexe - Exerciții recapitulative

Anexa 1 - Statistică descriptivă, exerciții recapitulative



Scurtă recapitulare

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

Începuturile biostatisticii au fost determinate de nevoia obținerii unor informații cantitative dintre cele mai simple, formulate de regulă sub forma „câți bolnavi ? ” , „câți decedați ? ” , etc. Cu timpul s-a constatat însă că asemenea metode sunt insuficiente pentru caracterizările fenomenelor, că există o variație în răspunsurile care se obțin între diverse măsurători sau, cu alte cuvinte, că fenomenele biologice sunt caracterizate prin variabilitate. Dar și în aceste condiții, 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 câștigat 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 și a calculului probabilităților în special.



Indicatori statistici

Principalii indicatori care caracterizează un șir de date sunt fie indicatori de tendință centrală, fie indicatori ce caracterizează împrăștierea datelor în jurul unei valori medii.

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

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

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

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

Mediana - este acea valoare din șirul de date care împarte în două părți egale șirul ordonat de valori (atenție, șirul 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 $k = \frac{n}{2} + 1$. Dacă n

este par, deci avem un număr par de valori, mediana este definită ca fiind

$$M_e = \frac{x_k + x_{k+1}}{2} \text{ unde}$$

$$k = n/2.$$

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

Amplitudinea - este diferența dintre valoarea maximă și cea minimă

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

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

Dispersia (varianța) notată s_x^2 este un indicator de împrăștiere 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 și valoarea medie a șirului 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) și de unitățile de măsură folosite. Se poate considera că un coeficient de variație sub 10% indică o dispersie mică (o împrăștiere), adică seria este omogenă. Un coeficient între 10% și 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 ușurează 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 aceleiași valori

Dacă datele pe care le studiem conțin valori care se repetă des, se obișnuiește să se grupeze datele care au aceeași valoare. Numărul de apariții ale unei valori anume se numește frecvența de apariție și se notează cu f_i .

Presupunem că în urma măsurătorilor am obținut șirul 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

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

$$\text{Mediana} - \text{este } x_k \text{ unde } k = \frac{\sum_{i=1,n} f_i + 1}{2}$$

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

Folosirea calculatorului în statistica descriptivă. Aplicații și exemple.

Apariția calculatoarelor a însemnat un moment crucial în evoluția analizei statistice. Puterea de calcul considerabilă a acestora a permis nu numai efectuarea unor analize statistice din ce în ce mai complexe, ci și punerea la punct a unor noi teste, respectiv modalități de analiză statistică.

În mod obișnuit, există două modalități de utilizare a calculatorului în analiza statistică a datelor, modalități ce diferă prin tipul de software folosit:

- *folosirea unor programe obișnuite de calcul tabelar* (fie comerciale - MS Excel, fie gratuite, în regim Open Source - Gnumeric etc) ce au încorporate însă și module pentru analiză statistică. Astfel de programe oferă acces însă doar la un număr limitat de indicatori și/sau teste statistice, începând cu calcularea unor indicatori din domeniul statisticii descriptive (medii, dispersii, intervale de încredere) și terminând cu o serie de teste statistice devenite “clasice” (testul t, testul CHI^2 , testul F etc). Principalul lor atu îl constituie simplitatea în utilizare, în special pentru utilizatorul deja familiarizat cu programele de calcul tabelar.

- *folosirea unor programe dedicate pentru analiza statistică*, software ce oferă de obicei posibilitatea efectuării unor analize statistice complexe,

folosind un număr însemnat de teste și diferite tipuri de abordări în mersul analizei. Principalul dezavantaj al acestui tip de software rezidă tocmai în complexitatea lui, utilizatorul trebuind să aibă un bagaj complet de cunoștințe teoretice și practice legate de analiza statistică ce urmează să o efectueze. Există un număr mare de astfel de programe, de la soluții comerciale, uneori costisitoare (SPSS, SAS, Statistica for Windows, GraphPad Instat, GraphPad Prism etc) până la soluții gratuite (EpiInfo, OpenEpi, SalStat etc).

Statistică descriptivă folosind MS Excel sau alte programe de calcul tabelar

Există două modalități de a face o serie de analize statistice folosind Excel:

Prima modalitate, mai laborioasă, presupune folosirea unor funcții speciale (utilitare) pentru analiza statistică existente în program alături de alte tipuri de funcții Excel (funcții pentru calcule matematice, financiare etc). Aceste funcții sunt disponibile prin intermediul opțiunii Function din meniul Insert.

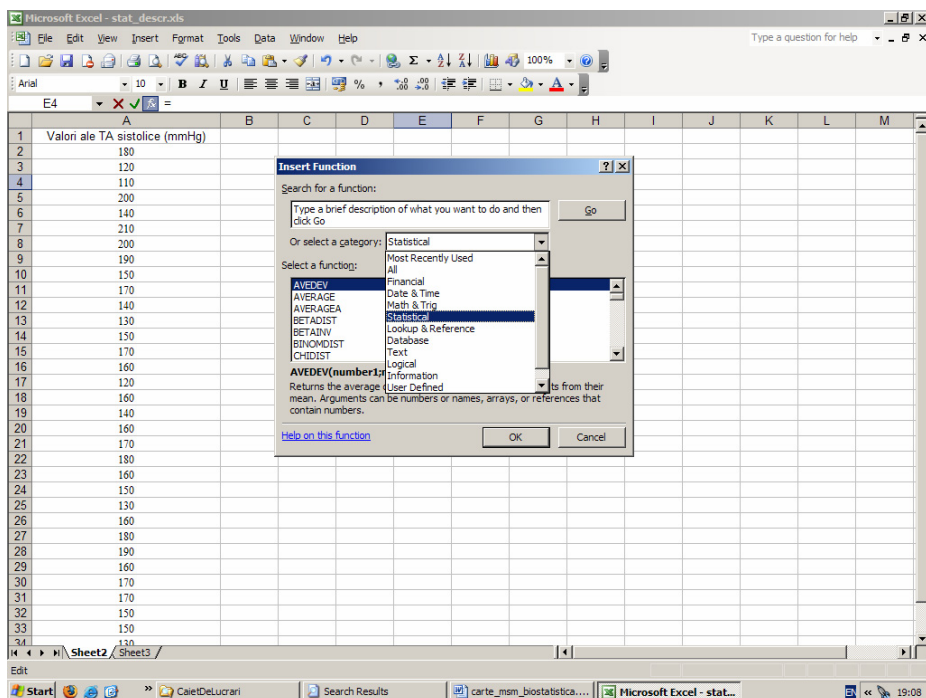


Figura A1.1. *Utilizarea funcțiilor statistice în MS Excel*

Principalele funcții (utilitare) Excel necesare calculării indicatorilor utilizați în statistica descriptivă sunt cele de mai jos:

A. Indicatori ai tendinței centrale

Media aritmetică – funcția AVERAGE()

Mediana – funcția MEDIAN()

Modul – funcția MODE()

B. Indicatori ai dispersiei (împrăstierii) datelor în jurul valorii medii

Dispersia (varianta) – funcția VAR() (cu variantele VAR, VAR A, VARP, VARPA)

Abaterea standard - funcția STDEV() (cu variantele STDEV, STDEVA, STDEVP, STDEVPA)

Asimetria (skewness) - funcția SKEW()

Kurtosis (aplatizarea - excesul) - funcția KURT ()

Valoare cea mai mică dintr-un set de valori – funcția MIN()

Valoare cea mai mare dintr-un set de valori – funcția MAX()

Lucrul cu funcțiile programului Excel presupune parcurgerea următorilor pași:

- alegerea unei celule unde va fi afișat rezultatul
- apelarea funcțiilor Excel prin opțiunea Function din meniul Insert
- selectarea tipului de funcție dorit (Math, Financial, Statistical etc)
- selectarea celulelor unde se află seria dorită de date
- dacă este cazul, setarea anumitor parametrii utilizați de respectiva funcție pentru efectuarea corectă a calcului (de pildă nivelul dorit de semnificație sau tipul de test, dacă sunt disponibile mai multe variante ale aceluiași test)
- în final, în celula dorită va fi afișat rezultatul calculului, în timp ce în bara cu formule va fi afișată funcția utilizată și setul de date folosit

Exemplu.

Fie o serie de 33 de date numerice obținute experimental (de exemplu valori ale tensiunii arteriale sistolice în mmHg): 180, 120, 110,

200, 140, 210, 200, 190, 150, 170, 140, 130, 150, 170, 160, 120, 160, 140, 160, 170, 180, 160, 150, 130, 160, 180, 190, 160, 170, 170, 150, 150, 130. Să se calculeze, folosind funcțiile din programul Excel media, mediana, modul, varianta, abaterea standard, valoarea maximă și minimă, pentru acest set de date.

Rezolvare.

- Se introduc datele în program

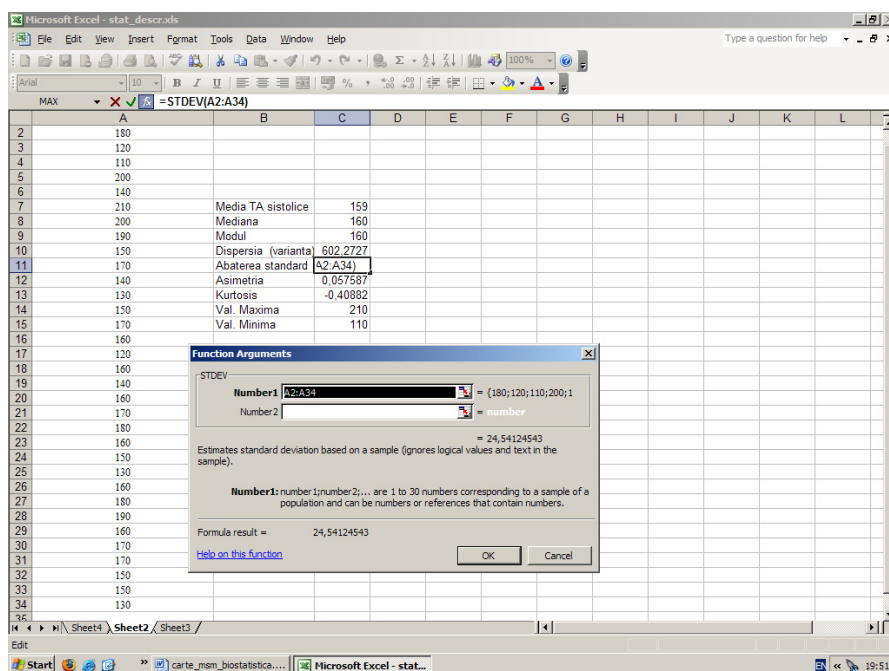


Figura A1.2. Introducerea datelor în fereastra de dialog în cazul funcțiilor statistice din MS Excel

Cea de-a doua modalitate, mult mai elegantă și cu o organizare mult mai elaborată a modului de prezentare a rezultatelor analizei, presupune instalarea folosirea unui modul special al programului Excel, modul ce,

odată instalat, poate fi apelat prin alegerea opțiunii Data Analysis din meniul Tools al programului Microsoft Excel.

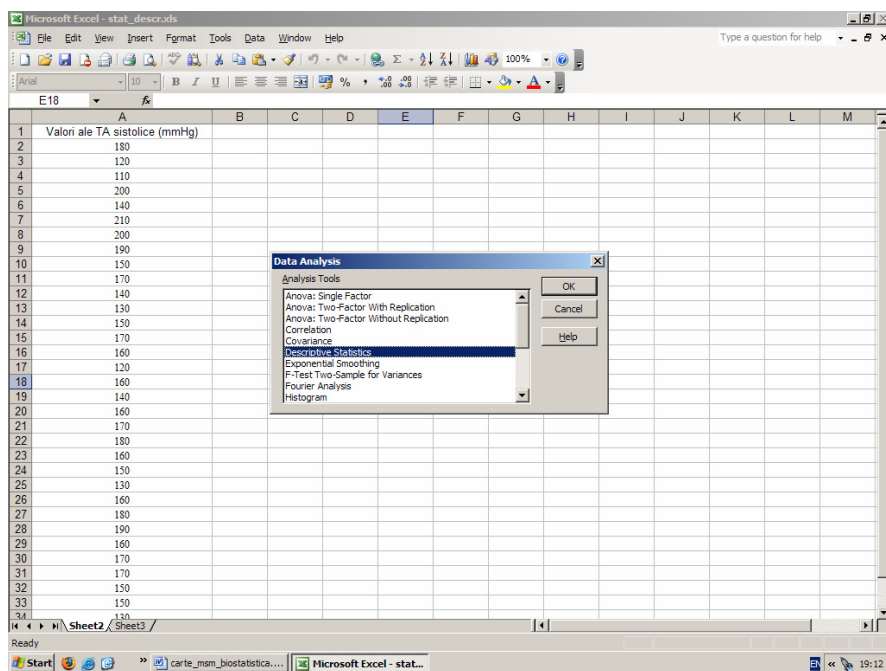


Figura A1.3. *Data Analysis ToolPak din MS Excel*

Statistică descriptivă – Exerciții recapitulative

1. Pentru stabilirea cantității de adenină dintr-o soluție dată, s-au făcut măsurători spectrofotometrice și s-au obținut 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 programului Excel, valoarea medie, dispersia, amplitudinea, abaterea standard, amplitudinea relativă, coeficientul de variație.

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

Greutatea (în grame)	Frecvența
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 și coeficientul de variație.

3. Să se calculeze valoarea medie, amplitudinea, amplitudinea relativă, dispersia, abaterea standard și coeficientul de variație al duratei de spitalizare în cazul unui grup de 200 de bolnavi internați 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

Anexa 2 - Teste statistice de semnificație, exerciții recapitulative

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 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 eșantioane de lucru: un grup de test extras din prima populație și un grup martor, extras din a doua populație. Se culeg și se înregistrează datele studiului. Se fac următoarele notații:

X_i reprezintă valorile înregistrate în grup 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:

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

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 valoare mai mică sau egală cu 0,5, atunci se respinge ipoteza nulă H_0 și se acceptă ipoteza alternativă H_1 , prin urmare există o diferență semnificativă între mediile caracteristicii în cele două populații.
- Dacă probabilitatea p are o valoare 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 ușor 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 ipotezei 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 și se presupune că populațiile au aceeași dispersie.

☐ **3** - dacă grupurile de date sunt independente și se presupune că populațiile au dispersii diferite.

Spre exemplu, dacă seriile de valori sunt conținute în zonele B7:B25 și E8:E35 și grupurile de date sunt independente, atunci conținutul funcției TTEST este:

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

Testul STUDENT – exemple și exerciții recapitulative

1. 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 creșterea secreției de calciu. Pentru a realiza acest lucru, mai întâi introduceți datele de mai sus într-o foaie de calcul tabelar și apoi aplicați testul Student cu un capăt (1 tails), de tipul 2 (two-sample equal variance) și ipoteza 0. Dacă probabilitatea P obținută este mai mică decât 0,05 atunci medicamentul are efect.

Ipoteze:

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

H_1 : medicamentul are efect în creșterea secreției de calciu.

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

Rezultate

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

2. 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 și de tipul 1 (grupuri dependente). Dacă probabilitatea p obținută este mai mică decât 0,05 atunci există diferențe semnificative.

3. 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 eficient ? Pentru a putea răspunde la această întrebare utilizați testul STUDENT cu două capete și pentru grupuri dependente (tipul 1) .

4. 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).

5. S-a măsurat uremia la două loturi de câte 10 bolnavi de gută, dintre care unii au fost tratați cu un anumit medicament, în timp ce ceilalți li s-a administrat un produs placebo; s-au obținut următoarele rezultate:

Nr. Subiect	Valoarea uremiei în lotul tratat (mg/l)	Valoarea uremiei 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).

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

Rezultatele obținute în urma analizării probelor sunt sintetizate 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).

7. S-a măsurat valoarea tensiunii arteriale sistolice la grup de pacienți diagnosticați cu stenoză și la un grup de pacienți asimptomatici. Să se

determine dacă media celor două serii de date diferă semnificativ. Se va utiliza testul STUDENT cu două capete și de tipul 2 (pentru grupuri independente).

Rezultatele măsurătorilor sunt sintetizate în tabelul următor.

Valoarea sistolică		
Pacientul	Simptomatici	Asimptomatici
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



Testul CHI^2

Testul CHI^2 este utilizat în analiza statistică în următoarele cazuri:

➔ în studiile epidemiologice pentru identificarea unei **asocieri** între un **factor de risc** și o **boală**. De exemplu, se poate aplica testul CHI pentru stabilirea unei eventuale legături între fumat și moartea prematură ca urmare a unei boli cardiovasculare, sau a unei legături între expunerea la o anumită substanță chimică și 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 și se înregistrează datele studiului. Subiecții sunt clasificați ca bolnavi sau nu, expuși la un anumit factor de risc sau nu, etc. Se stabilește numărul de subiecți care fac parte din fiecare clasă. Numărul de subiecți care fac parte din clasa i relativ la prima caracteristică și din clasa j relativ la a doua caracteristică se notează cu O_{ij} și se numește 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} și se calculează după formula:

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

➔ Se calculează valoarea statistică a testului CHI după formula:

$$\Rightarrow \chi_C^2 = \sum_{\substack{i=1,n \\ j=1,m}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

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

$$gl = (nr \text{ rânduri} - 1) \cdot (nr \text{ coloane} - 1)$$

➔ 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 $\leq 0,5$ se respinge ipoteza H_0 .

➔ Stabilirea concluziei testului CHI.

☞ Dacă probabilitatea p rezultată din calcul are o valoare mai mică sau egală cu 0,5 , atunci se respinge ipoteza nulă H_0 și se acceptă ipoteza alternativă H_1 , prin urmare există o asociere între cele două caracteristici studiate.

☞ Dacă probabilitatea p rezultată din calcul are o valoare mai mare decât 0,5 , atunci se acceptă ipoteza nulă H_0 , prin urmare nu există o asociere între cele două caracteristici studiate.

Utilizarea funcției CHITEST a programului Excel

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

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

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

Figura A2.1 . Gruparea datelor în MS Excel pentru testul CHI^2

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 A2.1)

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

Matricea valorilor estimate se contruiește 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 și anume: =CHITEST(A1:E3;A10:E12). Rezultatul întors de funcția CHITEST este valoarea p .

Testul CHI^2 – Exerciții recapitulative.

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

	Hipertensiune	Tesiune normală	Total
Cu contraceptive orale	8	32	40
Cu alte contraceptive	15	45	60
	23	77	100

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

Pentru a obține rezultatul studiului, mai întâi introduceți datele de mai sus într-o foaie de calcul tabelar. Apoi calculați frecvențele estimate astfel:

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

După aceea aplicați testul CHI. Dacă probabilitatea P obținută este mai mică decât 0,05 , atunci există diferență, deci există o legătură.

Ipoteze:

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

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

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

Dacă probabilitatea P obținută 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 și hipertensiune.

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

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

Stabiliți există o asocieră (legătură) între severitatea cancerului ovarian și nivelul de stres.

3. Se studiază asocieră 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

formeii 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 dorește 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 eșantion.

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 femeii negravidă, femeii cu sarcini normale în luna a 9-a și femeii cu disgravidii tardive, privind valorile medii, în g/zi ale aldosteronului, cortizonului și cortizolului.

Datele studiate sunt prezentate în tabelul următor:

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

7. 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 și cele care suferă de această boală:

Tratament	Holeră	
	+	-
+	5	431
-	9	291

Să se stabilească dacă tratamentul afectează numărul de persoane ce suferă de holeră, adică dacă există o asociere statistic semnificativă între numărul de persoane ce suferă de holeră și numărul de persoane supuse tratamentului.

7. În urma aplicării unui vaccin, s-a înregistrat numărul de persoane care s-au îmbolnăvit și care nu s-au îmbolnăvit. De asemenea, s-a înregistrat

și numărul persoanelor care s-au îmbolnăvit din rândul persoanelor nevaccinate. Se pune problema: diferențele între bolnavii vaccinați și cei nevaccinați sunt semnificative sau nu ?

Datele studiate sunt prezentate în tabelul următor:

	bolnavi	sănătoși
vaccinați	20	74
nevaccinați	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	
	Cu avorturi spontane	Fără avorturi spontane
expuse la pesticid	30	10
neexpuse	70	90

Să se stabilească existența unei eventuale legături între expunerea la pesticid și avortul femeilor.

9. Se studiază efectul obținut asupra numărului de carii prin efectuarea unui instructaj privind igiena orală unui număr de copii aleși aleator. La 50 de copii li s-a făcut un instructaj privind igiena orală iar la 50 de copii aleși 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:

Instructaj	<i>număr de carii noi</i>		
	0-1	2-3	4-5
+	30	15	5
-	20	15	15

10. S-a studiat asocierea dintre prezența anemiei la un lot de subiecți și 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 și 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

Anexa 3 – Corelații și regresii, exerciții recapitulative

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 greutateii, 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 medicamentului separat pe diferite regimuri alimentare ne conduce la concluzia că există o interacțiune între doi factori: regimul alimentar și medicamentul.

Asociere și cauzalitate – coeficientul de corelație

În acumularea și evidența datelor științifice apar o serie de probleme specifice, cum ar fi problema asocierii (dependenței) între două variabile. Se pune problema: există o dependență între sărăcie și 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 și 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 și poate varia între -1 și $+1$.

Dacă x_i sunt valorile măsurate ale variabilei X și y_i sunt valorile măsurate ale variabilei Y, atunci coeficientul de corelație se calculează astfel:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2} \sqrt{n(\sum y^2) - (\sum y)^2}}$$

unde n = numărul perechilor de date.

Putem întâlni următoarele situații:

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 și 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 și 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 și 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 și $+1$ sau între -1 și 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 și greutateii pentru un grup de persoane și calculăm coeficientul de corelație, vom obține o valoare pozitivă, dar o valoare mai mică decât 1.

Corelație și cauzalitate.

Problema determinării tăriei corelației dintre variabilele aleatoare este o problemă relativ dificilă, ce depinde de domeniul aplicațiilor, precum și de mulți alți factori. Variabilele psihologice sunt mai dificil de măsurat cu exactitate și 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 de altfel și avantajul că pot fi măsurate cu mai mare precizie.

Ca un exemplu, corelațiile dintre aptitudinile verbale și cele non-verbale la copiii școlari, măsurate cu ajutorul unor teste standard, variază între 0,44 și 0,77 depinzând mediul de proveniență al acestora și 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, aceasta 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 (diagramă), numit XY Scatter.

Între cele două variabile există o corelație puternică dacă punctele reprezentate grafic sunt grupate de-a lungul unei drepte (figura). 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 următoarea întrebare: 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 greși este mai mic, deci estimarea este mai sigură. Să ne reamintim câteva noțiuni importante:

- **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 și 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 depășește acest prag critic, variabilele studiate se consideră corelate.

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

Anexa 1 prezintă tabelul cu pragurile critice pentru nivelele de semnificație 0,10 ; 0,05 ; 0,02 și 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:

***H0:** cele două variabile studiate nu sunt corelate.*

***H1:** cele două variabile studiate sunt corelate.*

În continuare se efectuează pașii următori:

1. **Calculăm coeficientul de corelație r** asociat datelor x_i și y_i , cu ajutorul formulei prezentate mai sus sau cu ajutorul programului Microsoft Excel și 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 și 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 urmărește 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 și sunt mai puține șanse să greșim deoarece marja de eroare este mai mică.

4. În toate aceste cazuri se respinge ipoteza H_0 și 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 de semnificație 0,05 sau 0,01, care sunt considerate suficiente.

Să luăm un exemplu ipotetic: studierea corelației dintre nivelul de amfetamină din plasmă și intensitatea psihozei.

Datele rezultate în urma măsurărilor sun cele din tabelul următor:

Nr. subiect	Intensitatea psihozei	Concentrația amfetaminei 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 datelor cu ajutorul unei diagrame scatter-plot este cea din figură:

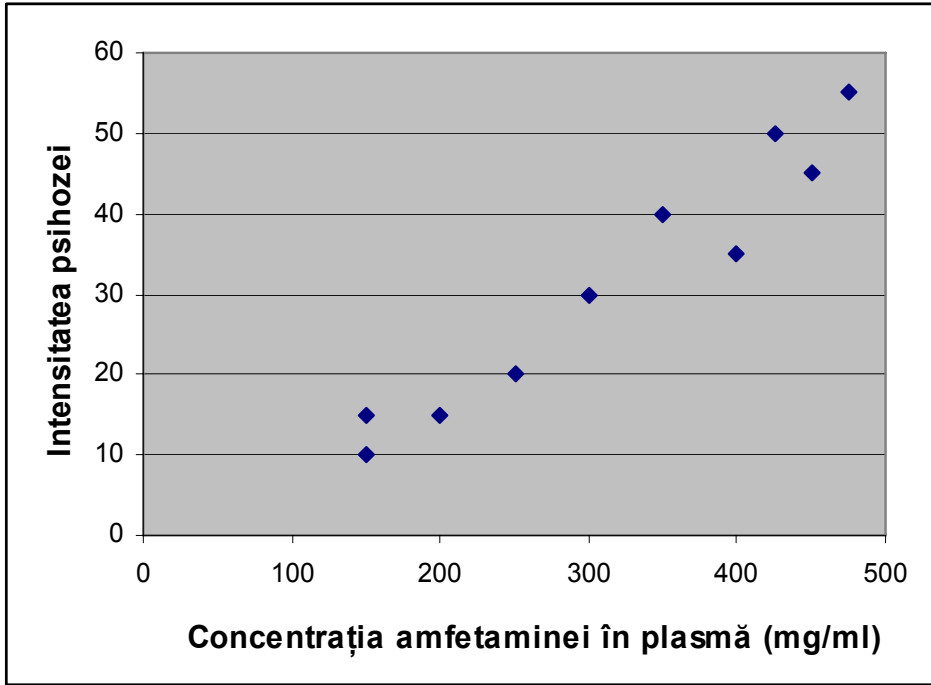


Figura A3.1. Reprezentarea grafică a datelor într-o diagramă scatter-plot

Coeficientul r rezultat din calcul este 0,96738 . Numărul gradelor de libertate este $10-2=8$. În anexa 1, pragul critic pentru 8 grade de libertate și nivelul de semnificație 0,05 este 0,6319, iar pentru nivelul de semnificație 0,01 este 0,7646.

REZULTAT

$r > 0,7646$ - se respinge ipoteza H_0 și se acceptă ipoteza H_1 cu un nivel de semnificație de 0,01.

CONCLUZIE

Intensitatea psihozei este corelată cu nivelul de Amfetamină din plasmă.

EXERCITII

1. Studiindu-se relația dintre doza unui medicament (exprimată în multipli ai unei doze minime) și durata bolii (exprimată prin numărul de zile de boală), s-a obținut următoarea relație:

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 și să se verifice dacă există o legătură între doza medicamentului și durata bolii.

2. Să se aprecieze existența și gradul legăturii dintre consumul de alcool (vin în litri) pe cap de locuitor, pe lună, și vârsta medie de debut a cirozei hepatice.

Datele colectate în urma studiului sunt prezentate în tabelul următor:

Nr. subiect	litri 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, sub formă de diagramă scatter.

3. Să se aprecieze gradul și sensul legăturii dintre temperatură și puls la un lot de 20 de bonavi. Datele studiului 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, sub formă de diagramă scatter.

4. 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 și menarha (vârsta de început a menstruației) și 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

Nr. subiect	vârsta	menarha
26	50	12
27	50	14
28	49	14

6. Într-un studiu cuprinzând cazuri de stenoză, s-au măsurat valorile IMT maxim și valoarea sistolică la diverși pacienți. Se pune problema există o legătură între valoarea sistolică și valoarea IMT maxim ? Datele colectate în urma efectuării studiului sunt prezentate în tabelul următor.

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.

7. O companie farmaceutică a încercat să evalueze relația dintre doza ingerată a unui nou medicament hipnotic și 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.

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

Datele colectate 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.

9. Într-un studiu cuprinzând multe cazuri, s-a descris relația dintre durata sarcinii exprimată în săptămâni și greutatea la naștere (g). Prezintă câteva dintre datele experimentale, care se referă la perioada între săptămâna a 26-a și săptămâna a 37-a. Se cere să se studieze statistic relația dintre cele două variabile. Datele colectate sunt cele din tabelul următor. 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

Anexa 4 – Exemple de studii epidemiologice. Studiu de tip Case-Control

↪ **Studiul influenței unor factori de risc biologici - infecția cu *Helicobacter pylori* – în apariția unor limfoame maligne (limfoame nonHodgkin –LNH) în cazul populației județului Mureș [22]**

Ipoteza de lucru

Conform datelor din literatura de specialitate, *H.pylori* a fost clasificat ca fiind un agent cancerigen cu rol în etiopatogenia cancerului gastric, dar constatări mai recente au adus în atenție această bacterie ca factor de risc în sfera etiopatogeniei limfoamelor.

Scopul studiului

Studiul prezenței infecției cronice cu *Helicobacter pylori* la pacienți cu limfoame nonHodgkin comparativ cu un lot martor.

Material și metodă:

Studiul efectuat este de tip Case-control prospectiv și prezintă următoarele caracteristici :

- este un studiu ce încearcă verificarea unor informații cunoscute cu privire la influența infecției cu *Helicobacter pylori*, ca factor de risc în apariția limfoamelor nonHodgkin
- studiul a avut ca principal criteriu de grupare boala (limfoamele nonHodgkin) și abia apoi expunerea (infecția)
- a fost precedat de o analiză descriptivă a cazurilor

- în ceea ce privește mărimea eșantionului, în cazul acestui tip de studiu am investigat un lot de 81 de bolnavi cu limfoame nonHodgkin, respectiv un lot martor (81 de subiecți)
- în paralel am urmărit vârsta, sexul și mediul de proveniență (urban/rural) al subiecților aflați în studiu, în încercarea de a constata dacă acești factori influențează în vreun fel apariția infecției cronice cu H.pylori
- analiza statistică preconizată a constatat în calcularea Odds Ratio (OR), respectiv a intervalului de încredere pentru acesta (CI); OR și CI au fost calculate folosind testul CHI^2 cu corecție Yates.

În acest studiu, pentru depistarea infecției cu H.pylori s-au efectuat determinări serologice de anticorpi de tipul IgG prin metoda ELISA, utilizând kituri imunoenzimatiche Dienes-Enzywell Helicobacter pylori, IgG. Infecția acută cu H.pylori induce apariția anticorpilor de tip IgA, în timp ce infecția cronică determină apariția anticorpilor de tip IgG.

Pentru acest studiu am ales infecția cronică cu H.pylori, care ar putea avea relevanță în etiopatogenia limfoamelor.

Pentru stabilirea infecției cu Helicobacter pylori s-au folosit metode serologice de tipul tehnicii ELISA pentru detectarea IgG anti-Helicobacter pylori.

Studiul s-a efectuat pe cazuistica Clinicii Medicale I Tg.Mureș pe perioada 2001-2002. Lotul martor provine din cazuistica Clinicii de Boli Infecțioase II Tg.Mureș unde s-au efectuat investigațiile pentru această categorie de pacienți *cu alte diagnostice decat cele hematologice.*

Diagnosticul de limfom nonHodgkin confirmat clinic și anatomopatologic, a fost completat de investigații serologice privind infecția cu H.pylori.

Rezultate și discuții.

Lotul cu LNH cuprinde 81 de pacienți, 44 (54,32 %) bărbați și 37 (45,68 %) femei. Vârsta acestora este cuprinsă între 20 și 78 de ani.

Tabel A4.1. Caracteristicile lotului cu LNH în funcție de sex

Lot cu LNH	Bărbați	Femei
81	44	37
100 %	54,32 %	45,68 %

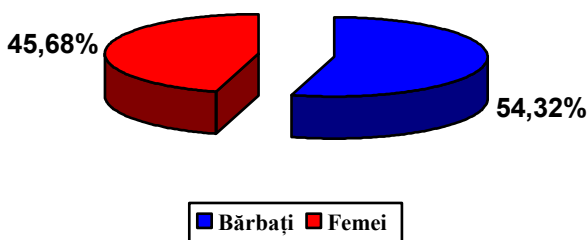
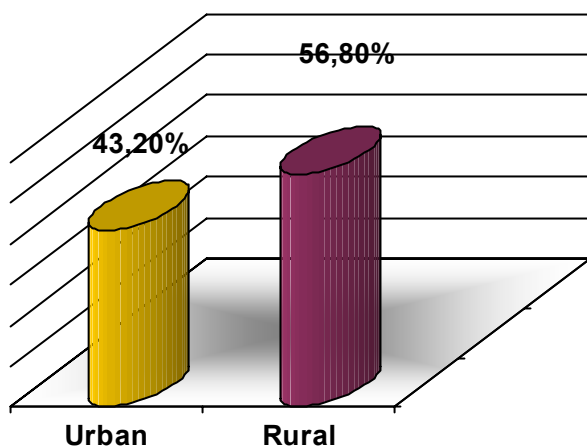


Figura A4.1. Repartiția procentuală pe sexe a lotului cu LNH

Un număr de 35 (43,20 %) de pacienți provin din mediul urban, iar restul 46 (56,80 %) din mediul rural.

Tabel A4.2 Caracteristicile lotului cu LNH după rezidență

Lot cu LNH	Urban	Rural
81	35	46
100 %	43,20 %	56,80 %

**Figura A4.2.** Repartiția procentuală a lotului cu LNH în funcție de rezidență

La pacienții cu LNH infecția cronică cu *Helicobacter pylori* este prezentă într-un procent ridicat, respectiv în 70,37 % din cazuri, în care rezultatele serologice au fost pozitive pentru IgG. Din totalul de 57 de bolnavi seropozitivi 34 sunt bărbați și 23 femei; 35 de bolnavi provin din mediul rural și 22 din mediul urban.

Tabel A4.3. *Caracteristicile bolnavilor seropozitivi din lotul cu LNH*

Infecția cu H.pylori+	Nr.cazuri LNH	Procente
Masculin	34	59,65 %
Feminin	23	40,35 %
Urban	22	38,60 %
Rural	35	61,40 %

La lotul martor, infecția bacteriană cronică este prezentă într-un procent ridicat, dar valoarea acestuia este totuși mai mică, de 60,49 % în comparație cu cea a lotului cu LNH respectiv 70,37 %. Dintre pacienții seropozitivi ai lotului de control, 29 au fost de sex masculin, iar 20 de sex feminin; 13 proveneau din mediul rural și 36 din urban.

Tabel A4.4. *Caracteristicile cazurilor seropozitive din lotul martor*

Infecția cu H.pylori+	Nr.cazuri martor	Procente
Masculin	29	59,18 %
Feminin	20	40,82 %
Urban	36	73,47 %
Rural	13	26,53 %

ANALIZA STATISTICĂ: Infecția cu *Helicobacter pylori* la pacienți cu LNH

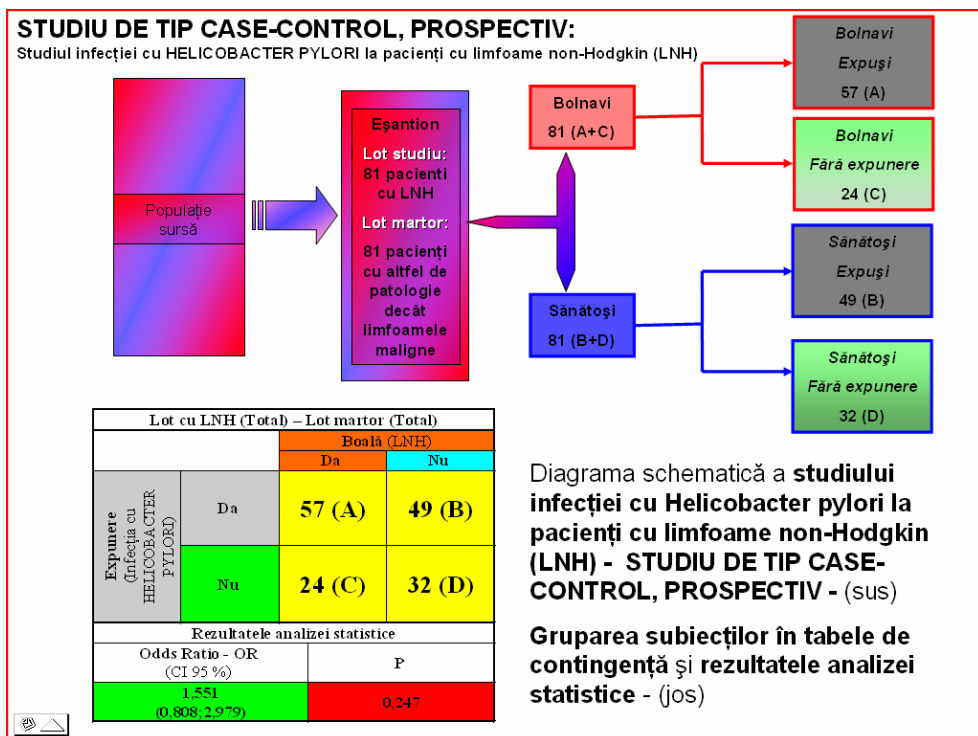


Figura A4.3. Diagrama schematică a studiului infecției cu *H.pylori* la pacienți cu LNH

Tabel A4.5. Analiza statistică : Lot cu LNH (Total) – Lot martor (Total)

Lot cu LNH (Total) – Lot martor (Total)			
		Boală (LNH)	
		Da	Nu
Expunere (Infecția cu)	Da	57	49

	Nu	24	32
Rezultatele analizei statistice			
Odds Ratio - OR (CI 95 %)		P	
1,551 (0,808;2,979)		0,247	

Se poate observa o **creștere nesemnificativă statistic** a ratei de șansă (Odds Ratio) OR=1,551 pe totalul cazurilor (comparație lot cu LNH -lot martor), tendință ce confirmă rezultatele altor studii.

Tabel A4.6. Analiza statistică : Lot cu LNH (Masculin) – Lot martor (Masculin)

Lot cu LNH (Masculin) – Lot martor (Masculin)			
		Boală (LNH)	
		Da	Nu
Expunere (Infecția cu HELICOBACTER PYLORI)	Da	34	29
	Nu	10	15
Rezultatele analizei statistice			
Odds Ratio - OR (CI 95 %)		P	
1,759 (0,686;4,508)		0,344	

Tabel A4.7. *Analiza statistică: Lot cu LNH (Feminin) – Lot martor (Feminin)*

Lot cu LNH (Feminin) – Lot martor (Feminin)			
		Boală (LNH)	
		Da	Nu
Expunere (Infecția cu HELICOBACTER PYLORI)	Da	23	20
	Nu	14	17
Rezultatele analizei statistice			
Odds Ratio - OR (CI 95 %)		P	
1,396 (0,553;3,530)		0,638	

În cazul analizei statistice efectuate pe loturi grupate după criterii demografice se constată o **creștere nesemnificativă statistic**, dar de magnitudine mare a OR, în cazul *populației masculine* (OR =1,759, comparativ cu OR =1,396, în cazul populației feminine).

Tabel A4.8. *Analiza statistică: Lot cu LNH (Urban) – Lot martor (Urban)*

Lot cu LNH (Urban) – Lot martor (Urban)			
		Boală (LNH)	
		Da	Nu
Expunere (Infecția cu	Da	22	36

	Nu	13	20
Rezultatele analizei statistice			
Odds Ratio - OR (CI 95 %)		P	
0,940 (0,391;2,260)		0,999	

Se constată o **asociere negativă, ne semnificativă statistic** factor de risc-boală, în cazul *populației urbane* (OR = 0,940).

Tabel A4.9. Analiza statistică: Lot cu LNH (Rural) – Lot martor (Rural)

Lot cu LNH (Rural) – Lot martor (Rural)			
		Boală (LNH)	
		Da	Nu
Expunere (Infecția cu HELICOBACTER PYLORI)	Da	35	13
	Nu	11	12
Rezultatele analizei statistice			
Odds Ratio - OR (CI 95 %)		P	
2,937 (1,041;8,284)		0,062	

Se constată o **creștere ne semnificativă statistic**, dar de magnitudine mare a OR, în cazul *populației rurale* (OR =2,937).

Concluzii

În cazul studiului privind influența infecției cronice cu *Helicobacter pylori* asupra apariției limfoamelor nonHodgkin, se poate observa o creștere nesemnificativă însă statistic a ratei de șansă - OR pe totalul cazurilor (OR=1,551), rezultate ce confirmă datele altor studii. O magnitudine mare a OR, nesemnificativă însă statistic, poate fi observată și în cazul *populației masculine* (OR=1,759), respectiv în cazul populației provenite din *mediul rural* ((OR=2, 937 – o posibilă explicație fiind o igienă mai precară în mediul rural). Trebuie remarcată în acest context și asocierea negativă nesemnificativă statistic între factorul de risc și boală, la populația din mediul urban.

Anexa 5 - Valori critice pentru testul F corespunzătoare pragului de semnificație $\alpha = 0,05$

	1	2	3	4	5	6	8	12	24	00
1.	161, ⁴	199, ⁵	215, ⁷	224, ⁶	230, ²	234, ⁰	238, ⁰	213, ⁰	249, ⁰	254, ²
2.	18,5	19,0	19,1	19,2	19,3	19,3	19,3	19,4	19,4	19,5
3.	10,1	9,55	9,28	9,12	9,01	8,94	8,84	8,71	8,64	8,53
4.	7,71	6,94	6,50	6,39	6,26	6,16	6,04	5,91	5,77	5,83
5.	6,61	5,79	6,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36
6.	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	8,84	3,67
7.	5,58	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23
8.	5,32	4,16	4,07	3,84	3,69	3,14	3,28	3,28	3,12	2,93
9.	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,07	2,90	2,71
10.	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,91	2,74	2,54
11.	4,84	3,98	3,59	2,38	3,20	3,09	2,95	2,79	2,61	2,40
12.	4,75	3,88	3,49	3,26	3,11	3,00	2,85	2,69	2,50	2,30
13.	4,67	3,80	3,41	3,18	3,02	2,92	2,77	2,60	2,42	2,21
14.	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,53	2,35	2,13
15.	4,51	3,68	3,29	3,06	2,90	2,79	2,74	2,48	2,29	2,07
16.	4,49	3,63	3,24	3,61	2,85	2,74	2,59	2,42	2,24	2,01
17.	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,38	2,19	1,96
18.	4,41	2,55	3,16	2,93	2,77	2,66	2,51	2,34	2,15	1,92
19.	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,31	2,11	1,88
20.	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,28	2,08	1,84
21.	4,32	3,47	3,07	2,84	2,68	2,57	2,42	2,25	2,05	1,81
22.	4,30	3,44	;	2,92	2,66	2,55	2,40	2,23	2,03	1,78
23.	4,28	3,42	3,03	2,80	2,64	2,53	2,38	2,20	2,00	1,76
24.	4,26	3,40	3,01	2,78	2,62	2,51	2,36	2,18	1,98	1,73
25.	4,24	3,38	2,09	2,76	2,80	2,49	2,34	2,16	1,96	1,71
26.	4,22	3,37	2,98	2,74	2,59	2,47	2,32	2,15	1,95	1,69
27.	4,21	3,35	2,96	2,73	2,57	2,46	2,30	2,13	1,93	1,67
28.	4,20	3,34	2,95	2,71	2,56	2,41	2,29	2,12	1,91	1,65

29.	4,18	3,33	2,93	2,70	2,54	2,43	2,28	2,10	1,90	1,64
30.	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,09	1,89	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,00	1,79	1,51
60	4,00	3,15	2,70	2,52	2,37	2,25	2,10	1,92	1,70	1,39
120	3,92	3,07	2,68	2,45	2,29	2,17	2,02	1,83	1,61	1,25
	3,81	2,99	2,60	2,37	2,21	2,09	1,94	1,75	1,52	1,00

Anexa 6 - Valorile critice pentru testul "t" Student corespunzătoare diferitelor praguri de semnificație α și numărului "n" al gradelor de libertate

Nivel de semnificație pentru testul bilateral									
$n \backslash \alpha$	0,50	0,20	0,10	0,05	0,02	0,01	0,002	0,001	0,0001
1	1,000	3,078	6,314	12,70	31,821	63,657	318,309	636,618	6366,198
2	0,816	1,886	2,290	4,303	6,965	9,925	22,327	31,598	99,992
3	0,765	1,638	2,353	3,182	4,541	5,841	10,214	12,924	28,000
4	0,741	1,533	2,132	2,77	3,747	4,604	7,173	8,610	15,544
5	0,727	1,476	2,015	2,571	3,365	4,032	5,893	6,869	11,178
6	0,718	1,440	1,943	2,44	3,143	3,707	5,208	5,959	9,082
7	0,711	1,415	1,895	2,365	2,998	3,499	4,785	5,408	7,885
8	0,706	1,397	1,860	2,30	2,896	3,355	4,501	5,041	7,120
9	0,703	1,383	1,833	2,262	2,821	3,250	4,297	4,781	6,594
10	0,700	1,372	1,812	2,228	2,764	3,169	4,144	4,587	6,211
11	0,697	1,363	1,796	2,201	2,718	3,106	3,025	4,437	5,921
12	0,695	1,356	1,782	2,17	2,681	3,102	3,930	4,318	5,694
13	0,694	1,350	1,771	2,16	2,650	3,055	3,852	4,221	5,513
14	0,692	1,345	1,761	2,145	2,624	2,977	3,787	4,140	5,363
15	0,691	1,341	1,753	2,131	2,602	2,947	3,733	4,073	5,239
16	0,690	1,337	1,746	2,12	2,583	2,921	3,686	4,015	5,134
17	0,689	1,333	1,740	2,11	2,567	2,898	3,646	3,965	5,014
18	0,688	1,330	1,734	2,101	2,552	2,878	3,610	3,922	4,966
19	0,688	1,328	1,729	2,093	2,539	2,861	3,579	3,883	4,897
20	0,687	1,325	1,725	2,086	2,528	2,845	3,552	3,850	4,837
21	0,686	1,323	1,721	2,080	2,518	2,831	3,527	3,819	4,784
22	0,686	1,321	1,717	2,07	2,508	2,819	3,505	3,792	4,736
23	0,685	1,319	1,714	2,069	2,500	2,807	3,485	3,767	4,693
24	0,685	1,318	1,711	2,06	2,492	2,797	3,467	3,745	4,654
25	0,684	1,316	1,708	2,060	2,485	2,787	3,450	3,725	4,619
26	0,684	1,315	1,706	2,05	2,479	2,779	3,435	3,707	4,587
27	0,684	1,314	1,703	2,052	2,472	2,771	3,421	3,690	4,558
28	0,683	1,313	1,701	2,04	2,467	2,763	3,408	3,674	4,530
29	0,683	1,311	1,699	2,045	2,462	2,756	3,396	3,659	4,506
30	0,683	1,310	1,697	2,042	2,457	2,750	3,385	3,646	4,482
35	0,682	1,306	1,690	2,030	2,438	2,724	3,340	3,491	4,389
40	0,681	1,303	1,684	2,021	2,423	2,704	3,307	3,551	4,321
45	0,680	1,301	1,679	2,01	2,412	2,690	3,281	3,520	4,269
50	0,679	1,299	1,676	2,009	2,403	2,678	3,261	3,496	4,228

60	0,679	1,296	1,671	2,00	2,390	2,660	3,232	3,460	4,169
70	0,678	1,294	1,667	1,994	2,381	2,648	3,211	3,435	4,127
80	0,678	1,292	1,664	1,990	2,374	2,639	3,195	3,416	4,096
90	0,677	1,291	1,662	1,987	2,368	2,632	3,183	3,402	4,072
100	0,677	1,290	1,660	1,984	2,364	2,626	3,174	3,390	4,053
120	0,677	1,289	1,658	1,980	2,358	2,617	3,160	3,373	4,025
200	0,676	1,286	1,653	1,972	2,345	2,601	3,131	3,310	3,970
500	0,675	1,283	1,648	1,965	2,334	2,586	3,107	3,310	3,922
1000	0,675	1,282	1,646	1,962	2,330	2,581	3,098	3,300	3,906
oo	0,675	1,282	1,645	1,960	2,326	2,576	3,090	3,290	3,891
$\frac{n}{a}$	0,25	0,10	0,05	0,025	0,01	0,005	0,001	0,0005	0,00005
<i>Nivel de semnificație pentru testul unilateral</i>									

Anexa 7 - Valorile critice pentru testul χ^2 corespunzătoare diferitelor praguri de semnificație α și numărului "n" al gradelor de libertate

$n \backslash \alpha$	0,990	0,975	0,950	0,900	0,100	0,050	0,025	0,010	0,001
1	0,0002	0,0010	0,0039	0,0198	2,71	3,84	5,02	6,63	10,83
2	0,02	0,05	0,10	0,21	4,61	5,99	7,38	9,21	13,82
3	0,12	0,22	0,35	0,58	6,25	7,81	9,35	11,34	16,27
4	0,30	0,48	0,71	1,06	7,78	9,49	11,14	13,28	18,47
5	0,55	0,83	1,15	1,61	9,21	11,07	12,83	15,09	20,62
8	0,87	1,24	1,64	2,20	10,64	12,59	14,45	16,81	22,46
7	1,24	1,69	2,17	2,83	12,02	14,07	16,01	18,47	24,32
8	1,65	2,18	2,73	3,49	13,36	15,51	17,53	20,09	26,13
9	2,09	2,70	2,33	4,47	14,68	16,92	19,02	21,67	27,88
10	2,56	3,25	3,94	4,87	15,99	18,31	20,48	23,21	29,59
11	3,05	3,82	4,57	5,58	17,27	19,67	21,92	24,72	31,26
12	3,57	4,40	5,23	6,30	18,55	21,03	23,34	26,22	33,91
13	4,11	5,01	5,89	7,04	19,81	22,36	24,71	27,89	34,53
14	4,66	5,63	6,57	7,79	21,06	23,67	26,12	29,14	36,12
15	5,23	6,26	7,26	8,55	22,31	25,00	27,49	30,58	37,70
16	5,81	6,91	7,96	9,31	23,54	26,30	28,84	32,00	39,25
17	6,41	7,56	8,67	10,08	24,77	27,59	30,19	33,41	40,79
18	7,01	8,23	9,39	10,86	25,99	28,87	31,53	34,80	42,31
19	7,63	8,91	10,12	11,65	27,20	30,14	32,85	36,19	43,82
20	8,26	9,59	10,85	12,44	28,41	31,41	34,17	37,57	45,32
21	8,90	10,28	11,59	13,24	29,61	32,67	35,48	38,93	46,80
22	9,54	10,98	12,34	14,04	30,81	33,92	36,78	40,29	48,27
23	10,20	11,69	13,09	14,85	32,01	35,17	38,08	41,64	49,73
24	10,86	12,40	13,85	15,66	33,20	36,41	39,37	42,98	51,18
25	11,52	13,12	14,61	16,47	34,38	37,65	40,65	44,31	52,62

26	12,20	13,84	15,38	17,29	.35,56	38,88	41,92	45,64	54,05
27	12,88	14,57	16,15	18,11	38,74	40,11	43,19	46,96	55,48
28	13,57	15,31	16,93	18,91	37,92	41,34	44,46	48,28	56,89
29	14,26	10,65	17,71	19,77	39,09	42,56	43,72	49,59	58,30
30	14,95	16,79	18,19	20,60	40,26	43,77	46,98	50,89	59,70

Anexa 8 – Soluțiile exercițiilor propuse în anexe

Soluții: Anexa 1 - Statistică descriptivă, exerciții recapitulative

1. Media : 88
Amplitudinea : 41
Amplitudinea relativă: 46,59 %
Dispersia: 219,3333
Deviația standard: $\pm 14,809$
Coeficientul de variație: 16,82%

2. Media : 3000
Amplitudinea : 400
Amplitudinea relativă: 13,33%
Dispersia: 12121,2121
Deviația standard: 110,096
Coeficientul de variație: 3,67%

3. Media : 29,84
Amplitudinea : 20
Amplitudinea relativă: 67,024 %
Dispersia: 13,48180905
Deviația standard: 3,671
Coeficientul de variație: 12,305%

Soluții: Anexa 2 - Teste statistice de semnificație, exerciții recapitulative

Testul Student

1. $p = 0,313$ - nu diferă semnificativ
2. $p = 0,0095$ - diferă semnificativ
3. $p = 0,135$ - nu diferă

4. $p = 0,000000000011$ - diferă semnificativ

5. $p = 0,0087$ - diferă semnificativ

6. $p = 0,060$ - nu diferă

7. $p = 0,0008$ - diferă semnificativ

Testul χ^2

1. $p = 0,560$ - nu există o legătură

2. $p = 0,674$ - nu există o legătură

3. $p = 0,252 \times 10^{-10}$ - există o legătură

4. $p = 0,032$ - există o legătură

5. $p = 0,000124$ - există o legătură

6. $p = 0,0705$ - nu există o legătură

7. $p = 0,00056$ - există o legătură

8. $p = 0,0000077$ - există o legătură

9. $p = 0,0301$ - există o legătură

10. $p = 0,0896$ - nu există o legătură

11. $p = 0,0048$ - există o legătură

Soluții: Anexa 3 - Corelație și regresie, exerciții recapitulative

1. $p = -0,979$ - corelate

2. $p = -0,949$ - corelate

3. $p = 0,901$ - corelate

4. $p = -0,092$ - necorelate

- 5. $p = 0,985$ - corelate
- 6. $p = - 0,0128$ - necorelate
- 7. $p = 0,900$ - corelate
- 8. $p = 0,915$ - corelate
- 9. $p = 0,994$ - corelate

Bibliografie selectivă

1. *** - *NIST/SEMATECH e-Handbook of Statistical Methods*, 2005, National Institute of Standards and Technology, <http://www.itl.nist.gov/div898/handbook/>
2. *** - *Farmacopeea Română*, ediția a-X-a, Editura Medicală București 1993
3. *** - *How Grubbs' test works*, articol disponibil online pe website-ul firmei GraphPad Software, http://www.graphpad.com/library/BiostatsSpecial/article_39.htm
4. *** - *Introduction to Epi principles*, Innovations in Health Education USA website, copyright Henry Ford Health System, 2004, carte disponibilă online pe website-ul IIHE, <http://www.iihe.org/education/lectures/epidemiology/default.htm>
5. *** - *Normality tests – use with caution*, articol disponibil online pe website-ul firmei GraphPad Software, http://www.graphpad.com/library/BiostatsSpecial/article_197.htm
6. Baker R. J. - *Basic principles of statistical analysis*, carte disponibilă online pe website-ul University of Saskatchewan, Canada, <http://homepage.usask.ca/~rjb609/stats.html>
7. Baron T. et al. – *Statistică teoretică și economică*, Editura Didactică și Pedagogică, București, 1996
8. Dallal G.E. - *The Little Handbook of Statistical Practice*, 2001, carte disponibilă online pe website-ul Tufts University, USA <http://www.tufts.edu/~gdallal/LHSP.HTM>

9. Douglas G. A. - *Practical Statistics for Medical Research*, CRC Press, 1990
10. Hopkins W. G. - *A New View of Statistics*, 2004, carte disponibilă online pe website-ul <http://www.sportsci.org/resource/stats/index.html>
11. Mărușteri M. – *Biostatistică - aplicații practice și exerciții recapitulative pentru studenții Școlii Doctorale*, curs online disponibil pe situl UMF Târgu Mureș, http://www.umftgm.ro/statdoct/biostatistica_lp.pdf
12. Mărușteri M. – *Biostatistică - note de curs pentru studenții Școlii Doctorale* - curs online disponibil pe situl UMF Târgu Mureș, http://www.umftgm.ro/statdoct/biostatistica_curs.pdf
13. Mărușteri M. – *Noțiuni de biostatistică*, în Maria T. Dogaru (sub redacția) – *Farmacologie experimentală*, Litografia UMF Târgu Mureș, 1998
14. McCurdy St., Patrick R. - *Clinical Epidemiology and Study Design*, 2006, carte disponibilă online pe website-ul UC Davis School of Medicine, USA, <http://som.ucdavis.edu/students/k30/folder.2004-06-21.4913658468/>
15. Motulsky H. - *InStat guide to choosing and interpreting statistical tests*, GraphPad Software, 2004, carte disponibilă online pe website-ul firmei GraphPad Software <http://www.graphpad.com/Downloads/InStat3.pdf>
16. Motulsky H. – *Intuitive Biostatistics*, Oxford University Press, 1995
17. Motulsky H., Christopoulos A. - *Fitting Models to Biological Data Using Linear and Nonlinear Regression*, Oxford University Press, 2004,

- carte disponibilă online pe website-ul firmei GraphPad Software
<http://www.graphpad.com/manuals/prism4/RegressionBook.pdf>
18. Saporta G., Ștefănescu Viorica – *Analiza datelor & informatică*, Editura Economică, București, 1996
19. Simionovici M., Cârstea Al., Vlădescu C. – *Cercetarea farmacologică și prospectarea medicamentelor*, Editura Medicală, București, 1983
20. Stockburger D. W. - *Introductory Statistics: Concepts, Models, And Applications*, 1996, carte disponibilă online pe website-ul Missouri State University USA, <http://www.psychstat.missouristate.edu/sbk00.htm>
21. Swinscow T.D.V. - *Statistics at Square One*, BMJ Publishing Group, 1997, carte disponibilă online pe website-ul
<http://bmj.bmjournals.com/collections/statsbk/index.shtml>
22. Tilincă Mariana, Mărușteri M., Brânzaniuc Klara – *Studiu privind efectul radiațiilor ionizante accidentale asupra populației județului Mureș*, Revista de Medicina si Farmacie Targu Mures, vol. 52/2006, pp. 117-122, ISSN 1221-2229
23. Varkevisser C. M., Pathmanathan Indra, Brownlee Ann - *Choosing A Significance Test*, în *Designing And Conducting Health Systems Research Projects: Volume 2*, 2003, carte disponibilă online pe website-ul International Development Center, Canada, http://www.idrc.ca/en/ev-33013-201-1-DO_TOPIC.html
24. Zar J. H. - *Biostatistical Analysis (4th Edition)*, Prentice Hall, 1998

ISBN (10) 973-7665-11-2
ISBN (13) 978-973-7665-11-9