UNIVERSITATEA DE MEDICINĂ ȘI FARMACIE "IULIU HAȚIEGANU" CLUJ-NAPOCA

Curs de Biostatistică Medicală

2018

Facultatea de Medicină

AUTORI: Tudor Drugan, Sorana D. Bolboacă, Daniel Leucuța, Cosmina Bondor, Tudor Călinici, Mădălina Văleanu, Horațiu Colosi, Mihaela Iancu, Dan Istrate

UMF
UNIVERSITATEA DE MEDICINĂ ŞI FARMACIE
IULIU HAŢIEGANU
CLUJ-NAPOCA

Universitatea de Medicină şi Farmacie
“Iuliu Haţieganu” Cluj-Napoca
Facultatea de Medicină

Curs de Biostatistică Medicală

ediția a doua revizuită și adăugită

Tudor Drugan
Daniel Leucuța
Tudor Călinici
Horațiu Colosi
Dan Istrate
Sorana D. Bolboacă
Cosmina Bondor
Mădălina Văleanu
Mihaela Iancu

Editura Medicală Universitară "Iuliu Haţieganu"
Cluj-Napoca
2018

Editura Medicală Universitară "Iuliu Haţieganu"

Cluj-Napoca

Curs de biostatistică medicală

Descrierea CIP a Bibliotecii Naţionale a României
Curs de biostatistică medicală / Tudor Drugan, Sorana D. Bolboacă, Daniel Leucuţa, ... - Ed. a 2-a, reviz. şi adăug.. - Cluj-Napoca : Editura Medicală Universitară "Iuliu Haţieganu", 2018
Conţine bibliografie
ISBN 978-973-693-861-0

I. Drugan, Tudor
II. Bolboacă, Sorana
III. Leucuţa, Daniel-Corneliu

57

Editura Medicală Universitară "Iuliu Haţieganu", Cluj-Napoca, tel. +40264596089 Universitatea de Medicină şi Farmacie "Iuliu Haţieganu", Cluj-Napoca,
400023, str. Victor Babeş nr. 8, tel. +40264597256
Coperta şi tiparul executat la S.C. Cartea Ardeleană S.R.L. Cluj-Napoca,
400210, str. Mecanicilor nr. 48, tel. +40364117246

PRINTED IN ROMÂNIA

Universitatea de Medicină şi Farmacie
“Iuliu Haţieganu” Cluj-Napoca
Facultatea de Medicină

Curs de Biostatistică Medicală

ediția a doua revizuită și adăugită

Tudor Drugan
Daniel Leucuța
Tudor Călinici
Horațiu Colosi
Dan Istrate
Sorana D. Bolboacă
Cosmina Bondor
Mădălina Văleanu
Mihaela Iancu

Editura Medicală Universitară "Iuliu Haţieganu"
Cluj-Napoca
2018

Mentorului nostru,

Prof. Dr. Ștefan Țigan

Cuprins

Prefață

"No amount of observations of white swans can allow the inference that all swans are white, but the observation of a single black swan is sufficient to refute that conclusion." - David Hume

Cursul de biostatistică medicală se adresează atât studenților Facultății de medicină, cât și medicilor care doresc să-și readucă aminte principalele noțiuni necesare pentru înțelegerea literaturii științifice medicale. Această carte pornește de la ideea că cititorul fără abilități matematice nu are cunoștințe anterioare de statistică și încearcă să introducă treptat noțiunile prezentate, pornind de la cele mai simple, până la cele care țin de matematicile superioare.

În cei peste 20 de ani de predare a acestei discipline, am constatat că interesul stârnit de statistică în rândul studenților și medicilor variază de la indiferență până la o profundă antipatie, astfel încât în această prefață o să încerc să ofer un răspuns la întrebările „De ce? De ce trebuie să cunoaștem statistica dacă suntem medici sau viitori medici? Nu am scăpat oare de matematică?" Aş vrea să ofer câteva argumente în favoarea biostatisticii, cu speranța că parcurgerea acestui curs va oferi și alte răspunsuri.

Francis Bacon (1561–1626) considera că fenomenele naturale pot fi observate, iar după un număr finit de observații repetate se poate generaliza o lege care să le descrie. Pornind de la această idee, un alt filozof al cunoașterii științifice, David Hume (1711–1776), a formulat următorul paradox: urmărind cum aterizează lebedele pe un lac, după un număr suficient de mare de observații, se poate generaliza faptul că toate lebedele sunt albe. La vremea aceea, această lege putea fi considerată universal valabilă, deoarece abia atunci un marinar adusese în Europa prima lebădă neagră, provenind din Noua Zeelandă. David Hume a enunţat continuarea paradoxului: ce se întâmplă dacă pe lac aterizează o lebădă neagră, ce mai rămâne din adevărul științific dedus anterior? Cum putem interpreta valoarea de adevăr în experimente în care rezultatul nu este întotdeauna același? Determinismul cauzal se aplică numai câteodată, sau fenomenele pot fi atât de complexe încât să nu percepem în totalitate mecanismele care le declanșează?

Privit din această perspectivă, adevărul științific nu mai are valoare absolută, nu mai suntem în Grecia antică să ne putem baza pe teoremele „provenite de la zei", iar fenomenele complexe dobândesc o valoare de adevăr temporară și statistică. Medicina s-a transformat dintr-o enciclopedie de reguli moștenite prin tradiția

Hipocratică sau enunţate în tratate, într-un ansamblu dinamic de cunoștințe aflate în continuă schimbare, într-o măsură atât de mare, încât aproape o zecime din noțiuni sunt perimate în fiecare an, cu toate că volumul literaturii științifice medicale crește cu 10% anual. În societatea contemporană, rutina medicală se bazează tot mai mult pe referințe științifice moderne, provenite din cercetare și tot mai puțin pe tradiționalele tratate și manuale.

Aproape fiecare gest al medicului ajunge să facă referire la elemente de statistică, asocierea simptomatologiei cu diagnosticul este determinată de probabilități, eficacitatea terapeutică este validată de teste statistice, subtila legătură dintre factorii de risc și îmbolnăviri este o măsură matematică... Cursul de biostatistică medicală încearcă să familiarizeze studentul cu noțiunea de adevăr validat statistic, etapă importantă în dezvoltarea gândirii medicale.

Înțelegerea profundă a naturii adevărului științific permite o abordare mai deferentă a actului medical. Prin fiecare capitol din această carte vom încerca să demonstrăm cele expuse mai sus, deoarece credem că dobândirea unui sistem de gândire probabilistic, bazat pe înțelegerea adevărului statistic, este în măsură să asigure o atitudine mai corectă în exercitarea profesiei medicale.

Față de prima ediție, în care fiecare autor a redactat un capitol, în această ediție, tot colectivul de autori a contribuit la modificarea și îmbunătățirea conținutului acestei lucrări, astfel încât nu mai putem reține atribuirea capitolelor individuale unui singur autor.

Cluj-Napoca,
8 octombrie 2018 Prof.Dr. Tudor Drugan

1. Noțiunile fundamentale ale statisticii

Obiective educaționale

După parcurgerea acestui capitol, studenții vor cunoaște noțiuni fundamentale ale statisticii: populație statistică, eșantion, eșantion reprezentativ, talie, eroare sistematică (bias), variabilă, date statistice.

Statistica este un domeniu ştiinţific care permite studiul parametrilor a căror proprietate este variabilitatea. Practic se studiază colecţii de observaţii efectuate asupra unor entităţi de aceeaşi natură, după una sau mai multe caracteristici variabile.

Domeniile de aplicabilitate ale statisticii sunt numeroase, dintre care se pot aminti: medicină, biologie, afaceri, marketing, economie, industrie, agricultură, învăţământ, sociologie.

1.1 Populația statistică

Statistica vizează studiul unui anumit număr de "obiecte" având aceeaşi natură. Pentru această colecţie de obiecte uneori se utilizează termenul de *mulţime statistică*.

În cele ce urmează, vom distinge două tipuri de mulţimi statistice:

- *populaţia statistică*
- *eşantionul* sau *selecţia*.

Populaţia statistică este o mulţime de elemente (obiecte sau subiecţi) care au anumite însuşiri (atribute sau caractere) comune, care formează obiectul unei analize statistice.

Numărul elementelor populaţiei se numeşte *volumul sau talia populaţiei.* Adesea volumul sau talia populaţiilor statistice este foarte mare sau chiar infinit.

Exemple:

O populaţie statistică poate fi:

- un grup de fiinţe umane (populaţia şcolară dintr-un oraş la un moment dat, populaţia vârstnică (peste 65 ani) dintr-o regiune la un moment dat);
- mulţime de obiecte (totalitatea fiolelor de penicilină produse de o anumită fabrică de medicamente);
- un grup de fenomene sau evenimente.

Un exemplu din domeniul biomedical este cel al unei populaţii de măsurători, care rezultă în urma măsurării repetate a unei mărimi (de exemplu, tensiunea arterială măsurată la unul sau mai mulţi indivizi în diferite momente).

Elementele unei populaţii statistice se numesc *unităţi statistice sau indivizi* ai populaţiei statistice.

În definirea populaţiilor statistice care intervin în studiile medicale trebuie stabilite cu claritate:

- *criteriile de includere* (condiţiile în care o entitate este ales pentru a deveni un element al populaţiei);
- *criteriile de excludere* (condiţiile în care o entitate care deși are criteriile de includere satisfăcute, se va exclude din mulțime).

1.2 Eșantionul

În general, cercetătorul nu studiază un caracter al populaţiei pe întreaga mulţime de elemente, din mai multe motive, dintre care menţionăm următoarele:

- Talia populaţiei poate fi foarte mare sau chiar infinită ceea ce face imposibilă o "observare" exhaustivă a întregii populaţii.
- Eşantioanele pot fi studiate mai rapid decât populaţiile, acesta fiind un motiv foarte important atunci când, de exemplu, se doreşte un răspuns rapid la o problemă.
- Studiul caracterului pe întreaga populaţie este frecvent imposibil, deoarece poate distruge populaţia. De exemplu, dacă se doreşte să se studieze durata medie de valabilitate a fiolelor fabricate de către o firmă farmaceutică, operând asupra tuturor fiolelor pentru verificarea lor, se distruge toată populaţia.
- În anumite situaţii nu se mai pot obţine informaţii decât despre o parte a populaţiei.
- Rezultatele observaţiilor pe eşantioane adesea sunt mai precise decât rezultatele bazate pe observarea populaţiei în totalitate, deoarece la nivelul unui eşantion se controlează mai uşor procesul şi tehnicile de observare, acestea menţinându-se cu un efort mai mic în standardele de eroare acceptate.
- Costul şi resursele necesare (umane, materiale etc.) pentru observarea exhaustivă a unei populaţii pot de asemenea să fie un motiv pentru utilizarea eşantioanelor.
- Uneori nu este etic să supunem riscului toată populația de subiecți – spre exemplu pentru evaluarea unor intervenții terapeutice.

Acestea sunt câteva raţiuni pentru care o populaţie este studiată cu ajutorul unei submulţimi a ei de talie mai mică, apărând necesitatea de a preleva numai o parte din această populaţie, submulţime denumită *eşantion*.

Prin *eşantion* (sau *selecţie*) extras dintr-o populaţie statistică se înţelege o submulţime finită a ei. Numărul elementelor submulţimii se numeşte *efectivul (talia sau volumul) selecţiei.*

Eşantionul, fiind, în general, mult mai mic decât populaţia din care provine (populaţie ţintă), el poate fi observat în totalitate, iar dacă este reprezentativ, se pot cunoaşte prin intermediul lui anumite proprietăţi ale populaţiei.

Un eşantion bun trebuie să constituie o imagine redusă cât mai adecvată şi fidelă a întregii populaţii pentru care se doreşte studierea unui caracter anume. În caz contrar, se spune că eşantionul este nereprezentativ (sau cu o eroare sistematică = "bias"). Alegerea eşantionului şi culegerea datelor necesare studiului propus constituie partea cea mai lungă şi mai laborioasă a multor studii. În scopul generalizării sau extrapolării la întreaga populaţie a rezultatelor obţinute pe eşantion (care este obiectivul statisticii inductive) este de dorit ca acesta să reprezinte cât mai bine posibil populaţia vizată.

Pentru ca un eşantion să fie *reprezentativ* pentru populaţia din care este extras, el trebuie să satisfacă două condiţii principale:

- condiţie de ordin *cantitativ: talia* sau *efectivul* eşantionului trebuie să fie suficient de mare;
- condiţie de ordin *calitativ:* eşantionul trebuie *extras aleator* (sau întâmplător) din populaţie.

1.3 Variabilele

Trăsătura comună a tuturor unităţilor unei populaţii ***M,*** care poate fi luată în considerare într-o analiză statistică poartă numele de caracteristică variabilă, pe scurt variabilă. Analiza statistică a unei populaţii se poate face după una sau mai multe variabile.

Exemplul 1.1.

Să presupunem că se dorește studierea numărului de leucocite la bolnavii internaţi într-un spital de boli infecţioase.

- Mulţimea bolnavilor internaţi într-o anumită perioadă formează o populaţie statistică;
- Fiecare bolnav este o unitate statistică;
- Numărul de leucocite ale bolnavului la internare este variabila studiată;
- Un eşantion din această populaţie statistică poate fi, de exemplu, mulţimea alcătuită din 50 de bolnavi internaţi, luaţi din trei în trei în ordinea internării.

Exemplul 1.2.

Să presupunem că se dorește studierea numărului pacienţilor consultaţi zilnic în cabinetele medicale dintr-o anumită zonă. Atunci:

- mulţimea cabinetelor reprezintă o populaţie statistică,
- fiecare cabinet este o unitate statistică,
- numărul de pacienţi consultaţi zilnic reprezintă o variabilă.

Precizarea exactă a populaţiei statistice prin stabilirea criteriilor de includere şi excludere este foarte importantă în proiectarea studiilor medicale.

Mulţimea de valori pe care o caracteristică le poate lua pentru fiecare unitate sau individ al unei populaţii statistice (sau eşantion), se numeşte variabilă definită pe populaţia statistică (sau pe un eşantion).

În realitate, variabila este o funcţie X: M $\rightarrow$ C , unde M este populaţia statistică, iar C este o mulţime în care caracteristica ia valori.

Variabilele statistice pot fi de două tipuri:

- calitative (atribut): asociate unor caracteristici care nu pot fi măsurate;
- cantitative: asociate unor caracteristici susceptibile de a fi măsurate.

Când variabilele statistice sunt de natură cantitativă, atunci mulţimea C este o mulţime de numere reale sau întregi, iar în caz că sunt de natură calitativă, C poate fi de regulă o mulţime finită, conţinând nivelurile calitative posibile ale caracteristicii.

1.3.1 Variabile calitative

Variabilele de tip calitativ sau atribut sunt nemăsurabile, finite, iar calculul mediei valorilor nu are sens. Este importantă definirea numărului şi tipurilor de clase pentru aceste variabile, adică a numărului de valori diferite pe care le pot lua.

Exemplul 1.3.

- sex – două clase: masculin/feminin
- starea de sănătate – trei clase: precară, bună, foarte bună

Variabilelor calitative pot fi clasificate în:

- variabile de tip nominal – subiecţii nu pot fi grupați în categorii ce au o anumită ordine (exemplu: rasa umană = rasa europeană (caucaziană),rasa mongoloidă și rasa australo-negroidă);
- nominale ordonate – grupează subiecţii în categorii ce pot fi ordonate (exemplu: grupa de vârstă = preșcolar, școlar, adolescent, adult, vârstnic);
- dicotomiale – subiecţii sunt întotdeauna grupaţi doar în două categorii (exemplu: fumător şi nefumător). Variabilele dicotomiale reprezintă un caz particular de variabile nominale.

1.3.2 Variabile cantitative

Variabile cantitative sunt asociate unor caracteristici ce pot fi măsurate, cuantificate. Ele pot fi:

- Variabile continue sunt asociate unor caracteristici măsurabile ce pot lua orice valoare dintr-un anumit interval. Calculul mediei are întotdeauna semnificaţie (exemplu: tensiunea arteriala, greutatea la naștere, perimetrul cranian, glicemia, valorile eritropoietinei la un grup de ciclişti etc.)

- Variabilele discrete sau discontinue iau valori numere întregi. Calculul mediei nu are întotdeauna semnificaţie (exemplu: numărul sarcinilor, numărul zilelor de spitalizare).

Transformarea variabilelor cantitative în variabile calitative este realizabilă însă cu pierdere de informație (exemplu: substituirea variabilei continue "nota la examen" cu o variabilă calitativă "calificativ"); Nu este posibilă transformarea variabilelor calitative în cantitative, chiar dacă codificarea lor este numerică (exemplu: stadiile tumorilor).

Observaţii:

Este mai simplu de lucrat cu variabile discrete decât cu variabile continue. Astfel, frecvent o variabilă continuă este transformată într-o variabilă discretă printr-un proces de *discretizare* sau *grupare în clase*. Această discretizare este cauzată şi de precizia aparatului de măsură utilizat, care transformă o variabilă continuă într-una discretă.

Variabilele de supravieţuire, ce intervin în anumite studii medicale, corespund timpului scurs între includerea unui subiect într-un studiu şi apariţia unui eveniment predefinit al studiului (exemplu: deces, metastază, complicaţie, simptom, semn). Aceste variabile sunt de tip cantitativ.

1.4 Date statistice

Legat de conceptul de variabilă definită pe o populaţie statistică, apare noţiunea de *date statistice*. Datele statistice sunt valori ale unei variabile (caracteristici) luate pe elementele unor eşantioane ale populaţiei.

Datele statistice care apar în domeniul medical au diverse provenienţe.

Un prim tip este cel al *datelor obţinute din măsurători*, care rezultă pe baza unor determinări cantitative ale unor proprietăţi susceptibile să varieze, în principiu de o manieră continuă, cum ar fi, spre exemplu, înălţimea, greutatea, presiunea sangvină, glicemia.

Alte date statistice rezultă din enumerarea indivizilor, operaţie care furnizează în mod necesar date întregi. Aceste *date de enumerare* se obţin de regulă ca fiind numărul de indivizi ai unor grupe, stabilite în urma unor operaţii de clasificare după anumite criterii.

Adesea, rezultatele de acest gen se exprimă şi sub forma de procente: în sângele unui anumit pacient s-au numărat 65,5% polinucleare, 8,2 % monocite şi 17,3% limfocite.

O altă categorie de date sunt *datele de înseriere (ordinale sau de ordonare)*, care reprezintă poziţia unor obiecte sau indivizi într-un "clasament" stabilit după anumite criterii.

Datele de ordonare sunt frecvent utilizate, de exemplu, în anumite studii de psihologie experimentală şi în particular, în cele privind educaţia. În domeniul medical, un exemplu de astfel de date îl constituie stadiile unei boli.

Clasificarea datelor statistice poate fi realizată ţinând seama de scalele de măsură utilizate. Astfel, se disting următoarele scale de măsură: scala nominală, scala ordinală și scala interval.

Scala *nominală* este o scală pentru variabilele calitative care pot lua un număr finit de valori, nu au nici o proprietate aritmetică, iar dacă admit o ordonare a valorilor aceasta nu are semnificație medicală.

Datele evaluate după o scală nominală sunt numite *observaţii calitative*, deoarece ele descriu o calitate a unei persoane sau obiect studiat. Unele dintre aceste scale au doar două valori şi atunci observaţiile sunt binare.

Multe dintre clasificările din domeniul medical sunt evaluate pe o scală nominală cum ar fi: rezultatul unui tratament medical, expunerea la un factor.

Exemplul 1.4.

- Tipul rasial: alb, negru, alt tip
- Sexul : bărbaţi, femei
- Clasificarea anemiilor: anemie microcitică (incluzând deficienţă de fier), anemie megaloblastică (incluzând deficienţa vitaminei B_{12}), anemie normocitică (adesea asociată cu o boală cronică).

Datele calitative sau nominale sunt prezentate de obicei cu ajutorul procentelor sau proporţiilor ce sunt incluse în tabele şi sunt reprezentate grafic prin diagrame cu bare sau coloane.

Scala *ordinală* este o scală utilizată în cazul variabilelor care pot lua valori într-o mulţime discretă finită de valori, care nu au nicio proprietate aritmetică, dar care însă posedă o anumită ordonare a acestor valori.

Exemplul 1.5.

- Stadiile unei boli: hipertensiune arterială în stadiul I, II, III.
- Severitatea crampelor la stomac: uşoară, moderată, puternică.
- Stadiile unei tumori (în funcţie de gradul de dezvoltare): carcinom: 0, I, II, III, IV, tumori colorectale: 1, 2, 3, 4.
- Scorul Apgar care descrie gradul de dezvoltare a unui nou născut utilizează o scală cu valori de la 0 la 10, scorurile înalte indicând o bună funcţionare cardiorespiratorie şi neurologică.

Deşi există o ordonare a diferitelor categorii, diferenţa între două categorii adiacente, respectiv lungimea intervalului dintre ele, nu are totdeauna aceeaşi semnificaţie sau valoare oriunde pe scală.

Scala *interval* este o scală utilizată în cazul variabilelor cantitative continue (ce pot lua valori într-un interval) şi pentru care diferenţa între două valori ale scalei are sens.

Exemplul 1.6.

Temperatura în grade Celsius în 5 zile consecutive:

Ziua	Luni	Marți	Miercuri	Joi	Vineri
Temperatura	20^0	25^0	18^0	24^0	20^0

Timpul (momentul).

Scala de tip *raţie sau raport* este utilizată în cazul variabilelor cantitative continue pentru care atât diferenţa cât şi câtul a oricăror două valori de pe scală au sens. Această scală are un zero absolut şi nu acceptă valori negative.

Exemplul 1.7.

- Greutatea în kilograme a cinci indivizi: 62, 61, 72, 71, 67;
- Înălţimea;
- Vârsta.

Scala *discretă* este folosită atunci când observaţiile numerice pot lua doar valori discrete.

Exemplul 1.8.

- Numărul de operaţii anterioare;
- Numărul de recidive;
- Numărul de factori de risc prezenţi la o persoană.

2. Statistica descriptivă

Obiective educaționale. După parcurgerea acestui capitol, studenții:

- Vor putea să recunoască simbolurile statisticii descriptive
- Vor putea calcula statisticile descriptive
- Vor putea descrie variabilele folosind măsuri ale tendinţei centrale, de localizare şi dispersie
- Vor identifica corect parametrii statistici care au sens să fie utilizați în funcţie de tipul variabilei
- Vor putea construi tabele de frecvenţă şi tabele de contingenţă
- Vor putea reprezenta corect grafic datele calitative şi cantitative

Primul pas în evaluarea statistică a datelor medicale este reprezentat de descrierea acestora într-o formă concisă care să reflecte corect datele brute (ex. valorile pentru fiecare variabilă colectate de la fiecare subiect inclus în studiu – vezi Exemplul 2.1). Dacă studiu s-a realizat pe un eşantion mic, acest lucru se poate face prin vizualizarea datelor brute. Această procedură nu se poate însă utilizabilă în cazul în care studiul este realizat pe un eșantion mare (ex. 200 subiecţi).

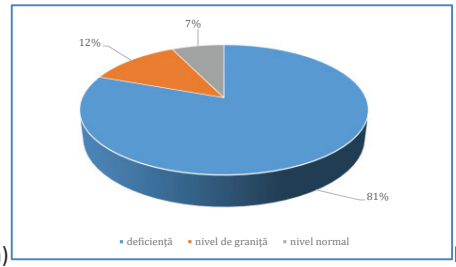
Exemplul 2.1.

Mahdavi și colaboratorii au evaluat nivelul stresului oxidativ şi al vitaminei C la 57 pacienți cu cancer și 22 subiecți indemni (subiecţi care nu prezentau nici o formă de cancer). Subiecții incluși în studiu au avut vârste cuprinse între 18 și 80 ani. Subiecții cu cancer au fost subdivizați în funcție de localizarea tumorii (cap și gât, gastro-intestinal, plămân). La toți subiecții incluși în studiu au fost determinate valorile sangvine pentru malondialdehida (MDA), statusul antioxidant total (TAS) și vitamina C.

Ce putem face cu datele brute colectate?

Primul pas în abordarea problemei este reprezentarea grafică a datelor. Pentru exemplificare se vor utiliza informaţiile cu privire la nivelul sangvin al vitaminei C (Figura 2.1).

Figura 2.1a este realizată doar pe datele subiecţilor din grupul caz şi arată că nivelul sangvin al vitaminei C la pacienţii cu cancer este sub nivelul normal în majoritatea cazurilor. Compararea între valorile serice ale vitaminei C la subiecţii cu cancer şi cei din grupul martor (Figura 2.1b) sugerează existenţa unei diferenţe aritmetice.

a) 

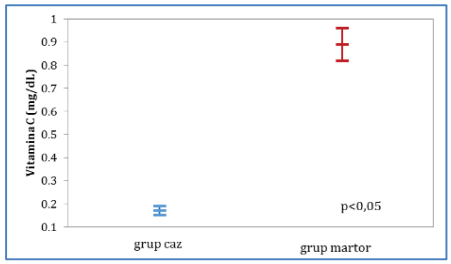
b) 

Fig. 2.1. Evaluarea nivelului sanguin al vitaminei C la pacienții cu cancer. a) Distribuția în funcție de nivelul deficienței; b) Nivele medii și intervale de încredere de 95%

Există diferite metode de prezentare a unei a datelor care se realizează cu ajutorul parametrilor statistici descriptivi, a tabelelor sau a reprezentărilor grafice. Acest capitol va introduce cele mai utilizate metode şi va prezenta avantajele şi dezavantajele acestora.

2.1 Simboluri şi abrevieri utilizate în statistica descriptivă

În statistica descriptivă se utilizează o serie de simboluri matematice şi statistice care trebuie cunoscute:

- Funcţii matematice:
  - adunarea ('+'), scăderea ('-'), înmulțirea ('*'sau '×'), împărțirea ('/')
  - suma (∑):
    - $\sum X_i$: Suma numerelor variabilei X; adună toate valorile variabilei X (i = 1, ..., n; unde n = volumul eșantionului)
    - $\sum X_i^2$: Suma pătratelor numerelor variabilei X; se ridică la pătrat fiecare valoare a variabilei X după care se adună toate aceste valori
    - $(\sum X_i)^2$: Suma numerelor variabilei X ridicată la pătrat; se adună toate valorile variabilei X după care se ridică la pătrat
  - simboluri de clasificare: mai mic (nu include numărul care este după simbol, '<'), mai mic sau egal (include numărul care este după simbol, '≤'), mai mare ('>'), mai mare sau egal ('≥')
- Simboluri utilizate în statistica descriptivă:
  - N = volumul populaţiei; numărul elementelor populaţiei
  - n = volumul sau talia eşantionului; numărul elementelor eşantionului
  - f = frecvenţa observată; fa = frecvenţa absolută; fr = frecvenţa relativă
  - μ = media populaţiei
  - $\overline{X}$ = media eşantionului
  - s^2 = variaţia eșantionului
  - S^2 = varianța sau variația de eșantionare
  - s = ecartul tip, sau deviaţia standard, sau abaterea standard
  - σ = ecartul tip sau deviaţia standard în populație

- S = ecartul tip sau deviația standard de eșantionare
- Q1 = cvartila 1 / percentila 25%
- Q3 = cvartile 3 / percentila 75%
- CV = coeficientul de variație

2.2 Serii statistice

Seria statistică care conține valorile caracteristicilor de interes (observate sau măsurate) pentru fiecare entitate din populație sau eșantion constituie seria de date statistice brute. O parte din datele brute posibile pentru scenariul din Exemplul 2.1 sunt prezentate în Tabelul 2.1.

Tabelul 2.1. Date brute pentru Exemplul 2.1

Grup	Vitamina C (mg/dl)	MDA (nmol/L)	TAS (mmol/L)
caz	0,16	7,08	1,32
caz	0,18	8,37	0,96
caz	0,17	6,92	1,34
caz	0,19	4,35	1,19
caz	0,18	6,60	0,86
caz	0,17	6,82	0,91
caz	0,18	7,49	0,93
caz	0,18	6,78	1,64
caz	0,16	6,40	1,49
caz	0,19	7,17	1,08
…	…	…	…
martor	0,76	0,43	1,18
martor	0,75	0,19	1,16
martor	0,93	0,24	1,18
martor	0,80	0,19	1,17
martor	0,88	0,37	1,19
…	…	…	…

MDA = malondialdehida; TAS = statusul antioxidant total

Clasificarea unei serii statistice se poate face în funcție de:

- Numărul de variabile luate în considerare simultan:

Numărul variabilelor	Seria statistică este:
1	Univariată
2	Bivariată
3	Trivariată
>3	Multivariată

- Numărul de variabile cantitative:

Numărul variabilelor cantitative	Seria statistică este:
1	Unidimensională
2	Bidimensională
3	Tridimensională
>3	Multidimensională

O seria statistică poate astfel să fie:

- Univariată: genul persoanelor consultate în serviciul de urgență în decurs de o oră a unei zile de sâmbătă

Gen (F/M)	F	F	M	M	M	F	M	F	M	M

- Unidimensională: valorile glicemiei (mg/dL) măsurate pe un eșantion de 10 pacienți cu disfuncție cognitivă

Glicemie (mg/dL)	83	104	89	100	90	88	88	82	90	112

- Bivariată, unidimensională: genul (F/M) și valorile presiunii arteriale sistolice (mmHg) (PAS) ale unui eșantion de 10 subiecți

Gen (F/M)	F	F	M	M	M	F	M	F	M	M
PAS (mmHg)	110	100	95	110	120	110	220	200	150	160

- Tridimensională: seria statistică din Exemplul 2.1.
- Multivariată, tridimensională: genul, obezitatea, vârsta, presiunea arterială sistolică (PAS) şi presiunea arterială diastolică (PAD) la un eșantion de 10 subiecți:

Gen (F/M)	F	F	F	M	F	F	F	F	M	F	F
Obezitate (da=1/nu=0)	0	0	0	1	0	0	0	0	1	0	0
Vârsta (ani)	66	55	46	74	62	59	59	60	60	53	79
PAS (mmHg)	140	123	149	147	127	130	111	140	117	115	112
PAD (mmHg)	103	102	83	84	74	86	101	100	117	109	83

2.3 Proporția, rata și raportul (rația)

În cazul **variabilelor calitative** se pot calcula și următorii parametri statistici: proporția, rata și raportul (rația).

2.3.1 Proporția

Proporția (frecvența) este raportul în care numărătorul face parte din numitor.

- Formula: a:(a+b) sau a/(a+b)
- Numitorul *b* nu include în mod obligatoriu subiecții numărătorului *a*.
- Poate lua valori între 0 și 1 (0 și 100 dacă se exprimă procentual).

Prevalența este proporția de indivizi dintr-o populație care au boala la un moment dat.

Prevalența = (numărul de cazuri de boală) / (total populație)

Incidența este proporția cazurilor noi dintr-o anumită afecțiune care apar într-o populație la risc într-un interval de timp.

Incidența = (numărul de cazuri noi într-o perioadă de timp) / (populația totală la risc în aceeași perioadă de timp)

Exemplul 2.2.

La serviciul de urgențe ale unui spital județean s-au prezentat pentru consultație 1200 pacienți. 420 au fost internați (200 femei și 220 bărbați).

Pentru Exemplul 2.2.:

- Proporția subiecților internați = 420/1200*100 = 35%
- Proporția subiecților de gen feminin internați = 200/420*100 = 48%
- Proporția subiecților de gen masculin internați = 220/420*100 = 52%

Proporția este raportul în care numărătorul face parte din numitor.

Parametrii utilizaţi în evaluarea semnelor și testelor diagnostice (sensibilitatea - Se, specificitatea - Sp, acurateţei - A, valoare predictivă pozitivă - VPP şi negativă- VPN) fac parte din categoria proporţiilor.

2.3.2 Rata

Rata reflectă riscul de apariție a unui eveniment în timp (ex. secundă/minut/oră/zi/săptămână/lună/an). Rata ia valori pozitive în intervalul [0, ∞).

Exemple: rata de morbiditate, rata de atac, rata de moralitate, rata de natalitate

Exemplul 2.3.

Într-un oraș cu o populație de 100.000 locuitori au fost înregistrați 200 nou-născuți vii în anul 1999. Care este rata de natalitate?

Soluția:

Rata de natalitate = 200/100.000 * 1.000 = 2‰

→ 2 nou născuți la o mie de locuitori

Un tip de indicator medical derivat din rate este riscul atribuabil (simbol RA). RA este diferența dintre rata de incidență la cei expuși (I_e) și rata de incidență la cei neexpuși (I_n). Riscul atribuabil procentual (RA%) este procentul incidenței patologiei de interes datorată expunerii:

$$RA\% = 100*[(I_e-I_n)/I_e]$$

RA% este procentul incidenţei patologiei de interes în grupul expus care va fi eliminat dacă expunerea încetează.

2.3.3 Raportul

Se calculează doar în cazul numerelor raționale pozitive *a* şi *b* unde $b \neq 0$.

- Formula:

  a:b sau a/b

- Numitorul *b* nu include în mod obligatoriu subiecții numărătorului *a*.

Exemplu 2.4.

Din cele 10 persoane consultate în 22 August 2016 de medicul de familie X, 2 au avut valorile presiunii arteriale sistolice (PAS) mai mari decât valorile normale. Care este raportul PAS normal / PAS patologic?

Soluţie:

Raportul PAS normal / PAS patologic = 8/2 = 4

→ la 4 subiecți cu PAS în limite normale există un subiect cu valori PAS mai mari decât valorile normale

Rata şansei (OR = odds ratio) și riscul relativ (RR) fac parte din această categorie.

- Riscul relativ (simbol RR) este raportul dintre rata de incidență la cei expuși și rata de incidență la cei neexpuși.
    - Un RR egal cu 1 indică faptul că riscul în grupul celor expuşi este egal cu riscul în grupul celor ne-expuşi.
    - Un RR > 1 indică faptul că expunerea este factori de risc pentru patologia de interes.
    - Un RR < 1 indică faptul că expunerea este un factor de protecţie pentru patologia de interes.

Exemplul 2.5.

S-a studiat asocierea între fumat şi cancerul pulmonar pe un eşantion de 1.000 subiecţi. În eşantionul studiat au existat 20 subiecţi cu cancer pulmonar, 12 dintre aceştia fiind şi fumători. 750 din subiecţii din studiu au fost şi nefumători şi fără cancer pulmonar. Evaluaţi riscul relativ şi riscul atribuabil.

Soluţie:

- Riscul în grupul celor expuşi = 12/242 = 0,05
- Riscul în grupul celor ne-expuşi = 8/758 = 0,01
- RR = 5 → riscul de a dezvolta cancer pulmonar la subiecţii fumători este de 5 ori mai mare comparativ cu riscul de a dezvolta cancer pulmonar la subiecţii nefumători.
- RA = 0,05-0,01 = 0,04 → dacă fumătorii se lasă de fumat incidenţa cancerului de plămân la fumători scade cu 0,04 la 100 de indivizi.
- RA% = 100*[(0,05-0,01)/0,05] = 80% → dacă fumătorii se lasă de fumat incidenţa cancerului de plămân la fumători se reduce cu 80%

2.4 Parametrii de centralitate

Problema de bază a statisticii poate fi enunțată după cum urmează:

Fie seria statistică formată din x_1, ..., x_n, unde x_1 este valoarea variabilei x (măsurată/observată) a primului subiect din eșantion şi n volumul/talia eșantionului. Presupunând că eșantionul a fost extras dintr-o populație P, care sunt concluziile care se pot trage asupra populației P prin studierea eșantionului?

Pentru a răspunde la această întrebare, datele trebuie sumarizate cât mai succint posibil, deoarece în studiile medicale numărul de subiecți din eșantion este de cele mai multe ori mare și e ușor să pierdem mesajul general privind datele brute. Diferiți parametri statistici sunt utilizați pentru a sumariza datele, aceşti parametri fiind aleşi în funcție de tipul variabilei (calitativă sau cantitativă). Parametrii statistici calculați pe datele unui eșantion se numesc *estimatori* punctuali deoarece se calculează pe baza eșantionului pentru a estima ce se întâmplă în populația P din care eșantionul a fost extras.

Parametrii de centralitate indică distribuția datelor seriei statistice în raport cu valorile centrale (parametrii de centralitate).

Media aritmetică, mediana și modulul sunt cei mai des folosiți parametri de centralitate, dar în această secțiune se vor discuta și media geometrică, media ponderată și amplitudinea.

2.4.1 Media aritmetică

Media aritmetică este una din cele mai vechi metode utilizate în prezentarea datelor observate/măsurate, conceptul fiind folosit de Hipparchus şi menţionat pentru prima dată într-un document ştiinţific în 1755.

Media aritmetică este suma tuturor valorilor seriei statistice împărțită la volumul/talia eșantionului. Calculul mediei aritmetice are semnificație doar dacă variabila este cantitativă.

Simbolul ***parametrului populației*** pentru media aritmetică este diferit față de simbolul ***estimatorului*** calculat pe ***eșantion***, dar formula de calcul este similară, singura diferență e dată de implicarea în calcul a volumului populației (N) în cazul parametrului populației și respectiv a volumului eșantionului (n) în cazul eșantionului.

Populație	**Eșantion**
$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$

unde N = volumul/talia populației; n = volumul/talia eșantionului, ∑ (sigma) = suma, este forma prescurtată pentru ($x_1 + x_2 + ... + x_n$).

Exemplul 2.6.

Fie seria statistică unidimensională formată din greutatea la naștere (exprimată în grame) a primilor 10 copiilor cu restricție de creștere intrauterină născuți în 2015 la Spitalul Clinic Județean de Urgență Cluj-Napoca, Clinica Ginecologie I.

i	1	2	3	4	5	6	7	8	9	10
x_i	1900	1800	2800	2800	3500	3800	2400	2500	3500	3600

Care este valoarea mediei aritmetice pentru această serie statistică?

Soluția:

$\sum X_i$ = 1900+1800+2800+2800+3500+3800+2400+2500+3500+3600

$\sum X_i$ = 28600

$$\overline{X} = \frac{1900 + 1800 + ... + 3600}{10} = \frac{28600}{10} = 2860\ g$$

Media aritmetică:

- Este un estimator care ia în considerare toate valorile seriei statistice (vezi formula de calcul).
- Valoarea mediei aritmetice se situează printre valorile seriei statistice. Pentru datele din Exemplul 2.6, greutatea la naștere ia valori de la 1800 grame (valoarea minimă) la 3800 grame (valoarea maximă) iar valoarea mediei aritmetice este de 2860 grame.
- Valorile extreme influențează media aritmetică. Dacă al șaselea subiect inclus în studiu ar avea valoarea greutății la naștere egală cu 800 g, media aritmetică a eșantionului ar avea valoarea de 2560 g.
  Este suficientă modificarea unei singuri valori a seriei statistice pentru ca media aritmetică să se modifice!
- Suma diferențelor dintre valorile individuale ale seriei şi medie este egală cu zero:

$$\sum_{i=1}^{n} \left(x_i - \overline{X}\right) = 0$$

  Pentru datele din Exemplul 2.6:
  (1900-2860) + (1800-2860) + (2800-2860) + (2800-2860) + (3500-2860) + (3800-2860) + (2400-2860) + (2500-2860) + (3500-2860) + (3600-2860) =
  -960-1060-60-60+640+940-460-360+640+740 = 0

- Modificarea datelor brute are efect asupra mediei aritmetice:

- Schimbarea originii scalei de măsură: dacă la fiecare valoare a seriei statistice de adună o constantă *c*, valorile seriei statistice devin $x_i + c_1$ iar valoarea mediei aritmetice este dată de formula:

$$\overline{X}' = \overline{X} + c_1$$

Această proprietate a mediei aritmetice este foarte importantă, deoarece în unele cazuri schimbarea originii scalei este necesară dar raportarea rezultatelor trebuie făcută raportat la scala originală.

- Transformarea scalei de măsură: dacă fiecare valoare a seriei statistice se înmulțește cu o valoare constantă c_2, valorile seriei statistice devine $x_i \cdot c_2$ (vezi Exemplul 2.7), iar media aritmetică este dată de formula:

$$\overline{X}' = \overline{X} \cdot c_2$$

Proprietatea se aplică și dacă se dorește modificarea simultană atât a originii cât și transformarea scalei.

Exemplul 2.7.

Dacă avem un eșantion cu temperaturi exprimate în grade Celsius cu media aritmetică egală cu 43°C, care este media aritmetică exprimate în grade Fahrenheit (°F)?

Soluție:

Fie y_i temperatura în °F care corespunde temperaturii în °C pentru x_i.
Formula de calcul:

- pentru valorile seriei statistice este: $y_i = 9/5 * x_i + 32$
- pentru media aritmetică: $\overline{X}' = \frac{9}{5} \quad 43 + 32 = 77,4 + 32 = 109,4$

Media aritmetică este un parametru de centralitate care este influenţat de valorile seriei statistice. În cazuri extreme, majoritatea datelor seriei statistice pot să fie mai mici sau mai mari decât media aritmetică, iar una sau două valori să fie de partea opusă majorității datelor. Cu toate acestea, media aritmetică este cel mai utilizat parametru de centralitate.

2.4.2 Mediana

Mediana este o mărime de centralitate, pe locul doi în ceea ce privește popularitatea, fiind valoarea care împarte seria statistică în două părți egale.

Mediana (Me):

- Are formulă de calcul diferită pentru un eșantion care are un număr par de observații (ex. 2, 4, 6, ...) față de un eșantion care are un număr impar de observații:

n=par	n=impar
$Me = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$	$Me = x_{\frac{n+1}{2}}$

- În cazul în care volumul eșantionului este impar, valoarea medianei va fi egală cu una din valorile seriei statistice.
- Valoarea Me nu este afectată de valorile extreme ale seriei de date.
- Valoarea medianei poate fi nereprezentativă pentru distribuția datelor seriei dacă valorile individuale nu se grupează înspre valoarea centrală.
- Mediana este o măsură de tendință centrală care minimizează suma valorilor absolute ale abaterilor de la o valoare *x* de pe dreapta numerelor reale.

$$\sum_{i=1}^{n} |x_i \quad Me| = \sum_{i=1}^{n} |x_i \quad \overline{X}|$$

Calcularea medianei se va exemplifica prin folosirea datelor din Exemplul 2.6

- Ordonăm datele crescător:

i	1	2	3	4	5	6	7	8	9	10
x_i	1800	1900	2400	2500	2800	2800	3500	3500	3600	3800

- Volumul eșantionului este par (n=10), deci mediana este dată de formula:

$$Me = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} = \frac{x_5 + x_6}{2} = \frac{2800 + 2800}{2} = 2800$$

Exemplul 2.8.

Fie seria statistică formată din valorile greutății la naștere exprimată în grame la primii 9 subiecți cu restricție de creștere intrauterină prezentați în Exemplul 2.6. Care este valoarea medianei pentru această serie statistică?

Soluție:

În acest caz, seria ordonată este:

i	1	2	3	4	5	6	7	8	9
x_i	1800	1900	2400	2500	2800	2800	3500	3500	3600

Valoarea medianei va fi dată de formula:

$$Me = x_{\frac{n+1}{2}} = x_{\frac{9+1}{2}} = x_5 = 2800\,g$$

2.4.3 Modulul

Denumit și valoare modală, modulul este valoarea care apare cu cea mai mare frecvență în seria statistică. Nu există nici o formulă de calcul pentru valoarea modulului, valoarea lui se obține prin numărarea valorilor identice din seria statistică.

În funcție de numărul de valori cel mai frecvente, seria statistică poate să fie:

- Unimodală:

i	1	2	3	4	5	6	7	8	9	Modulul
x_i	140	140	160	170	180	180	180	190	190	180

- Bimodală:

i	1	2	3	3	4	5	6	7	8	9	Modulul
x_i	110	110	120	120	120	130	140	140	140	160	120, 140

- Multimodală:

i	1	2	3	4	5	6	7	8	9	10	Modulul
x_i	110	110	110	120	120	120	130	140	140	140	110, 120, 140

2.4.4 Alți parametri de centralitate

Media geometrică

Media geometrică (MG), introdusă de William Stanley Jevons în 1863, este utilizată în situațiile în care modificările urmărite sunt exponențiale sau în cazul distribuțiilor asimetrice care devin sau nu simetrice prin logaritmarea datelor brute. Media geometrică se utilizează frecvent în microbiologie/parazitologie și în imunologie. Calcularea mediei geometrice nu are sens dacă datele brute pot lua valoarea zero sau pot lua valori negative.

Formula: $MG = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$

Exemplul 2.9.

O specie a unei bacterii își crește populația cu 20% în prima oră, 30% în cea de-a doua oră, respectiv cu 50% în cea de-a treia oră. Estimați creșterea procentuală medie.

Soluția:

	Prima oră	A doua oră	A treia oră
Rata de creștere	1,2 (100*1,2=120)	1,3 (120*1,3=156)	1,5 (156*1,5=234)

MG = $(1,2*1,3*1,5)^{1/3} = (2,34)^{1/3} = 1,3276$

Rata medie de creștere a bacteriei într-o perioadă de 3 ore este de 32,75%, ducând la o populație de 234 de bacterii, dacă s-a plecat de la o populație de 100 de bacterii.

Media ponderată

Media ponderată a fost introdusă de Roger Cotes în 1712 și este parametrul de centralitate care se utilizează pentru a pondera valorile individuale în funcție de un criteriu definit (ex. anumite valori din seria statistică sunt mai importante decât celelalte). Media ponderată își găsește utilitatea în cazul meta-analizelor.

În calcularea mediei ponderate fiecare valoare x_i este înmulțită cu o pondere p_i pozitivă, care indică importanța valorii respective în raport cu celelalte valori:

$$\overline{X}_{ponderată} = \frac{\sum_{i=1}^{n} (p_i \quad x_i)}{\sum_{i=1}^{n} p_i}$$

Media ponderată este utilizată în clasificarea studenților și acordarea burselor (se utilizează media generală care este o medie aritmetică dintre media aritmetică și media ponderată).

Valoarea centrală

Formula de calcul pentru valoarea centrală este:

$$\text{Valoarea centrală} = (x_{min} + x_{max})/2$$

unde $x_{min} = \min \{x_1, ..., x_n\}$ şi $x_{max} = \max \{x_1, ..., x_n\}$

Valoarea centrală este un estimator rar folosit ca măsură de centralitate deoarece:

- are eficientă redusă, luând în considerare doar valorile extreme ale seriei statistice:

i	1	2	3	4	5	6	7	8	9	10	Valoarea centrală
x_i	1800	1900	2400	2500	2800	2800	3500	3500	3600	3800	=(1800+3800)/2 = 2800

- are putere redusă, deoarece valoarea se modifică semnificativ în condițiile existenței valorilor extreme:

i	1	2	3	4	5	6	7	8	9	10	Valoarea centrală
x_i	1800	1900	2400	2500	2800	2800	3500	3500	3600	3800	=(800+3800)/2 = 2300

- nu este reprezentativă pentru o serie statistică deoarece ia aceeași valoare pentru serii statistice cu valori minime și maxime identice.

Fie două serii statistice formate din notele la examenul practic a două grupe cu 10 studenți:

i	1	2	3	4	5	6	7	8	9	10	Valoarea centrală
x_i	8	7	8	9	9	10	8	9	7	4	=(4+10)/2 = 7

i	1	2	3	4	5	6	7	8	9	10	Valoarea centrală
x_i	6	6	5	7	4	10	4	8	9	4	=(4+10)/2 = 7

Valoarea centrală este identică pentru ambele grupe, dar distribuția notelor este diferită (vezi Figura 2.2).

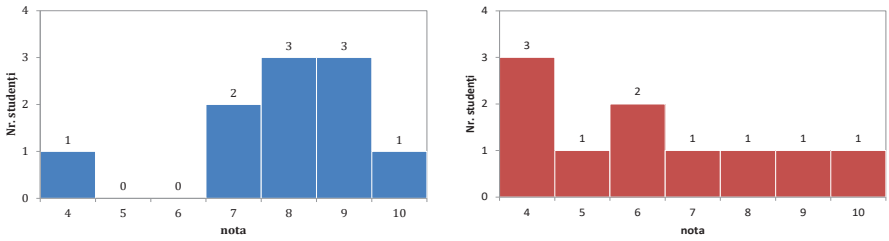


Fig. 2.2. Distribuția notelor pentru cele două serii statistice: majoritatea notelor din serie statistică au valori de la 7 la 10 (stânga); notele de 4 şi respectiv 6 au o frecvenţă mai mare, iar restul notelor au frecvența egală cu 1, distribuția notelor din a doua seria statistică fiind uniformă (dreapta)

2.4.5 Avantaje și dezavantaje ale parametrilor de centralitate

Principalele avantaje și dezavantaje ale estimatorilor de centralitate sunt redate în Tabelul 2.2.

Tabelul 2.2. Avantaje și dezavantaje ale diferiților parametri de centralitate

Denumire	Avantaje	Dezavantaje
Media aritmetică	Utilizează toate datele Ușor de aplicat (formulă simplă)	Influențată de valori extreme. Nu este utilă pentru date discrete sau calitative și dacă dacă datele nu au o distribuție normală.
Mediana	Nu este influențată de valori extreme. Nu este influențată de asimetria datelor.	Ignoră majoritatea datelor din serie.
Modulul	Aplicabil și variabilelor calitative .	Ignoră majoritatea datelor din serie.
Media geometrică	Aplicabilă datelor cu distribuție asimetrică.	Este indicat a se aplica doar dacă transformarea datelor prin logaritmare produce o distribuție normală.
Media ponderată	Cuantifică importanța relativă a fiecărei observații.	Este necesară cunoașterea sau estimarea ponderilor.

2.5 Parametrii de dispersie

În această secțiune se vor descrie parametri de cuantificare ai variabilității. În literatura de specialitate descrierea variabilelor cantitative se realizează cu ajutorul unui parametru de centralitate și a unui parametru de dispersie, cei mai utilizați fiind media aritmetică și deviația standard.

Parametrii de dispersie descriu variabilitatea sau dispersia datelor unei serii statistice:

- Oferă informații privind împrăștierea sau aglomerarea datelor.

- Permit stabilirea reprezentativității măsurilor de centralitate. De exemplu, semnificația unei medii ca și valoare reprezentativă pentru o variabilă depinde de gradul de împrăștiere a valorilor individuale brute în jurul ei.
- Au un rol important în estimarea parametrilor statistici și în inferența statistică.

Parametrii de dispersie care vor fi exemplificați în această secțiune sunt amplitudinea, cvartilele și percentilele, varianța și deviația standard și respectiv coeficientul de variație.

Exemplul 2.10 pune în evidență importanța dispersiei datelor pentru două serii statistice cu medii artimetice identice.

Exemplul 2.10.

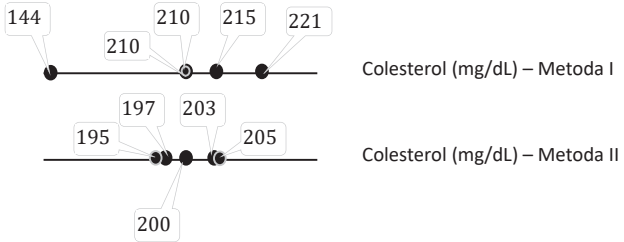
Colesterolul sangvin a fost măsurat prin două metode diferite (unitatea de măsură mg/dL) la un eșantion de 5 persoane. Valorile obținute au fost:

i	1	2	3	4	5	$\bar{X}$
x_i-metoda I (mg/dL)	210	144	210	215	221	200
x_i-metoda II (mg/dL)	203	205	195	197	200	200

Care metodă de măsurare a colesterolului este mai bună?

Soluția:

Media aritmetică este identică pentru cele două metode (200 mg/dL), dar dacă ne uităm la modalitatea în care datele sunt distribuite (distanța dintre valorile individuale și valoarea mediei sau a medianei), se poate observa că valorile date de cele două metode sunt diferite.



Valorile obținute prin prima metodă au o variabilitate/dispersie mai mare comparativ cu valorile obținute prin aplicarea celei de a doua metode.

2.5.1 Amplitudinea

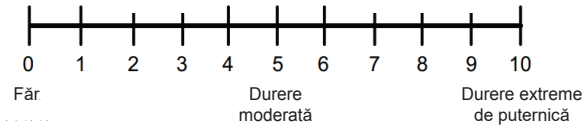
Amplitudinea este cel mai simplu parametrul de dispersie.

- Definiție: diferența dintre valoarea maximă și minimă a observațiilor seriei statistice.

- Formula de calcul: $A = X_{man}-X_{min}$, unde A = amplitudinea, X_{max} = valoarea cea mai mare a seriei statistice, X_{min} = valoarea cea mai mică a seriei statistice.
- Avantaj: ușor de calculat.
- Dezavantaje:
    - Amplitudinea este sensibilă la valorile extreme: dacă valoarea minimă a colesterolului măsurat prim metoda I este egală cu 100, A=221-100=21 comparativ cu valoarea obținută pentru datele din **Exemplul 2.10** Metoda I – A=221-144 = 77)
    - Amplitudinea este influențată de volumul eșantionului și tinde să fie mai mare pe măsură ce volumul eșantionului crește. Deoarece valoarea amplitudinii fluctuează cu volumul eșantionului, amplitudinea nu se utilizează în compararea grupurilor care au volume de eșantion diferite.
    - Amplitudinea nu ne dă informații cu privire la modalitatea în care datele variază în jurul valorii centrale (Figura 2.3).

Exemplu 2.11.

S-a evaluat durerea percepută de un eșantion de subiecți cu boală cronică inflamatorie autoimună (spondilită anchilozantă sau artrită psoriazică, *n*=50) și respectiv un eșantion de subiecți cu poliartrită reumatoidă (*n*=50). Pentru cuantificarea durerii s-a utilizat scala numerică vizuală de durere:



Care este valoarea amplitudinii scorului de durere pentru pacienții cu spondilită anchilozantă sau artrită psoriazică, respectiv pentru pacienții cu poliartrită reumatoidă? Este dispersia datelor identică la cele două eșantioane?

Soluția:

Ambele grupuri investigate au avut scoruri de durere de la 4 la 10, deci amplitudinea durerii a fost de 6 atât la subiecții cu boală inflamatorie autoimună cronică cât și la cei cu poliartrită reumatoidă (Figura 2.3). Cu toate acestea dispersia scorurilor de durere este diferită pentru cele două grupuri: durerea extrem de puternică este mai frecventă în grupul celor cu poliartrită reumatoidă, iar pentru cei cu boală inflamatori autoimună cronică se obțin mai degrabă scoruri moderate de durere.

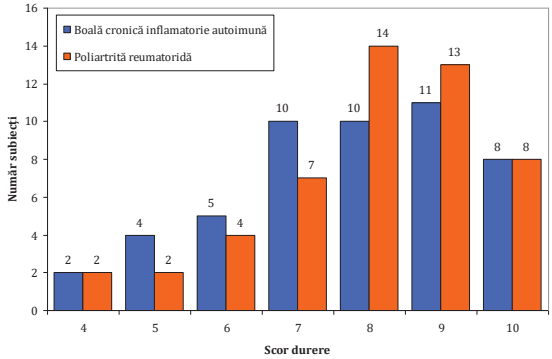
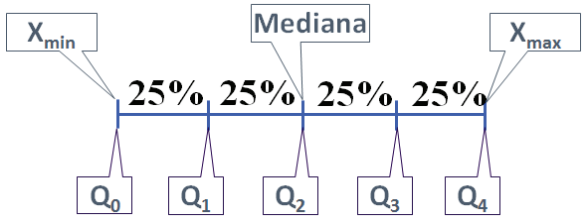


Fig. 2.3. Distribuția scorului de durere la un eșantion de 50 subiecți cu spondilită anchilozantă (boală cronică inflamatorie autoimună) și respectiv un eșantion de 50 subiecți cu poliartrită reumatoidă

2.5.2 Percentile

În cazul percentilelor, seria statistică este împărțită în părți egale în funcție de volumul eșantionului. Cele mai utilizate percentile sunt cvartilele (Q1 = percentila de 25%, Q2 = percentila de 50% (mediana) și Q3 = percentila de 75%) și decilele (percentilele de 10%, 20%, ..., 90%).

În cazul cvartilelor, seria statistică este împărțită în 4 părți egale:



- Cvartila zero (Q0) are valoarea egală cu valoarea minimă din seria statistică.
- Cvartila 1 (Q1) este acea valoare care împarte seria statistică în două, astfel încât 25% dintre datele seriei să fie mai mici sau egale cu valoarea Q1, iar 75% mai mari decât valoarea Q1.
- Cvartila 2 (Q2) este acea valoare care împarte seria statistică în două părți egale. Are aceeași valoare ca și mediana.
- Cvartila 3 (Q3) este acea valoare care împarte seria statistică în două, astfel încât 75% din datele seriei sunt mai mici sau egale cu valoarea Q3, iar 25% mai mari decât valoarea Q3.
- Cvartila 4 (Q4) este egală cu valoarea maximă.

Cvartilele și respectiv percentilele sunt utilizate în sumarizarea datelor atunci când parametrul de centralitate care descrie cel mai bine seria statistică este mediana. Are sens calcularea cvartilelor și/sau percentilelor pentru variabile ordinale (calitative), interval (cantitative) sau raport (cantitative).

Exemplul 2.12.

Fie greutatea la naștere a primilor 20 nou-născuți cu restricție de creștere intrauterină cu vârstă gestațională egală cu 40 și respectiv 41 săptămâni incluși într-un studiu.

i	1	2	3	4	5	6	7	8	9	10
x_i	2450	2500	1900	2700	2650	2190	2600	2800	2600	2490

i	11	12	13	14	15	16	17	18	19	20
x_i	2200	2350	2700	2650	2660	2700	2600	2900	2350	2490

Ce putem spune despre dispersia datelor acestei serii statistice?

Soluția:

i	x_i
1	2450
2	2500
3	1900
4	2700
5	2650
6	2190
7	2600
8	2800
9	2600
10	2490
11	2200
12	2350
13	2700
14	2650
15	2660
16	2700
17	2600
18	2900
19	2350
20	2490

ordonare

x_i
1900
2190
2200
2350
2350
2450
2490
2490
2500
2600
2600
2600
2650
2650
2660
2700
2700
2700
2800
2900

Q1 = 2425

Me = $(x_{10}+x_{11})/2$
Me = 2600

Q3 = 2670

A = 2900-1900 = 1000

Pentru această serie statistică, interpretarea parametrilor este următoarea:

- Jumătate din subiecți au avut greutatea la naștere mai mică sau egală cu 2600g.
- 25% din subiecți au avut greutatea la naștere mai mică sau egală cu 2425g.
- 75% din subiecți au valoarea greutatea la naștere mai mică sau egală cu 2670g.
- Diferența dintre valoarea maximă și minimă la naștere a fost de 1000g.

2.5.3 Varianța și deviația standard

În exemplul 2.10 diferența dintre valorile colesterolului măsurate prin cele două metode apare deoarece valorile obținute prin aplicarea metodei II sunt mai apropiate de valoarea mediei. Dacă parametrul de centralitate este media aritmetică, o mărime simplă care cuantifică dispersia datelor este media distanțelor față de medie:

$$d = \frac{\sum_{i=1}^{n} \left(x_i - \overline{X}\right)}{n}$$

Această măsură nu este potrivită, deoarece suma diferențelor dintre valorile individuale ale serie și medie aritmetică este zero. Pentru datele din Exemplul 2.10 (metoda I/II):

- Metoda I:
  - d = (210-200) + (144-200) + (210-200) + (215-200) + (221-200)
  - d = 10-56+10+15+21 = 0
- Metoda II:
  - d = (203-200) + (205-200) + (195-200) + (197-200) + (200-200)
  - d = 3 + 5 – 5 - 3+0 = 0

Această mărime nu permite evaluarea diferenței variabilității celor două metode.

O mărime alternativă este reprezentată de media deviației care este un indicator de dispersie ameliorat în raport cu amplitudinea. Se poate calcula media abaterii de la medie sau media abaterii de la mediană:

- media abaterii de la media aritmetică:

$$\frac{\sum_{i=1}^{n} \left|x_i - \overline{X}\right|}{n}$$

- media abaterii de la mediană:

$$\frac{\sum_{i=1}^{n} \left|x_i - Me\right|}{n}$$

Pentru valorile colesterolului determinate prin cele două metode valorile mediilor abaterilor sunt:

- Metoda I:
  - Media abaterii de la medie: (|210-200| + |144-200| + |210-200| + |215-200| + |221-200|)/5 = (10+56+10+15+21)/5 = 112/5 = 22,40
  - Media abaterii de la mediană: (|210-210| + |144-210| + |210-210| + |215-210| + |221-210|)/5 = (0+66+0+5+11)/5 = 82/5 = 16,40
- Metoda II:
  - Media abaterii de la medie: (|203-200| + |205-200| + |195-200| + |197-200| + |200-200|)/5 = (3 + 5 + 5 + 3 + 0)/5 = 16/5 = 3,20
  - Media abaterii de la mediană: (|203-200| + |205-200| + |195-200| + |197-200| + |200-200|)/5 = (3 + 5 + 5 + 3 + 0)/5 = 16/5 = 3,20

Media abaterii este o mărime de dispersie care permite compararea celor două metode de măsurare a colesterolului: dispersia valorilor față de medie este de 7 ori mai mare pentru determinările prin aplicarea primei metode comparativ cu determinările prin cea de-a doua metodă. Dispersia valorilor față de mediană este de 5 ori mai mare pentru determinările prin aplicarea primei metode comparativ cu determinările prin cea de-a doua metodă.

Metoda cea mai cunoscută de determinare a dispersiei unei serii statistice care urmează distribuția normală este însă deviația standard (cunoscută și ca abatere standard sau ecart tip).

Varianța, media deviațiilor pătratice față de media aritmetică, are următoarea formulă de calcul:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n}$$

Unitatea de măsură a variației este pătratul unității de măsură a datelor brute. **Abaterea sau deviația standard a populației**, cunoscută și sub denumirea de **ecart tip**, este o mărime a dispersiei care are aceeași unitate de măsură ca și datele brute și are următoarea formulă de calcul:

$$s = \sqrt{s^2}$$

În statistica inferențială pentru o mai bună aproximare formula variației utilizează *n-1* (corecția Bessel) la numitor și rezultă astfel **varianţa de eșantionare** sau **varianța**:

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n-1}$$

Abaterea sau deviația standard a eșantionului este o mărime a dispersiei care are aceeași unitate de măsură ca și datele brute și are următoarea formulă de calcul:

$$S = \sqrt{S^2}$$

Termenul de deviaţie standard a fost introdus de Karl Pearson în cadrul unei conferinţe din 1893. Simbolul grecesc asociat deviaţiei standard a populaţiei (σ) a fost introdus de William Sealy Gosset (cunoscut în statistică şi sub pseudonimul 'Student') și se folosește în unele tratate.

Tabelul 2.3. Terminologia utilizată pentru variație

Statistică descriptivă (eșantion sau întreaga populație)	**Statistică inferențială (aproximare pentru întreaga populație pe baza eșantionului)**
Varianța (s)	Varianţa (S) Varianţa de eşantionare
Ecartul tip (s^2) Abaterea (deviația) standard a populației	Abaterea (deviația) standard (S^2)

Valorile variației și a deviației standard ale colesterolului pentru cele două metode utilizate sunt:

- Metoda I:
  - varianța: $S^2 = [(210-200)^2 + (144-200)^2 + (210-200)^2 + (215-200)^2 + (221-200)^2]/(5-1) = [100 + 3136 + 100 + 225 + 441]/4 = 4002/4 = 1000,5$
  - deviația standard: $S = \sqrt{1000,5} = 31,63$
- Metoda II:
  - varianța: $S^2 = [(203-200)^2 + (205-200)^2 + (195-200)^2 + (197-200)^2 + (200-200)^2]/(5-1) = [9 + 25 + 25 + 9 + 0]/4 = 68/4 = 17$
  - deviația standard: $S = \sqrt{17} = 4,12$

Interpretând rezultatele deviației standard, putem afirma că variabilitatea valorilor colesterolului măsurat cu ajutorul primei metode este de 7 ori mai mare comparativ cu variabilitatea valorilor măsurate cu ajutorul celei de a doua metode.

Proprietăți ale variației și ale deviației standard:

- Schimbarea originii seriei statistice nu are efect nici asupra valorii variației și nici asupra valorii deviației standard (vezi exemplul 2.13).

Exemplu 2.13.

Fie seria statistică a greutății la naștere a 10 copii cu restricție de creștere intrauterină exprimată în grame (x_i).

i	1	2	3	4	5	6	7	8	9	10
x_i (g)	1800	1900	2400	2500	2800	2800	3500	3500	3600	3800
$y_i = x_i$ -100 (g)	1700	1800	2300	2400	2700	2700	3400	3400	3500	3700

Cântarul cu care au fost măsurate greutățile s-a dovedit a fi decalibrat, arătând cu 100 de grame mai mult decât valoarea reală. Cum afectează această decalibrare valoarea mediei, a variației și a deviației standard?

Soluție (pentru x_i):

- media aritmetică: (1800 + 1900 + 2400 + 2500 + 2800 + 2800 + 3500 + 3500 + 3600 + 3800)/10 = 28600/10 = 2860
- varianța: $[(1800-2860)^2 + (1900-2860)^2 + (2400-2860)^2 + (2500-2860)^2 + (2800-2860)^2 + (2800-2860)^2 + (3500-2860)^2 + (3500-2860)^2 + (3600-2860)^2 + (3800-2860)^2]/9] = (1123600 + 921600 + 211600 + 129600 + 3600 + 3600 + 409600 + 409600 + 547600 + 883600)/9 = 4644000*9 = 516000$
- deviația standard: = 516000 = 718,33

Soluție (pentru y_i):

- media aritmetică: (1700 + 1800 + 2300 + 2400 + 2700 + 2700 + 3400 + 3400 + 3500 + 3700)/10 = 27600/10 = 2760
- varianța: $[(1700-2760)^2 + (1800-2760)^2 + (2300-2760)^2 + (2400-2760)^2 + (2700-2760)^2 + (2700-2760)^2 + (3400-2760)^2 + (3400-2760)^2 + (3500-2760)^2$

+ (3700-2760)^2]/9 = (1123600 + 921600 + 211600 + 129600 + 3600 + 3600 + 409600 + 409600 + 547600 + 883600)/9 = 4644000/9 = 516000

- deviația standard: =SQRT(516000) = 718,33

Se poate, astfel, observa că media aritmetică a seriei y_i este cu 100 mai mică comparativ cu media aritmetică a seriei x_i, adică cu valoare cu care aparatul a fost decalibrat. Decalibrarea aparatului nu a avut însă efect nici asupra variației și nici asupra deviației standard, valorile acestora fiind identice pentru seriile x_i și respectiv y_i.

- Transformarea scalei de măsură prin înmulțirea fiecărei valori cu o constantă determină modificarea variației și a deviației standard după cum urmează: Fie seriile statistice $x_1, \dots x_n$ și respectiv $y_1, \dots y_n$ unde $y_i = c*x_i$ (i = 1, …, n; c>0)

$$S^2_y = c^2*S^2_x$$
$$S_y = c*S_x$$

Exemplul 2.14.

Fie seria statistică a greutății la naștere a 10 copii cu restricție de creștere intrauterină exprimată în grame (x_i) și respectiv kg ($y_i = x_i*0,001$, c=0,001):

i	1	2	3	4	5	6	7	8	9	10	Varianța	Deviația standard
x_i (g)	1800	1900	2400	2500	2800	2800	3500	3500	3600	3800	516000	718,331
y_i (kg)	1,8	1,9	2,4	2,5	2,8	2,8	3,5	3,5	3,6	3,8	0,516	0,718

Care este media aritmetică și varianța pentru x_i și y_i?

Soluție:

Pentru x_i:

- media aritmetică: = (1800 + 1900 + 2400 + 2500 + 2800 + 2800 + 3500 + 3500 + 3600 + 3800)/10 = 28600/10 = 2860
- varianța: [(1800-2860)^2+ (1900-2860)^2 + (2400-2860)^2 + (2500-2860)^2 + (2800-2860)^2 + (2800-2860)^2 + (3500-2860)^2 + (3500-2860)^2 + (3600-2860)^2 + (3800-2860)^2]/(10-1) = (1123600 + 921600 + 211600 + 129600 + 3600 + 3600 + 409600 + 409600 + 547600 + 883600)/9 = 4644000/9 = 516000
- deviația standard: =SQRT(516000) = 718,33

Pentru y_i:

- media aritmetică: = (1,8 + 1,9 + 2,4 + 2,5 + 2,8 + 2,8 + 3,5 + 3,5 + 3,6 +3,8)/10 = 28,6/10 = 2,86
- varianța: = [(1,8-2,86)^2 + (1,9-2,86)^2 + (2,4-2,86)^2 + (2,5-2,86)^2 + (2,8-2,86)^2 + (2,8-2,86)^2 + (3,5-2,86)^2 + (3,5-2,86)^2 + (3,6-2,86)^2 +(3,8-2,86)^2]/(10-1) = (1,1236 + 0,9216 + 0,2116 + 0,1296 + 0,0036 + 0,0036 + 0,4096 + 0,4096 + 0,5476 + 0,8836)/9 = 4,644/9 = 0,516
- deviația standard: =SQRT(0,516) = 0,7183

Analizând rezultatele se observă că la transformarea scalei de măsură cu o constantă se modifică media aritmetică, varianța și deviația standard. Valorile seriei statistice transformate se pot obține prin înmulțirea valorilor mediei, a variației și respectiv a deviației standard cu valoarea constantei utilizată la transformarea scalei de măsură (în cazul nostru c = 0,001) pentru media aritmetică și respectiv deviația standard și respectiv cu pătratul constantei pentru varianță.

Media aritmetică și deviația standard sunt cele mai utilizate mărimi de centralitate și dispersie raportate în literatura de specialitate. Alegerea utilizării lor se face în cazul în care datele sunt normal distribuite.

Pentru datele care sunt normal distribuite (media aritmetică, mediana și modulul au aproape aceeași valoare), deviația standard are o semnificație aparte. Un anumit procent din date se regăsesc în intervalele medie ± o deviație standard, medie ± 2 deviații standard și respectiv medie ± 3 deviații standard (vezi Figura 2.4).

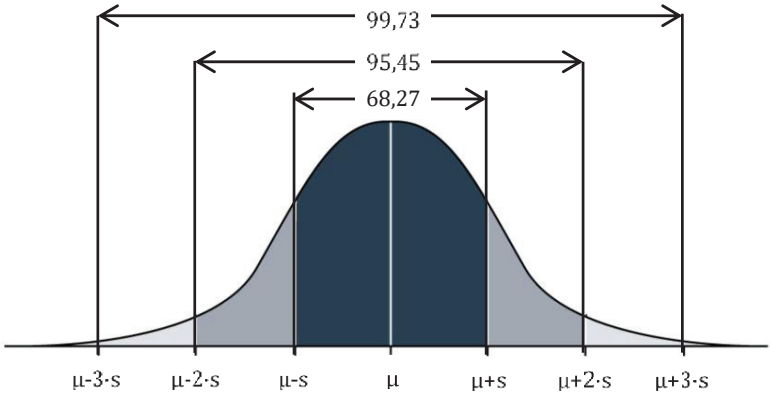


Fig. 2.4. Distribuția datelor: media (μ) și deviația standard (s)

2.5.4 Coeficientul de variație

Coeficientul de variație este o mărime a dispersiei relative introdusă de Karl Pearson și este utilizată în caracterizarea datelor cantitative.

Formula de calcul al coeficientului de variație este:

$$CV(\%) = \frac{S}{X} \times 100$$

- Coeficientul de variaţie este independent de unitatea de măsură a datelor seriei statistice.
- Calcularea are sens doar pentru variabile cantitative pe scala raport.
- Se poate utiliza pentru a compara două sau mai multe serii statistice cu unități de măsură diferite (ex. indicele de masă corporală și presiunea arterială sistolică).

Exemplul 2.15

Pentru un eșantion de 10 pacienți s-a măsurat greutatea (kg), înălțimea (cm) și presiunea arterială sistolică (PAS, mmHg). Utilizând datele obținute la măsurarea greutății și înălțimii s-a determinat valoarea indicelui de masă corporală (kg/m^3). Datele seriei statistice sunt:

i	1	2	3	4	5	6	7	8	9	10
x_i: PAS (mmHg)	220	180	100	130	120	110	140	160	120	160
y_i: IMC (kg/m^3)	28	30	21	27	25	20	30	25	24	20

Se dorește compararea variabilității presiunii arteriale sistolice cu cea a indicelui de masă corporală.

Soluție:

Pentru determinarea coeficientului de variație este necesară calcularea inițială pentru fiecare serie statistică (variabilă) a mediei și deviației standard:

x_i (PAS(mmHg)):

- media aritmetică: = (220 + 180 + 100 + 130 + 120 + 110 + 140 + 160+ 120 + 160)/10 = 1440/10 = 144 (mmHg)
- varianța: [(220-144)^2 + (180-144)^2 + (100-144)^2 + (130-144)^2 + (120-144)^2 + (110-144)^2 + (140-144)^2 + (160-144)^2 + (120-144)^2 + (160-144)^2]/(10-1) = (5776 + 1296 + 1936 + 196 + 576 + 1156+ 16+ 256 + 576 + 256)/9 = 12040/9 = 1337,78
- deviația standard: =SQRT(1337,78) = 36,58
- CV(%) = 36,58/144*100 = 25,40

y_i (IMC – kg/m^3):

- media aritmetică: = (28 + 30 + 21 + 27 + 25 + 20 + 30 + 25 + 24 + 20)/10 = 250/10 = 25
- varianța: = [(28-25)^2 + (30-25)^2 + (21-25)^2 + (27-25)^2 + (25-25)^2 + (20-25)^2 + (30-25)^2 + (25-25)^2 + (24-25)^2 + (20-25)^2]/(10-1) = (9 + 25 + 16 + 4 + 0 + 25 + 25 + 0 + 1 + 25)/9 = 130/9 = 14,44
- deviația standard: = $\sqrt{14,44}$ = 3,8
- CV(%) = 3,8/25*100 = 15,20

Împrăștierea relativă a presiunii arteriale sistolice este mai mare comparativ cu împrăștierea relativă a indicelui de masă corporală, indicând astfel o variabilitate mai mare a valorilor presiunii arteriale comparativ cu valorile indicelui de masă corporală.

În interpretarea coeficientului de variație se folosesc următoarele reguli empirice:

- CV<10%: populația poate fi considerată omogenă
- 10%≤ CV < 20%: populația poate fi considerată relativ omogenă
- 20%≤ CV < 30%: populația poate fi considerată relativ eterogenă

- CV ≥ 30%: populația poate fi considerată eterogenă

Limitele utilizării coeficientului de variație sunt:

- calcularea coeficientului de variație nu are sens dacă scala de măsură a variabilei cantitative este interval.
- dacă datele nu au distribuție normală parametrul de centralitate indicat a fi raportat este mediana ca mărime de centralitate, iar ca alternativă la coeficientul de variație se recomandată utilizarea intervalului dintre cvartila 1 și cvartila 3 (coeficientul de variație al cvartilelor):

$$CQV(\%) = (Q3-Q1)/(Q3+Q1)*100$$

2.5.5 Eroarea standard

Eroarea standard este o mărime a dispersiei utilizată în statistica inferențială, mai specific în calcularea intervalului de încredere asociat estimatorului punctual, iar notaţia ei a fost introdusă de Carl Friedrich Gauss în 1816.

Eroarea standard are formulă de calcul diferită pentru diferitele tipuri de variabile (ex. variabile cantitative – eroarea standard a unei medii sau a diferenţei dintre două medii; variabile calitative – eroarea standard a proporţiei). În cazul în care variabila este cantitativă formula de calcul a erorii standard este:

$$ES = \frac{S}{\sqrt{n}}$$

unde s = deviaţia standard a eşantionului, n = volumul/talia eşantionului.

2.6 Parametrii de simetrie și boltire

2.6.1 Asimetria

Asimetria indică pentru o serie statistică deviația de la simetrie și respectiv direcția acesteia (pozitivă – la dreapta sau negativă – la stânga).

Valorile mediei aritmetice, a medianei și a modulului pot aduce și ele informații cu privire la asimetria unei serii statistice (vezi Figura 2.5).

Asimetria unei serii statistice se poate evalua cu ajutorul cvartilelor. Dacă Q2-Q1 ≈ Q3-Q2 (unde '≈' se citește *aproximativ egal cu*) distribuția este aproximativ simetrică. Opus, dacă Q2-Q1 e diferită de Q3-Q2 distribuția este asimetrică. În cazul unei serii statistice cu asimetrie pozitivă valoare modului este mai mică decât valoarea medianei care la rândul ei este mai mică față de valoarea mediei aritmetice. Opus, în cazul unei serii statistice cu asimetrie negativă valoare modului este mai mare decât valoarea medianei care la rândul ei este mai mare față de valoarea mediei aritmetice.

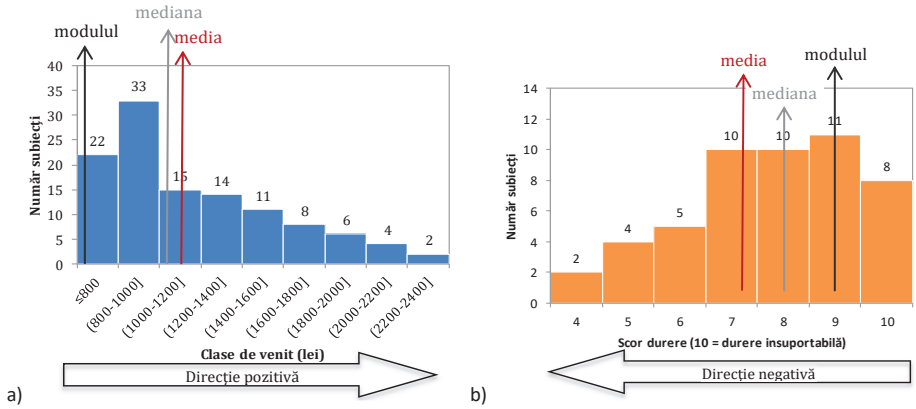


Fig. 2.5. a) asimetrie pozitivă: distribuția venitului unui eșantion de 115 subiecți; b) asimetrie negativă: distribuția scorului de durere post injectare moldamin pe un eșantion de 50 pacienți

Formula asimetriei (cunoscută și sub denumirea de moment de ordin 3) este dată de media cuburilor deviațiilor standardizate de la medie:

$$\alpha_3 = \frac{\sum_{i=1}^{n}\left[\frac{(x_i - \bar{X})}{S}\right]^3}{n}$$

Pentru interpretarea asimetriei se folosesc valorile din intervalele următoare:

- Valorile mai mici de -1 sau mai mari de 1 indică o distribuție asimetrică
- Valorile cuprinse în intervalele (-1; -0,5] respectiv [0,5; 1) indică o distribuția moderat asimetrică
- Valorile cuprinse în intervalul (-0,5; 0,5) indică o distribuție aproximativ simetrică

2.6.2 Boltirea

Boltirea este o mărime care cuantifică forma unei serii sau distribuții de date. Boltirea măsoară înălțimea unei distribuții în comparație cu o distribuție normală. Boltirea mai este cunoscută și ca moment de ordin 4, iar valoarea pentru o distribuție normală standard este egală cu 3. Din acest motiv, programele statistice calculează excesul de boltire dat de formula:

$$\alpha_4 = \frac{\frac{1}{n}\cdot\sum_{i=1}^{n}(x_i - \bar{X})^4}{S^4} - 3$$

Interpretare valorii boltirii (figura 2.6):

- Distribuția cu boltirea ≅ 3 (excesul de boltire = 0) se numește mezocurtică.
- Boltirea <3 (excesul de boltire < 0, valoare negativă) indică o distribuție platicurtică.
- Boltirea >3 (excesul de boltire > 0) indică o distribuție leptocurtică.

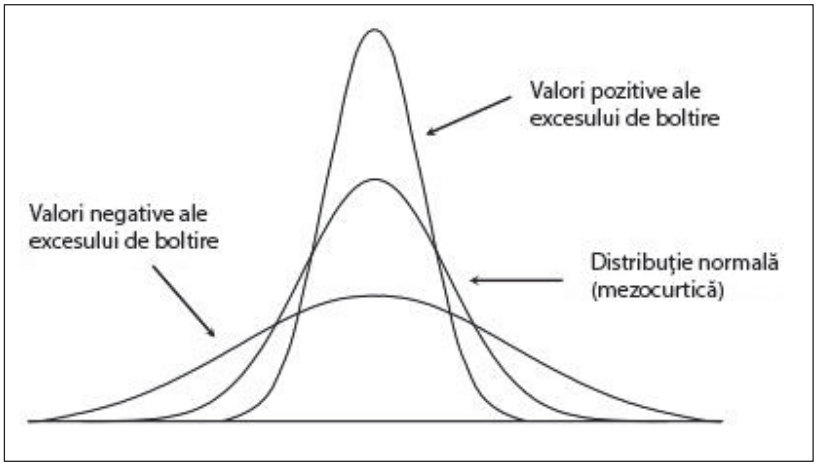


Fig. 2.6. Boltirea pentru diferite distribuții

2.6.3 Alegerea parametrilor de centralitate și dispersie

Parametrii de centralitate și dispersie se pot calcula pe orice tip de variabilă indiferent de scala de măsură. Dar calcularea lor are sens doar pentru anumite variabile și respectiv scale de măsură (vezi Tabelul 2.4).

Tabelul 2.4. Tipul variabilei, al scalei de măsură și alegerea parametrilor descriptivi

Parametrul	Date calitative		Date cantitative	
	Nominale	Ordinale	Interval	Raport
Media	Nu	Nu	Da*	Da*
Mediana	Nu	Da	Da	Da
Modulul	Da	Da**	Da**	Da**
Amplitudinea	Nu	Da	Da	Da
Deviația standard	Nu	Nu	Nu	Da*

* Dacă datele sunt normal distribuite

** Nu este recomandat

2.7 Reprezentarea tabelară a datelor medicale

Tabelul este cea mai simplă metodă de sumarizare a datelor unei serii statistice și se poate face atât pentru datele calitative, cât și pentru datele cantitative. Metodele de sumarizare tabelară cele mai des utilizate sunt reprezentate de tabelul de frecvență și tabelul de contingență și vor fi prezentate în această secțiune.

Prezentarea datelor brute nu va duce la o interpretare corectă a acestora, interpretarea fiind cu atât mai dificilă cu cât numărul de date din seria statistică este mai mare. Sumarizarea tabelară permite reducerea datelor și prezentarea acestora într-o formă cât mai simplă posibilă care însă să ajute la identificarea (dacă există) tiparului pe care datele îl urmează.

Există o serie de avantaje ale sumarizării datelor cu ajutorul tabelele:

- Rezultatele sunt prezentate într-o formă compactă.
- Rezultatele sunt ușor de interpretat.
- Permite observarea existenței unor modele.
- Permite compararea a două sau mai multe grupuri.
- Permite utilizarea parametrilor cantitativi.

Bineînțeles că acest proces nu este lipsit de dezavantaje. Detaliile datelor brute se pierd în momentul sumarizării și acest proces este ireversibil.

Indiferent de forma sumarizării tabelare, tabelele trebuie să aibă următoarele caracteristici:

- Trebuie să se explice singure. Citirea tabelului trebuie să fie posibilă fără a citi textul care însoțește tabelul. Astfel:
    - Semnificațiile abrevierilor și a simbolurilor utilizate în tabel trebuie definite la subsolul acestuia.
    - Rândurile și coloanele trebuie definite. De asemenea, trebuie specificat, acolo unde este cazul, care sunt unitățile de măsură utilizate.
    - Tabelul trebuie să aibă un titlu clar și la subiect.
    - Este recomandat ca tabelele să aibă rânduri și/sau coloane de sinteză (cu totaluri).
- Dacă datele nu sunt ale dvs. trebuie să obțineți acordul de utilizare al acestora de la persoana sau instituția care deține drepturile de proprietate intelectuală. În acest caz aveți, de asemenea, obligația de a menționa la subsolul tabelului sursa datelor.

2.7.1 Tabelul de frecvență

Tabelul de frecvență se utilizează pentru sumarizarea unei singure variabile (serie statistică unidimensională sau univariată).

Exemplul 2.16.

Greutatea la naștere a primilor 120 copii cu restricție de creștere intrauterină incluși în studiu sunt redate mai jos:

740	890	1600	1600	2100	1800	2150	2140	2940	2400	1900	2460
900	990	1560	1900	1990	1660	2200	2100	2400	2490	2460	2600
990	1170	1490	1300	1870	2400	2300	2300	1900	1780	2600	2200
950	1249	1300	1560	2160	1990	2150	1850	2400	2200	2600	2300
780	1100	1400	1100	1800	2300	2200	2550	2200	2330	2490	2480
890	1300	1450	1900	1990	2300	2200	2400	1990	1800	2600	2100
900	1100	1970	1900	990	2700	2400	2500	2000	2450	2500	2650
960	1350	1490	1900	2250	2100	2350	2000	2500	1900	2400	2490
900	1350	1550	2000	1680	2280	2100	2300	2200	2490	2200	2700
800	1225	1250	2100	1600	2430	1900	2490	2450	2700	2700	2650

Care este forma cea mai potrivită de sumarizare tabelară a acestor date?

Datele brute din tabelul anterior pot fi sumarizate cu ajutorul tabelului de frecvență.

Tabelul de frecvență se poate utiliza atât la sumarizarea datelor calitative (nominale sau ordinale) cât și a celor cantitative (discrete sau continue). Dacă datele sunt cantitative continue, înainte de sumarizare trebuie stabilite clasele de frecvență. Clasele de frecvență se pot stabili în funcție de diferite criterii, putând fi alese arbitrar sau după criterii cu semnificație clinică (ex. greutate la naștere mică < 2500g, foarte mică < 1500g, extrem de mică < 1000g).

Organizarea în tabelul de frecvență a datelor se face prin includerea valorilor distincte pe care variabila le poate lua (pentru variabilele calitative, scala nominală sau ordinală, respectiv pentru datele cantitative discrete) sau respectiv clasa de frecvență (pentru variabilele cantitative continue – se aplică obligatoriu, dar se poate aplica și celor discrete) și respectiv a frecvenței de apariție. În construirea unui tabel de frecvență, frecvențele raportate pot fi:

- Absolute: număr de cazuri care iau valoarea de interes sau au valoare în clasa de frecvență de interes.
- Relative: numărul de cazuri/total cazuri.
- Cumulate (crescător/descrescător):
    - Frecvența absolută cumulată crescător asociată unei valori *x* este egală cu suma frecvențelor absolute ale tuturor valorilor seriei care sunt mai mici sau egale cu valoarea *x*.

- Frecvența absolută cumulată descrescător asociată unei valori *x* este egală cu suma frecvențelor absolute ale tuturor valorilor seriei care sunt mai mari sau egale cu valoarea *x*.
- Frecvența relativă cumulată crescător asociată unei valori *x/n* este egală cu suma frecvențelor relative ale tuturor valorilor seriei care sunt mai mici sau egale cu valoarea *x*.
- Frecvența relativă cumulată descrescător asociată unei valori *x/n* este egală cu suma frecvențelor relative ale tuturor valorilor seriei care sunt mai mari sau egale cu valoarea *x*.

Datele brute din Exemplul 2.16. se vor sumariza folosind următoarele clase de frecvență: greutate normală (≥ 2500g), greutate mică (de la 1500g la 2499g), greutate foarte mică (de la 1000g la 1499g) și respectiv greutate extrem de mică (<1000g):

Clasa de frecvență	Frecvența absolută	Frecvența relativă (%)	Frecvența absolută cumulată crescător	Frecvența relativă cumulată crescător
<1000g	13	=13/120*100 = 11	=13	=11
1000-1499	16	=16/120*100 = 13	= 16+13 = 29	=13+11 = 24
1500-2499	76	= 76/120*100 = 63	= 76+29 =105	= 63+24 = 87
≥2500	15	=15/120*100 = 13	= 15+105 = 120	=13+87 = 100
Total	120	100		

Analizând tabelul de frecvență anterior putem caracteriza seria statistică:

- Majoritatea subiecților cu restricție de creștere intrauterină au greutate mică la naștere.
- Eșantionul studiat conține un număr/procent similar de nou născuți cu greutate mică la naștere și respectiv greutate considerată în limite normale (13%).
- Numărul de copii cu greutate mai mică decât normală la naștere a fost egal cu 105.
- 24% din nou născuții din eșantion au avut greutate foarte mică sau extrem de mică la naștere.

Clasele de frecvență se pot scrie și cu ajutorul parantezelor rotunde și drepte, fiecare cu semnificație diferită. Dacă se utilizează parantezele drepte valorile sunt incluse în clasa respectivă de frecvență; dacă se utilizează paranteze rotunde valoarea nu este inclusă în clasa de frecvență respectivă. Clasificarea severității obezității prin utilizarea indicelui de masă corporală utilizează următoarele clase de frecvență:

Descriere	**Clase de frecvență (kg/m^2)**
Subponderal	<18,5
Greutate normală	[18,5–25)
Supraponderal	[25–30)
Obezitate moderată (obezitate clasa I)	[30–35)
Obezitate severă (obezitate clasa II)	[35–40)
Obezitate foarte severă (obezitate clasa II)	≥ 40

Clasele de frecvență trebuie construite în așa fel încât o valoare a seriei statistice să aparțină unei singure clase (de ex. o valoarea de 25kg/m^2 aparține doar clasei de supraponderali, deoarece în această clasă paranteza este dreaptă).

Tabelul de frecvență se construiește similar și dacă datele sunt calitative (ordinale sau nominale).

Exemplul 2.17.

Scorul de durere la prezentarea la fizioterapie a 100 pacienți cu artroză a fost colectate prin utilizarea scalei vizuale de durere (0=fără durere, 10 = durere extrem de puternică). Datele brute colectate sunt prezentate mai jos:

9	8	9	6	5	8	10	10	10	9
7	10	9	10	10	8	8	10	8	10
8	9	9	8	6	9	8	7	8	10
8	8	9	10	10	8	10	7	9	8
8	8	9	9	7	6	10	6	10	9
9	9	8	8	8	7	8	6	8	9
10	9	7	10	10	9	7	10	9	7
8	8	6	7	10	6	7	9	7	8
9	7	8	9	9	9	8	9	8	9
7	7	8	9	8	9	9	10	9	8

Realizați pentru aceste date tabelul de frecvență pentru a evidenția frecvențele absolute cumulate crescător și descrescător.

Soluție:

Tabelul de frecvență asociat seriei statistice din Exemplul 2.17:

Scor durere	Nr. cazuri	%	Frecvența absolută cumulată crescător	Frecvența absolută cumulată descrescător
0	0	0	=0	=100-0=100
1	0	0	=0+0	=100-0=100
2	0	0	=0+0	=100-0=100
3	0	0	=0+0	=100-0=100
4	0	0	=0+0	=100-0=100
5	1	1	=1+0=1	=100-1=99
6	7	7	=7+1=8	=99-7=92
7	14	14	=14+8=22	=92-14=78
8	29	29	=29+22=51	=78-29=49
9	29	29	=29+51=80	=49-29=20
10	20	20	=20+80=100	=20-20=0
Total	100	100		

Analiza datelor din tabelul de frecvență evidențiază următoarele:

- Pacienții cu artroză au un scor al durerii de la moderat (x_{min} = 5) la durere extrem de puternică (x_{max} = 10).
- a99% din pacienții au un scor al durerii mai mare sau egal cu 5.

- 50% din pacienți resimt o durere puternică (scor durere = 8) sau foarte puternică (scor durere = 9).
- 100% din pacienți au scorul durerii mai mic sau egal cu 10, ceea ce indică absența datelor lipsă.

Tabelul de frecvență se poate utiliza pentru sumarizarea datelor sau pe baza lui se poate realiza reprezentarea grafică a rezultatelor.

2.7.2 Tabelul de contingență

Tabelul de contingență permite sumarizarea în același timp a mai mult de o variabilă, cu scopul de a se observa relația dintre variable. Cu cât numărul de variabile e mai mare complexitatea tabelului crește.

Forma generală a tabelului de contingență este r×c unde *r* este numărul de coloane și *c* este numărul de rânduri. Pe coloane, respectiv rânduri se va sumariza una sau mai multe variabile. Numărul de coloane și respectiv rânduri este dat de tipul variabilei sumarizate. Dacă variabila este calitativă ordinală sau nominală, numărul de coloane/rânduri este egal cu numărul de valori posibile pe care variabila de interes le ia în eșantion. În cazul în care variabilele de interes sunt dicotomiale (pot lua doar două valori: Da/Nu, Sănătos/Bolnav, Prezent/Absent, etc.) tabelul de contingență devine de 2×2, tabel frecvent utilizat în studiile medicale.

Exemplul 2.18.

Fie seria statistică formată din gen (F/M) și valoarea scorului de durere a 100 pacienți cu artroză:

F	9	F	9	M	4	M	9	F	7	F	10	M	9	F	10	M	9	M	6
M	9	M	6	F	4	M	5	F	7	F	5	F	9	F	9	M	8	F	9
M	7	M	9	M	5	M	8	F	6	F	9	F	8	M	7	F	9	F	8
M	10	M	8	M	9	M	8	M	6	F	10	F	8	F	8	F	9	F	8
M	10	M	7	M	5	F	10	F	9	F	8	F	8	F	7	F	6	F	9
F	7	F	8	M	8	M	8	F	10	F	8	M	8	M	7	F	8	F	7
F	10	M	9	M	8	M	7	M	7	F	5	F	2	F	9	M	6	M	8
M	8	M	10	M	9	M	10	M	5	F	7	F	4	F	10	F	10	M	6
M	7	F	9	M	10	M	8	M	6	F	9	F	10	F	10	M	7	M	7
M	7	F	8	M	9	M	7	F	6	F	9	F	10	M	8	F	8	F	9

Realizați tabelul de contingență pentru această serie statistică.

Soluția:

Tabelul de contingență asociat acestei serii statistice se numește tabel de contingență observat (construit pe baza datelor colectate în urma studierii eșantionului) și este de 8×2 deoarece scorul de durere ia valori de la 2 la 10, iar genul poate fi F (feminin) sau M (masculin).

Gen / Scor durere	F	M	Total
2	1	0	1
4	2	1	3
5	2	4	6
6	3	6	9
7	6	11	18
8	12	12	24
9	15	9	24
10	11	5	16
Total	52	48	100

Tabelul de contingență prezentat anterior prezintă frecvențele absolute (număr subiecți) ale subiecților care îndeplinesc concomitent două criterii (au un anumit scor al durerii și respectiv un anumit gen). În cazul tabelelor de contingență este util să avem totalurile atât pe rânduri cât și pe coloane. Astfel:

- Avem 52 de subiecți de gen feminin și 48 de subiecți de gen masculin.
- Scorul cel mai mic de durere este resimțit de un pacient de gen feminin.
- În eșantionul studiat există 11 pacienți de gen feminin și 5 pacienți de gen masculin care resimt o durere extrem de puternică (scor durere = 10).

Tabelul de contingență își regăsește utilitatea atât în descrierea datelor, cât și în inferența statistică, unde pe baza tabelului de contingență observat se construiește tabelul de contingență așteptat (sau teoretic) și se compară dacă există diferență semnificativă statistic între tabelul observat și așteptat, putându-se astfel trage concluzii în ceea ce privește existența sau nu a dependenței dintre variabilele din tabelul de contingență (vezi capitolul *Compararea variabilelor calitative*).

Tabelul de contingență de 2×2 sumarizează două variabile dicotomiale. A fost studiat un eşantion de 440 subiecţi, 104 cu obezitate şi 137 cu diabet. 43 din subiecţii cu obezitate 43 prezentau şi diabet. Reprezentarea tabelară a datelor observate este redată în Tabelul 2.5.

Tabelul 2.5. Obezitate versus diabet: tabel de contingență de 2×2

	Diabet=prezent	Diabet=absent	Total rând
Obezitate=prezentă	43	61	104
Obezitate=absentă	94	242	336
Total coloană	137	303	440

Așa cum a fost specificat anterior, tabelul de contingență poate să sumarizeze mai mult de două variabile (vezi Exemplul 2.19).

Exemplul 2.19.

Fie seria statistică trivariată formată din variabilele gen (F/M), obezitate (da/nu) și fumat (da/nu). Au fost colectate datele brute pentru un eșantion de 440 de subiecți, iar rezultatele sunt sumarizate în tabelul următor:

Gen	Fumat	Obezitate = prezentă	Obezitate = absentă	Total rânduri
F	da	3	28	**31**
	nu	7	172	**179**
Total F		***10***	***200***	***210***
M	da	33	39	**72**
	nu	61	97	**158**
Total M		***94***	***136***	***230***
Total coloane		**104**	**336**	**440**

Așa cum se poate observa din analiza datelor prezentate în tabelul anterior (Exemplul 2.19.) complexitatea a crescut, iar informația din tabel e destul de greu de citit. Analizând tabelul nu putem să vedem fără a face alte calcule dacă, de exemplu, frecvența fumătorilor în rândul subiecților de gen feminin cu obezitate este sau nu mai mare comparativ cu frecvența fumătorilor de gen masculin obezi.

2.8 Reprezentarea grafică a datelor medicale

Reprezentarea grafică a datelor medicale este metoda care permite prezentarea rapidă și cu un bun impact vizual a rezultatelor. Ca și la reprezentarea tabelară a datelor rezultate din cercetarea medicală, reprezentarea grafică trebuie să:

- se explice singură,
- aibă titlul concis și informativ,
- să prezinte definiția axelor și a unităților de măsură (acolo unde este cazul),
- să prezinte legende explicative (dacă este cazul),
- să prezinte sursa datelor dacă datele nu sunt personale și acordul de publicare în conformitate cu dreptul de proprietate intelectuală.

Există o serie de recomandări care trebuie urmate în crearea reprezentărilor grafice. Evergreen și Emery au publicat în 2014 un ghid de evaluare a unei reprezentări grafice care cuprinde următoarele criterii:

- Textul. Deoarece reprezentarea grafică nu trebuie să conțină mult text, textul care însoțește reprezentarea grafică trebuie să fie concis și la subiect.
  - Titlul: este indicat să fie concis (6-12 cuvinte) și specific (conține mesajul pe care reprezentarea grafică îl transmite).
  - Subtitlul și/sau adnotările: au scopul de a aduce informații explicative și de a crește puterea de explicare a graficului.

- Mărimea textului: titlul trebuie să fie mai mare decât subtitlul care trebuie să fie mai mare decât definiția axelor. Cea mai mică dimensiune trebuie să fie de 9 pentru graficele care sunt inserate în text, respectiv 20 pentru graficele din prezentările PowerPoint.
- Orientarea textului: cu excepția definițiilor de axe, textul din reprezentarea grafică trebuie să fie orizontal.

- Gama de culori:
  - Culorile se folosesc pentru a pune în evidență modele.
  - Utilizarea culorilor permite diferențierea și în cazul în care tipărirea se face alb-negru.
  - Atenție la culorilor a două sau mai multe clase alăturate: evitați combinațiile roșu-verde, galbe-albastru.
  - Contrastul dintre fundalul graficului și culorile utilizate în reprezentarea grafică trebuie să fie maxim. Se recomandă folosirea negrului sau a culorilor închise pentru text și alb sau culori pastelate transparente pentru fundal.
- Fundalul și subdiviziunile graficului:
  - Excesul de linii (ex. subdiviziuni ale axelor, etc.) dau impresia de dezordine. Eliminați-le dacă nu sunt necesare în interpretarea graficului.
  - Dacă subdiviziunile axelor sunt necesare, acestea trebuie să fie cu gri nu cu negru.
  - Graficul trebuie să facă parte integrantă din document, așa încât nu este recomandată utilizarea delimitării acestuia cu linii.
  - Reprezentarea grafică este prin definiția bidimensională (o axă OX și o axă OY). Evitați utilizarea unei axe OY suplimentare sau reprezentarea grafică tridimensională.
- Mesajul. Aranjarea necorespunzătoare a datelor în reprezentările grafice în cel mai bun caz pot induce confuzie, iar în cel mai rău caz pot induce în eroare cititorul. Realizați reprezentarea grafică astfel încât vizualizarea datelor să fie ușoară.
  - Datele se ordonează astfel încât mesajul transmis să fie logic. Ordonarea se poate face în funcție de frecvența categoriilor (ex. de la cea mai mare la cea mai mică pentru datele ordinale), clasele de frecvență (ex. histograma), perioadele de timp (ex. graficul de tip linie), alfabetic, etc.
  - Nu se încarcă reprezentarea grafică cu desene sau forme geometrice în scop decorativ.
  - Reprezentarea grafică trebuie să ilustreze rezultatele semnificative sau concluziile studiului.
  - Se utilizează reprezentarea grafică care se potrivește cel mai bine tipului de date.
  - Se utilizează nivelul de precizie potrivit. De exemplu, graficul de tip bare atunci când precizia este importantă respectiv graficul de tip plăcintă sau cerc atunci când precizia nu este așa de importantă.
  - Elementele individuale ale reprezentării grafice (ex. textul, aranjamentul, culorile, etc.) trebuie să contribuie la consolidarea mesajului care se dorește a fi transmis.

Reprezentarea grafică a datelor se face prin alegerea acelei reprezentări care transmite cel mai bine mesajul de interes. Întotdeauna există mai mult decât o alternativă în realizarea reprezentării grafice.

Același tip de reprezentare grafică poate fi utilizată pentru ilustrarea unei variabile calitative sau cantitative. În această secțiune vor fi prezentate cele mai utilizate modalități de reprezentare grafică a variabilelor.

În principiu, pentru a putea reprezenta grafic o variabilă calitativă sau o variabilă cantitativă cu ajutorul claselor de frecvență utilizând Microsoft Excel, trebuie inițial realizat tabelul de frecvență pentru variabila de interes.

Exemplul 2.20.

Au fost investigate caracteristicile antropometrice ale nou născuților cu restricție de creștere intrauterină prin evaluarea genului (F/M), greutății (g), lungimii (cm), a perimetrului cranian (cm) și a traumatismului la naștere. Studiul a fost de tip caz-martor cu potrivire în ceea ce privește genul și vârsta de gestație.

Care sunt reprezentările grafice potrivite pentru ilustrarea datelor din acest exemplu?

2.8.1 Graficul de tip sectorial

O diagramă de tip sectorial (Pie/Doughnut) este o reprezentare grafică circulară (*Pie*) sau de tip 'gogoașă' (*Doughnut*) utilizată pentru a vizualiza părți ale întregului. Nu se regăsește foarte frecvent în literatura de specialitate, dar este frecvent utilizată în prezentarea rezultatelor în tezele de licență și respectiv disertație. În graficul de tip *Pie* se reprezintă o singură variabilă (figura 2.7).

Exemplul 2.20.1.

Reprezentați grafic distribuția pe gen a subiecților incluși în studiul din Exemplul 2.20.

Indiferent de tipul de grafic utilizat pentru a reprezenta distribuția pe gen avem nevoie de tabelul de frecvență.

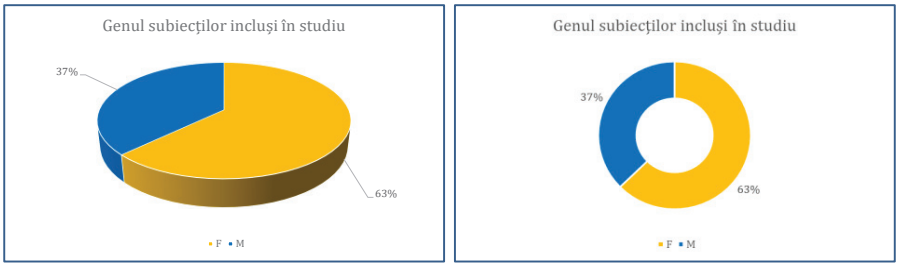


Fig. 2.7. Reprezentarea grafică a distribuției genului în eșantionul studiat: stânga: grafic de tip circular (*Pie*); dreapta: grafic de tip 'gogoașă' (*Doughnut*)

O reprezentare grafică sectorială aparte este graficul de tip Sectorial-Sectorial *(Pie to Pie*, exemplul 2.21, figura 2.8b) sau Coloană-Sectorial (*Bar to Pie*, figura 2.8c), reprezentare în care categoriile care au procentele mici sunt încadrate într-o singură categorie și reprezentate grafic distinct.

Exemplul 2.21.

A fost evaluată modalitatea de tratament primit în urgență pe un eșantion de subiecți cu patologii cardiace. Datele sunt sumarizate în Tabelul 2.5.

Tabelul 2.5. Modalitatea de aplicare a tratamentului (iv = intravenos, po = per oral) în funcție de diagnostic și gen

Diagnostic	Tratament iv			Tratament po			Total
	F	M	Total iv	F	M	Total po	
Cardiomiopatie dilatativă	1	1	2	0	0	0	2
Cardiopatie ischemică	7	5	12	12	13	25	37
Hipertensiune arterială	5	3	8	10	12	22	30
Infarct miocardic	1	0	1	0	0	0	1
Total	14	9	23	22	25	47	10

Reprezentați grafic distribuția diagnosticului în serviciul de urgență pe eșantionul studiat.

În acest caz avem două categorii de diagnostic cu puține cazuri (cardiomiopatia dilatativă şi respectiv infarctul miocardic), iar utilizarea unui grafic sectorial simplu va împărți întregul în 4 părți, vizualizarea fiind astfel dificilă (Figura 2.8a). În acest caz, se poate utiliza graficului sectorial de tip *Pie to Pie* (Figura 2.8b) sau *Pie to Bar* (Figura 2.8c).

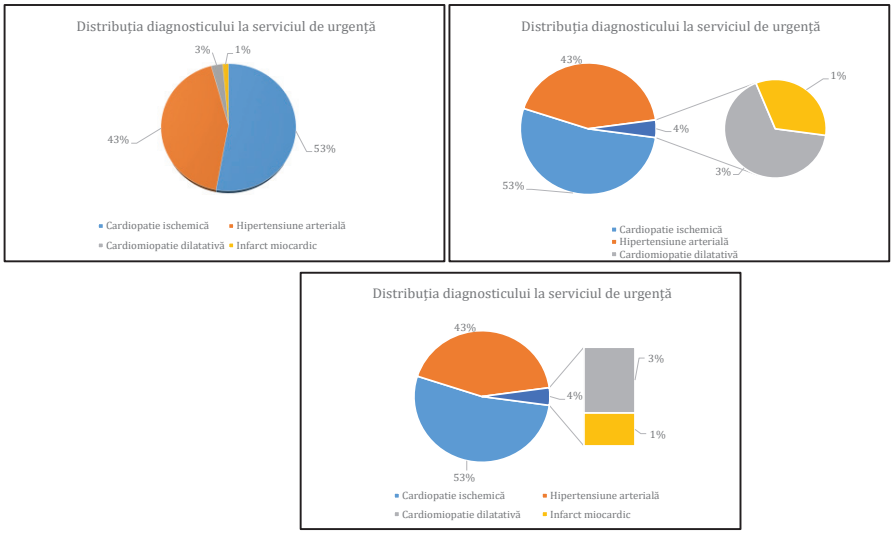


Fig. 2.8. Reprezentarea grafică a distribuției diagnosticului: a) (stânga sus) grafic sectorial; b) (dreapta-sus) grafic de tip Pie to Pie; c) (jos) grafic de tip Pie to Bar

Graficul de tip *Doughnut* permite reprezentarea a mai mult de o variabilă (vezi Exemplul 2.22).

Exemplul 2.22.

Scorul de durere la trei pacienți cu artroză care s-au prezenta pentru efectuarea tratamentului fizioterapic au fost înregistrate la prezentare, după prima cură de fizioterapie (post-FT1) şi respectiv după a doua cură de fizioterapie (post-FT2). Datele sunt sumarizate în Tabelul 2.6.

Tabelul 2.6. Efectul fizioterapiei asupra durerii la pacienţii cu artroză. Scorul de durere măsurat pe scala vizuală (0 = fără durere, 10 = durere extrem de puternică)

Vârsta (ani)	Genul	Scor durere		
		Iniţial	Post-FT1	Post-FT2
42	F	10	10	6
52	F	10	9	8
65	F	9	8	2

Reprezentaţi grafic aceste date.

Avem de reprezentat grafic datele care aparţin la 3 subiecţi fiecare având 3 observaţii. O alternativă de reprezentare grafică, nu neapărat cea mai reprezentativă, este graficul de tip *Doughnut* (Figura 2.9).

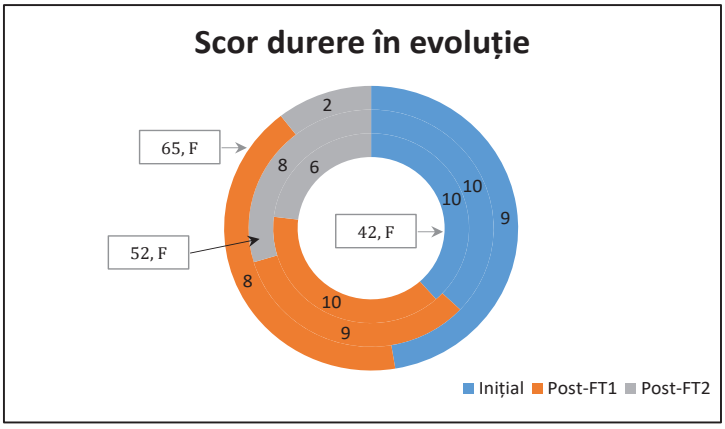


Fig. 2.9. Evoluţia scorului de durere post-fizioterapie la 3 pacienţi cu artroză

Din figura 2.9 se poate observa că scorul de durere a fost la evaluarea iniţială în categoria extrem de puternică la toţi subiecţii, la doi din trei subiecţi durerea scade după prima cură de fizioterapie, urmată de scădere după cea de-a doua cură de fizioterapie. Această evoluţie este însă greu de vizualizat pe acest tip de reprezentare grafică.

2.8.2 Graficul de tip coloane/bare

Graficul de tip coloane (bare verticale) sau bare (orizontale) este utilizat pentru reprezentarea diferitelor categorii, având aplicabilitate în reprezentarea uneia sau mai multor variabile. Înălțimea coloanei/barei este egală cu cantitatea din categoria dată. Graficul de tip coloane se utilizează dacă textul asociat denumirii coloanei este scurt (vezi Figura 2.10a); în caz contrar se utilizează graficul de tip bare (vezi Figura 2.10b).

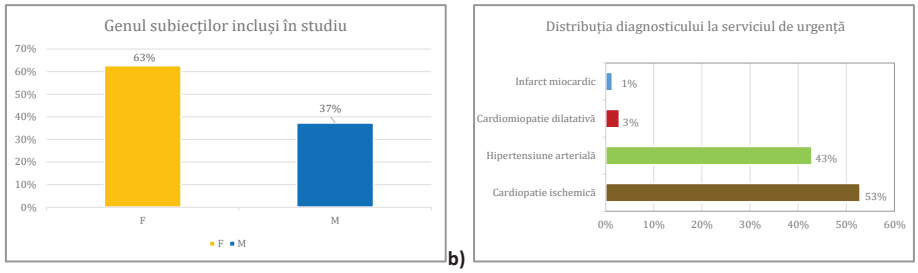


Fig. 2.10. Reprezentarea grafică de tip: a) coloane (datele din Exemplul 2.19.1, o singură variabilă calitativă dicotomială); b) bare (datele din Exemplul 2.20, o singură variabilă calitativă nominală)

Într-un grafic de tip coloane se pot reprezenta una (figura 2.10) sau mai multe variabile (figura 2.11).

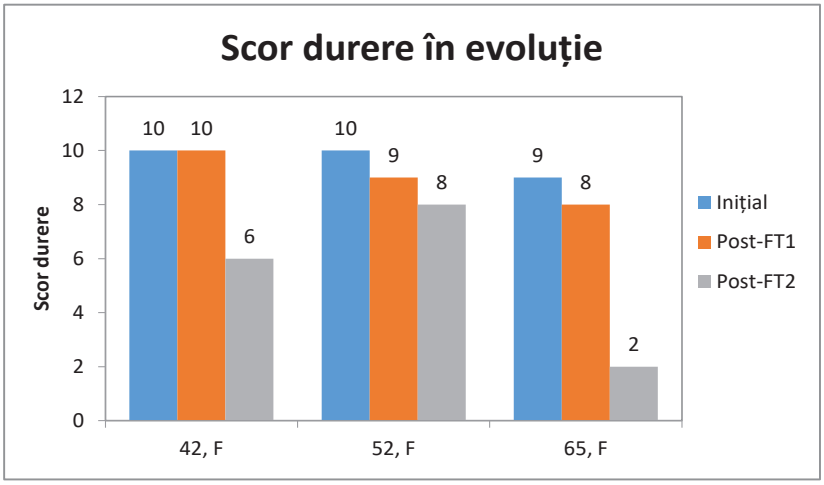


Fig. 2.11. Evoluţia scorului de durere pentru datele din exemplul 2.22

Graficul de tip coloane permite prezentarea frecvenţelor absolute (nr. de cazuri, figura 2.11) sau relative (figura 2.12, *bare suprapuse reprezentate procentual până la 100% - stacked bar*).

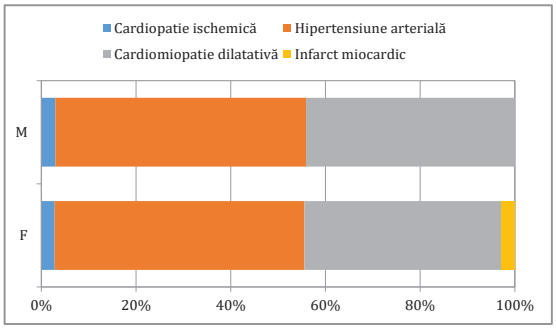


Fig. 2.12. Diagnosticul în funcţie de gen pentru datele dineExemplul 2.21

2.8.3 Histograma

Reprezentarea grafică a distribuției frecvențelor valorilor din cadrul unei serii statistice se face cu ajutorul histogramei. Histograma este un grafic de tip coloane în care nu există spaţiu între coloane. Axa OX a histogramei este dată de clase corespunzătoare claselor de frecvenţă. Valoarea fiecărei clase de frecvență este dată de o bară verticală. Histograma poate fi construită cu clase egale sau ne-egale respectiv prin utilizarea frecvenţelor absolute sau relative.

Exemplul 2.20.2.

Reprezentaţi grafic distribuţia greutăţii la naştere a subiecţilor cu restricţie de creştere intrauterină din exemplul 2.20.

Reprezentarea grafică se va face pentru următoarele clase de frecvenţă exprimate în grame: ≤1000; (1000-1500]; (1500-2500] şi >2500 (figura 2.12). Putem opta să reprezentăm grafic frecvenţele absolute (numărul de cazuri) sau relative (%) pentru fiecare clasă în parte.

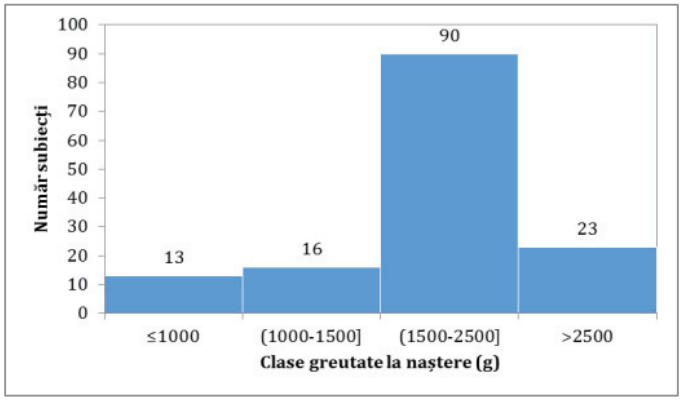


Fig. 2.13. Histograma greutăţii la naştere a copiilor cu restricţie de creştere intrauterină din exemplul 2.20

2.8.4 Grafice de tip poligon de frecvenţă

Reprezentarea grafică a distribuţiei unei serii statistice cantitative se poate face şi cu ajutorul poligonului de frecvenţă. Spre deosebire de histogramă, poligonul de frecvenţă permite reprezentarea simultană a două grupuri folosind aceleaşi clase de frecvenţă. Această modalitate de reprezentare grafică prezintă un interes deosebit deoarece: forma sa dă o imagine globală a distribuţiei observate (care poate fi "comparată" cu o distribuţie teoretică) și permite vizualizarea observaţiilor "aberante" (apărute prin erori sistematice de codificare şi de măsurare sau introduse de subiecţi cu anumite particularităţi) chiar dacă acestea sunt atenuate prin gruparea în clase.

Exemplul 2.20.3.

Reprezentaţi grafic distribuţia greutăţii la naştere a subiecţilor cu (grup caz) şi respectiv fără (grup martor) restricţie de creştere intrauterină din exemplul 2.20.

Reprezentarea grafică se va face utilizând următoarele clase de frecvenţă: 1000, (1000-1500], (1500-2000], (2000-2500], (2500-3000], (3000-3500], >3500 (vezi figura 2.14).

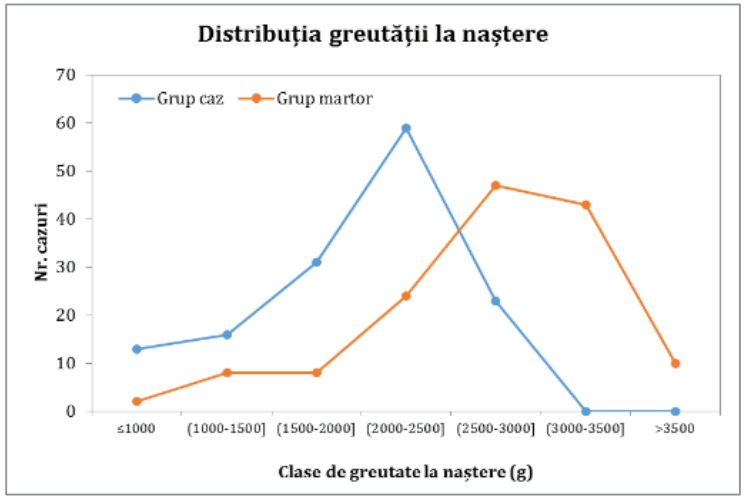


Fig. 2.14. Poligoane de frecvenţă pentru variabila greutate la naştere pentru nou-născuţii cu (Grup caz) şi fără (Grup martor) restricţie de creştere intrauterină din exemplul 2.20

2.8.5 Grafice de tip cutie cu mustăți

Graficele de tip cutie cu mustăți, *boxplot* în limba engleză, asociază categoriilor reprezentate segmente de dreaptă și dreptunghiuri cu diferite semnificaţii aparţinând parametrilor statistici descriptivi. În cazul acestui tip de reprezentare grafică, putem avea o variabilă cantitativă pentru un eşantion (rar utilizat) sau o variabilă cantitativă pentru mai mult de un grup, caz în care avem variabila cantitativă de interes şi o variabilă de grupare

(calitativă). Sunt frecvent utilizate pentru a arăta diferenţele între grupuri în ceea ce priveşte o variabilă cantitativă.

Exemplul 2.20.4.

Reprezentaţi grafic utilizând diagrama de tip *boxplot* greutatea la naştere a subiecţilor cu şi respectiv fără restricţie de creştere intrauterină din Exemplul 2.20.

Obţinerea acestui tip de reprezentare grafică este foarte uşoară cu ajutorul programelor dedicate de statistică dar există și implementări în ultimele versiuni ale Microsoft Excel (vezi figura 2.15).

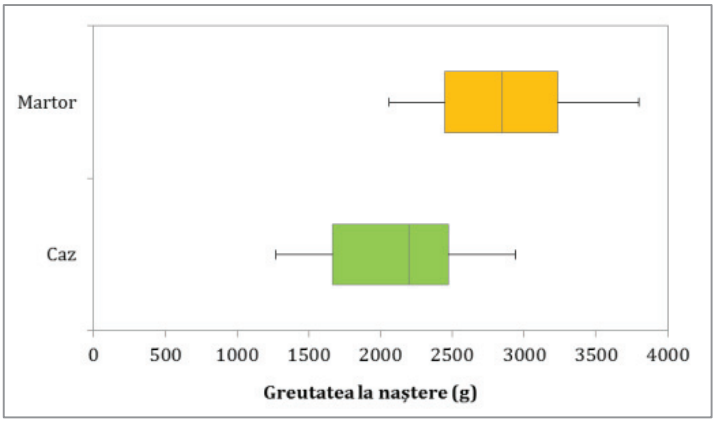


Fig. 2.15. Grafic de tip boxplot pentru variabila greutate la naştere pentru nou-născuţii cu (grup caz) şi fără (grup martor) restricţie de creştere intrauterină din exemplul 2.20. Valoarea din mijlocul dreptunghiului este mediana, dreptunghiul este dat de valorile cvartilei 1 şi respectiv 3 iar barele extreme sunt date de valoarea minimă şi respectiv maximă

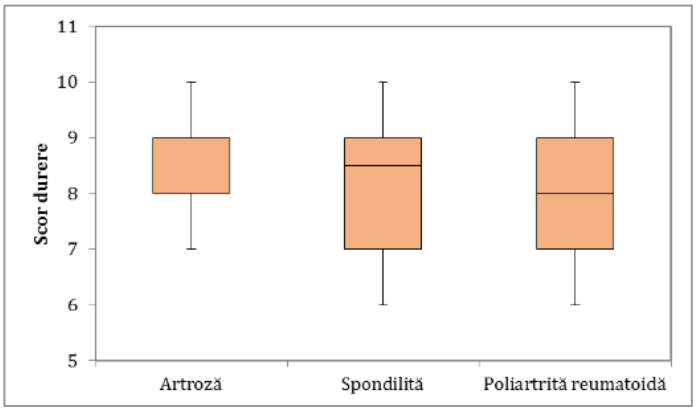


Fig. 2.16. Grafic de tip boxplot pentru scorul durerii (scala de la 0 = fără durere la 10 = durere extrem de puternică) la pacienţii cu artoză, spondilită anchilozantă şi respectiv cu poliartrită reumatoidă. Valoarea din mijlocul dreptunghiului este mediana şi lipseşte la pacienţii cu artroză deoarece valoarea medianei este egală cu valoarea cvartilei 1

Pentru utilizarea acestui tip de grafic se utilizează media dacă datele cantitative urmează distribuția normală. În caz contrar, valoarea minimă, valoarea primei cvartile, mediana, valoarea celei de a 3-a cvartile şi valoarea maximă se utilizează pentru construirea graficului de tip *boxplot*. Orientarea graficului poate să fie orizontală (figura 2.15) sau verticală (figura 2.16).

2.8.6 Graficul de tip nor de puncte

Pentru prezentarea grafică a relaţiilor sau asocierilor dintre două variabile cantitative perechi (de exemplu, lungime la naştere – perimetru cranian la naştere) se utilizează diagrama de tip nor de puncte (*scatter*).

Fiecare punct din reprezentarea grafică este caracterizat prin două valori, o valoare pe axa OX şi cea de-a doua valoare pe axa OY. Dacă există o relaţie de dependenţă cunoscută între cele două variabile atunci pe axa OX (lungimea în cazul datelor din figura 2.17) se vor pune valorile variabilei independente iar pe axa OY (perimetrul cranian în cazul datelor din figura 2.17) se vor pune valorile variabilei dependente.

Pe graficul de tip nor de puncte se poate prezenta unul (figura 2.17a) sau mai multe grupuri (figura 2.17b) pentru care au fost studiate aceleaşi variabile perechi.

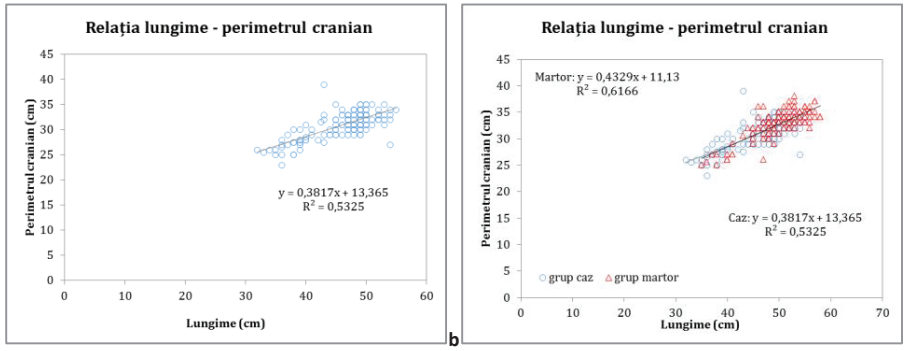


Fig. 2.17. Relaţia dintre lungimea şi perimetrul cranian la naştere pentru datele studiului prezenta în exemplul 2.20: a) grupul cu restricţie de creştere intrauterină (caz); b) grupul cu (caz) şi fără (martor) restricţie de creştere intrauterină.

Pe graficul de tip nor de puncte pot exista informaţii cu privire la dreapta de regresie şi a coeficientului de determinare (r^2) noţiuni care vor fi introduse în capitolul *Corelaţii şi Regresii.*

2.8.7 Graficul de tip linie

Graficul de tip linie permite vizualizarea a două variabile (una de tip timp) şi este utilizat pentru ilustrarea modificărilor în timp care apar la unul sau mai multe grupuri.

Exemplul 2.23.

S-a investigat numărul de cazuri de tuberculoză (TBC) din Republica Moldova şi România în perioada 2004-2014. Sumarizarea datelor colectate este prezentată în Tabelul 2.7.

Tabelul 2.7. Numărul de cazuri de tuberculoză raportate la 100000 locuitori

An / Ţara	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Republica Moldova	5141	4990	4857	4464	4471	4135	4233	4409	4485	4058
România	26106	24295	22590	21724	20643	18590	17045	16107	15523	14861

Reprezentarea grafică a evoluţiei cazurilor de TBC în timp pentru cele două ţări este redată în figura 2.18.

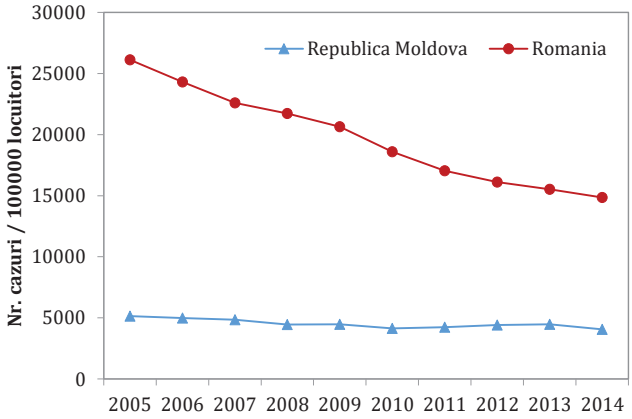


Fig. 2.18. Evoluţia pe parcursul a 10 ani a cazurilor de TBC în România şi Republica Moldova

2.8.8 Graficul de tip arie

Graficul de tip arie permite vizualizarea sumarizărilor cantitative şi arată importanţa relativă a valorilor fiind utilizat pentru a compara doua sau mai multe cantităţi.

Exemplul 2.24.

S-a evaluat scorul de durere al unui eşantion de 7 paciente de gen feminin cu modificări artrozice la includerea în studiu, după prima cură de fizioterapie, la 6 luni de la prima cură de fizioterapie şi respectiv la după cea de-a doua cură de fizioterapie. Scala vizuală a fost utilizată pentru a cuantifica durerea (scor 0 = fără durere, scor 10 = durere extrem de puternică) iar datele brute sunt prezentate în tabelul 2.8.

Tabelul 2.8. Scorul durerii la pacienţi cu artroză

	Scor durere			
Vârsta (ani)	**Iniţial**	**Post fizioterapie 1**	**6 luni**	**Post fizioterapie 2**
65	9	8	2	2
73	7	6	1	2
72	8	7	2	2
50	7	6	5	4
52	10	9	9	8
41	8	8	2	2
76	9	8	3	2

Reprezentarea grafică a evoluţiei scorurilor de durere este redată în figura 2.19.

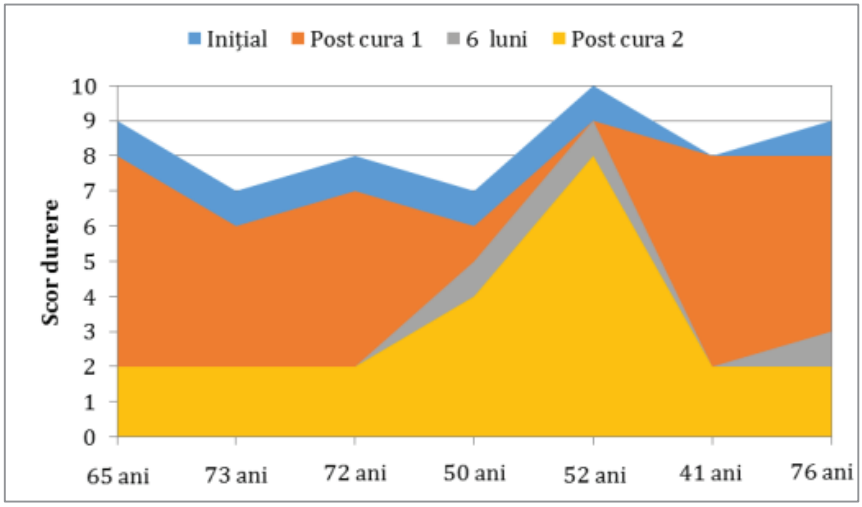


Fig. 2.19. Evoluţia durerii în contextul curei de fizioterapie

Reprezentarea grafică arată fără nici o excepţie diminuarea durerii pacientelor cu artroză atât la finalizarea primei cure de fizioterapie cât şi la 6 luni după prima cură de fizioterapie.

2.8.9 Graficul de tip baloane

Graficul de tip baloane permite reprezentarea grafică simultană a 3 dimensiuni. Diametrul balonului indică valoarea celei de a treia dimensiuni în timp ce prima şi cea de-a doua dimensiune sunt reprezentate pe axa OX şi respectiv OY de poziția centrului acestuia.

Exemplul 2.25.

S-a realizat un studiu pentru a evalua costul tratamentului cu antibiotice pentru tusea convulsivă pe o durată de 5 ani (2008-2012). Datele sunt prezentate în Tabelul 2.9.

Tabelul 2.9. Costuri ale antibioterapiei în tusea convulsivă: 2008-2012

Anul	Nr. cazuri tuse convulsivă	Costul mediu al antibioterapiei
2012	82	2952
2011	86	3096
2010	29	1044
2009	10	360
2008	51	1836

Reprezentarea grafică asociată datelor din Tabelul 2.9 este redată în figura 2.20. Graficul nu este foarte intuitiv deoarece aria fiecărei sfere este proporțională cu costul mediu, deci variază cu radicalul valorii.

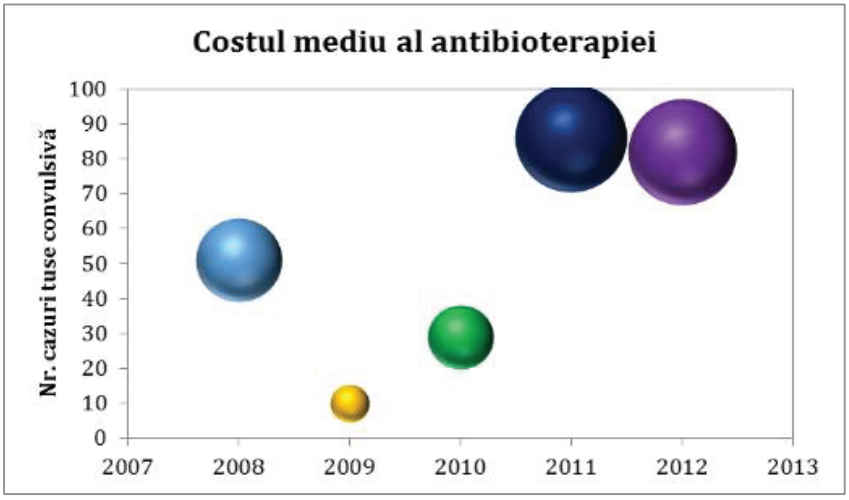


Fig. 2.19. Costul mediu al antibioterapiei per ani pentru tusea convulsivă

2.9 Noțiuni de reţinut

- Proporţia, rata, procentul sunt utilizate pentru sumarizarea (rezumarea, sintetizarea) datelor cantitative (nominale sau ordinale).
- Modalitatea de sumarizare a datelor este dictată de tipul variabilei şi scala de măsură. Acelaşi principiu se aplică şi în cazul calculării parametrilor statistici descriptivi.
- Media aritmetică este un parametru de centralitate caracteristic datelor cantitative.
- Mediana este parametrul de centralitate care caracterizează datele ordinale sau cele cantitative preferată pentru acele date care nu urmează distribuţia normală.

- Deviaţia standard este un parametru de dispersie care caracterizează modalitatea în care datele seriei statistice se distribuie în jurul mediei.
- Coeficientul de variaţie este o mărime de dispersie relativă care permite compararea a două sau mai multe variabile cu sau fără diferite unităţi de măsură.
- Tabelele de frecvenţă arată numărul de observaţii sau % observaţiilor care au o anumită caracteristică.
- Histograma, poligonul de frecvenţă şi *boxplot-ul* permit reprezentarea şi vizualizarea distribuţiei datelor cantitative.
- Graficul de tip nor de puncte prezintă relaţia dintre două caracteristici cantitative.
- Rata descrie numărul de evenimente care apar într-o perioadă definită de timp, de spațiu și dintr-o pupulație definită.
- Relaţia dintre două caracteristici calitative se descrie cu ajutorul indicatorilor de tip raport sau rație (ex.riscul relativ, rația şansei, etc.).

2.10 Exerciţii propuse

1. A fost evaluat un eşantion de 44 subiecţi pentru a identifica dacă fumatul este factor de risc pentru cancerul de plămâni. Eşantionul a cuprins 10 subiecți cu cancer pulmonar, 9 dintre aceștia fiind fumători. 20 din subiecții fără cancer pulmonar erau, de asemenea, fumători. Calculați riscul relativ.

2. Un medic de familie are pe liste 1800 persoane (918 de gen feminin). 506 dintre persoanele de gen feminin au avut vârste cuprinse între 15 și 45 ani în anul 2015 și 115 dintre acestea au născut. Au fost înregistrate 52 infarcte miocardice și un număr de 89 decese în cabinetul acestui medical de familie pe parcursul anul 2015.

1. Care este raportul femei/bărbați?
2. Care a fost rata de natalitate?
3. Calculați rata de fertilitate.
4. Calculați data de morbiditate pentru infarctul miocardic.
5. Calculați rata de mortalitate înregistrată în cabinetul acestui medic de familie.

3. S-a investigat efectul terapiei asistate de animale asupra calității vieții pacienților cu Altzheimer. Douăzeci de participanți cu vârste cuprinse între 58 și 88 ani au fost incluși în studiu (12 femei și 8 bărbați).

1. Care este raportul femei/bărbați?
2. Care este raportul bărbați/femei?

4. Valorile tensiunii arteriale sistolice exprimate în mmHg la un eșantion de 10 pacienţi sunt: 120, 100, 110, 120, 130, 160, 130, 120, 140, 160. Media aritmetică a tensiunii arteriale este:
A. 110
B. 120
C. 135

D. 130
E. Nici un răspuns nu este corect

5. Valorile presiunii arteriale sistolice exprimate în mmHg la un eşantion de 10 pacienţi sunt: 120, 100, 110, 120, 130, 160, 130, 120, 140, 160. Mediana presiunii arteriale sistolice este:
A. 130
B. 125
C. 120
D. 140
E. 135

6. Valorile presiunii arteriale sistolice exprimate în mmHg la un eşantion de 10 pacienţi sunt: 120, 100, 110, 120, 130, 160, 130, 120, 140, 160. Valoarea centrală a serie este:
A. 60
B. 30
C. 260
D. 130
E. 120

7. Valorile presiunii arteriale sistolice exprimate în mmHg la un eşantion de 10 pacienţi sunt: 120, 100, 110, 120, 130, 160, 130, 120, 140, 160. Media aritmetică, mediana, modulului şi valoarea centrală sunt:
A. 129 – 125 – 120 – 130
B. 130 – 125 – 130 – 125
C. 120 – 130 – 120 – 125
D. 129 – 130 – 120 – 130
E. 125 – 125 – 120 – 130

8. Valorile presiunii arteriale sistolice exprimate în mmHg la un eşantion de 10 pacienţi sunt: 120, 100, 110, 120, 130, 160, 130, 120, 140, 130. Seria statistică este:
A. unimodală
B. univariată
C. bimodală
D. trimodală
E. nici un răspuns nu este corect

9. Valorile presiunii arteriale sistolice exprimate în mmHg la un eșantion de 10 pacienți sunt: 120, 100, 110, 120, 130, 160, 130, 120, 140, 160. Varianța eșantionului este egală cu:
A. 349
B. 387,78
C. 3490
D. 3787,8
E. 350

10. Valorile presiunii arteriale sistolice exprimate în mmHg la un eșantion de 10 pacienți sunt: 120, 100, 110, 120, 130, 160, 130, 120, 140, 160. Deviația standard este egală cu:
A. 20

B. 349
C. 378,78
D. 19,69
E. 350

11. Valorile presiunii arteriale sistolice exprimate în mmHg la un eșantion de 10 pacienți sunt: 120, 100, 110, 120, 130, 160, 130, 120, 140, 160. Coeficientul de variație este egal cu:
A. 0,15
B. 15%
C. 0,16
D. 16%
E. 0,13

12. Valorile presiunii arteriale sistolice exprimate în mmHg la un eșantion de 10 pacienți sunt: 120, 100, 110, 120, 130, 160, 130, 120, 140, 160. Din analiza coeficientului de variație (CV=15%) rezultă că seria este:
A. relativ omogenă
B. omogenă
C. eterogenă
D. relativ eterogenă
E. parțial omogenă

13. Valorile presiunii arteriale sistolice exprimate în mmHg la un eșantion de 10 pacienți sunt: 120, 100, 110, 120, 130, 160, 130, 120, 140, 160. Amplitudinea serie statistice este egală cu:
A. 60
B. 160
C. 160
D. 130
E. Nici un răspuns nu este corect

14. Valorile presiunii arteriale sistolice exprimate în mmHg la un eşantion de 10 pacienţi sunt: 120, 100, 110, 120, 130, 160, 130, 120, 140, 160. În care din intervalele următoare ne aşteptăm să găsim 95,5% din observaţii:
A. 109,31 – 148,69
B. 89,62 – 168,38
C. 69,92 – 188,08
D. 109 - 148
E. 89 - 168

15. Valorile presiunii arteriale sistolice exprimate în mmHg la un eşantion de 10 pacienţi sunt: 120, 100, 110, 120, 130, 160, 130, 120, 140, 160. Se ştie că valoarea cvartilei 1 (Q1) este egală cu 120, a cvartilei 2 (Q2) este egală cu 125 şi a cvartilei 3 (Q3) este egală cu 137.5. Care din următoarele afirmaţii sunt corecte:
A. Q2 - Q1 = 5
B. Q3 – Q2 = 12
C. Q3 – Q2 = 12,5
D. distribuţia este aproximativ simetrică

E. distribuţia este asimetrică

16. Valorile presiunii arteriale sistolice exprimate în mmHg la un eşantion de 10 pacienţi sunt: 120, 100, 110, 120, 130, 160, 130, 120, 140, 160. Care din următoarele afirmaţii sunt corecte:
A. media aritmetică = mediana = valoarea modală
B. valoarea modală < mediana < media aritmetică
C. valoarea modală > mediana > media aritmetică
D. seria este asimetrică la dreapta/negativă
E. seria este asimetrică la stânga/pozitivă

17. Următoarele valori reprezintă zilele de incubaţie de la posibilul contact cu agentul etiologic până la manifestarea unei boli infecţioase: 7, 3, 5, 9, 10, 6, 8, 4, 5, 3, 7, 6, 5, 4, 8, 8, 7, 10, 10, 3, 3, 5, 6, 7, 8. Cărei din valorile de mai jos îi corespunde frecvenţa relativă de 0,16:
A. 3
B. 5
C. 10
D. 8
E. 7

18. Următoarele valori reprezintă zilele de incubaţie de la posibilul contact cu agentul etiologic până la manifestarea unei boli infecţioase: 7, 3, 5, 9, 10, 6, 8, 4, 5, 3, 7, 6, 5, 4, 8, 8, 7, 10, 10, 3, 3, 5, 6, 7, 8. Cărei din valorile de mai jos îi corespunde frecvenţa relativă cumulată crescător de 68%:
A. 7
B. 5
C. 10
D. 8
E. Nu se poate determina pe baza informaţiilor disponibile

19. S-a studiat asocierea dintre stres şi hipertensiunea arterială. Au fost incluse în studiu 500 persoane, 220 prezentau hipertensiune arterială, dintre aceştia 100 raportând diferite nivele de stres. Au fost identificaţi 210 pacienţi fără tensiunea arterială şi fără stres. Frecvenţele absolute observate pentru adevărat pozitivi – falşi pozitivi – falşi negativi – adevărat negativi în tabelul de contingenţă asociat problemei sunt:
A. 100 – 70 – 120 – 210
B. 100 – 120 – 70 – 210
C. 100 – 210 – 120 – 70
D. 100 – 120 – 210 – 70
E. Nici un răspuns nu este corect

20. Reprezentarea grafică a datelor cantitative perechi se realizează prin:
A. Diagramă de tip „nor de puncte” (Scatter)
B. Diagramă de tip bare
C. Diagramă de tip coloane
D. Diagramă sectorială (plăcintă sau pie)
E. Diagramă de tip linii

21. Care este tipul de reprezentare grafică pentru vizualizarea relației dintre greutate (kg) și colesterol (mg/dL) la un eșantion de 380 pacienți?
A. Diagrama de tip "nor de puncte"
B. Grafic de tip bare
C. Grafic de tip linii
D. Histograma
E. Gafic sectorial

22. Care este tipul de reprezentare grafică pentru vizualizarea distribuţiei pe clase de frecvenţă a variabilei tensiune arterială sistolică pentru un eşantion de 1000 pacienţi?
A. Diagrama de tip "nor de puncte"
B. Grafic de tip bare
C. Grafic de tip linii
D. Histograma
E. Grafic sectorial

23. Calculaţi pentru datele din Exemplul 2.18. raportul F/M.

24. Care este rata scorului de durere egal cu 10 la femei faţă de bărbaţi pentru datele din Exemplul 2.18.

25. Calculaţi următorii parametri pentru datele din Tabelul 2.4:
1. Prevalenţa diabetului.
2. Proporţia celor cu obezitate în grupul subiecţilor fără diabet.
3. Proporţia celor cu diabet în grupul subiecţilor obezitate.
4. Proporţia celor fără obezitate în grupul subiecţilor fără diabet.
5. Proporţia celor fără diabet în grupul subiecţilor cu obezitate.

26. Reprezentaţi grafic datele prezentate în Tabelul 2.4.

27. Reprezentaţi grafic datele din Exemplul 2.18.

28. Calculaţi pentru datele din Exemplul 2.16:
1. Media aritmetică
2. Media geometrică
3. Mediana
4. Modulul
5. Valoarea primei cvartile
6. Valoarea celei de a treia cvartile
7. Deviaţia standard
8. Coeficientul de variaţie
9. Amplitudinea
10. Eroarea standard

29. Care este parametrul de centralitate care caracterizează cel mai bine seria statistică din Exemplul 2.16?
Argumentaţi răspunsul.

30. Care este parametrul de dispersie care caracterizează cel mai bine seria statistică din Exemplul 2.16?
Argumentaţi răspunsul.

Exerciții bazate pe interpretarea articolelor medicale

31. Scenariul pentru acest exerciţiu este construite pe baza studiului publicat de Lee și co-autorii (Lee JS, Park SY, Kim JS, You JY, Ju YS, Eom JS. The clinical effectiveness of oseltamivir in mild cases of pandemic influenza A H1N1 2009 infection. Scand J Infect Dis. 2012;44(8):595-9.).

S-a investigat eficiența medicamentului Oseltamivir în tratamentul infecțiilor cu virusul gripal AH1N1. Au fost incluși în studiu 90 subiecți cu tratamentul de interes și 72 subiecți cu tratament simptomatic. În ambele grupuri tratamentul a fost urmat la domiciliu. Principala variabilă urmărită a fost durata infecției cu virusul gripal. S-au raportat următoarele rezultate:

Tratament	**Durata gripei (zile)**	**Timpul până la *mă simt bine* (zile)**	**Timpul până la reluarea activităților curente (zile)**
Oseltamivir	6,50 ± 3,75	1,70 ± 1,57	7,13 ± 2,61
Simptomatic	7,04 ± 3,75	2,00 ± 2,12	7,58 ± 2,71

1. Identificați semnificația parametrilor numerici din tabelul prezentat anterior.
2. Care este modalitatea cea mai ilustrativă de reprezentare grafică a diferențelor duratei gripei la cele două grupuri?
3. Care este modalitatea cea mai ilustrativă de reprezentare grafică a diferențelor timpului până la *mă simt bine* la cele două grupuri?
4. Dar pentru timpul până la reluarea activității curente?
   Argumentaţi răspunsurile dvs.

32. Scenariul pentru acest exercițiu este construite pe baza studiului publicat de Joshi și co-autorii (Joshi M, Chandra D, Mittadodla P, Bartter T. The impact of vaccination on influenza-related respiratory failure and mortality in hospitalized elderly patients over the 2013-2014 season. Open Respir Med J. 2015;9:9-14). Scopul studiului a fost de a descrie caracteristicile clinice și post tratament ale pacienților cu infecție virală la un anumit spital la subiecţii vaccinați sau nu antigripal.

1. Se dorește reprezentarea grafică a statusului de vaccinat/nevaccinat pe de o parte și internat/tratat la domiciliu pe de altă parte. Care este modalitatea cea mai ilustrativă de reprezentare grafică?

2. Care este modalitatea cea mai ilustrativă de reprezentare grafică a diferențelor în ceea ce privește co-morbiditățile dintre grupul celor vaccinați și respectiv grupul celor nevaccinați?
3. Cum reprezentăm grafic diferențele de vârstă între cei internați și cei tratați la domiciliu?
4. Care sunt parametrii descriptivi pe care i-am putea raporta pentru variabila vârsta la internare pentru cei vaccinați și respectiv pentru cei nevaccinați?

Argumentaţi răspunsurile dvs.

33. Scenariul pentru acest exercițiu este construit pe baza studiului publicat de Hutchinson și co-autorii (Hutchinson PJ, Kolias AG, Timofeev IS, Corteen EA, Czosnyka M, Timothy J, et al. Trial of Decompressive Craniectomy for Traumatic Intracranial Hypertension. N Engl J Med. 2016). Au fost incluși în studiu pacienți cu traumatism cranian și hipertensiune intracraniană refractară la tratament. Rezultatul primar urmărit a fost scorul Glasgow extins evaluat la 6 și 12 luni (scală de 8 puncte, deces – stare vegetativă – dizabilitate severă membre inferioare – dizabilitate severă membre superioare – dizabilitate moderată membre inferioare – dizabilitate moderată membre superioare – recuperare membre inferioare bună – recuperare membre superioare bună). Eșantionul studiat a fost împărțit în două grupuri, un grup a beneficiat doar de tratament medical iar la cel de-al doilea grup s-a practicat craniectomie decompresivă.

1. Care este parametrul/parametrii descriptivi pe care i-aţi calcula pentru a identifica diferenţele între succesul tratamentului definit ca „recuperare bună”?
2. Care sunt parametrii statistici pe care i-ați calcula pentru variabila vârstă?
3. Care sunt parametrii statistici pe care i-ați calcula pentru variabila gen?
4. Cum ați reprezenta grafic scorul Glasgow la 6 luni pentru cele două grupe?
5. Cum ați reprezenta grafic scorul Glasgow la 6 și 12 luni pentru cele două grupe?

Argumentați răspunsurile dvs.

34. Scenariul pentru acest exercițiu este construite pe baza studiului publicat de Knight şi co-autorii (Knight BA, McIntyre HD, Hickman IJ, Noud M. Qualitative assessment of user experiences of a novel smart phone application designed to support flexible intensive insulin therapy in type 1 diabetes. BMC Med Inform Decis Mak. 2016;16:119.). Aceştia au evaluat feedback-ul cu privire la utilizarea unei aplicații mobile care permite calcularea dozei de insulină necesară pacienților cu diabet zaharat de tip I în cazul terapiei flexibile injectabile multiple. Au fost incluși în studiu 7 adulți care au participat la sesiunile de instruire în utilizarea aplicației și interviul de evaluare al acesteia.

1. Care este parametrul/parametrii descriptivi pe care i-aţi calcula pentru a sumariza vârsta subiecților incluși în studiu?
2. Care este parametrul/parametrii descriptivi pe care i-aţi calcula pentru a sumariza durata diabetului?
3. Cum ați reprezenta grafic gradul de satisfacție al subiecților incluși în studiu cu primire la utilitatea aplicației testate? Argumentaţi răspunsurile dvs.

35. Scenariul pentru acest exercițiu este construite pe baza studiului publicat de Pils și co-autorii (Pils S, Promberger R, Springer S, Joura E, Ott J. Decreased Ovarian Reserve Predicts Inexplicability of Recurrent Miscarriage? A Retrospective Analysis. PLoS One. 2016;11(9):e0161606). Aceștia au evaluat hormonul anti-Mullerian, hormonul de stimulare foliculară, hormonul luteinizant și estradiolul la femei cu avorturi spontane recurente (cauze necunoscute vs. idiopatic).

1. Care este modalitatea în care ați reprezenta grafic motivele pierderii sarcinii pe grupul în care acestea se cunosc?
2. Care este parametrul/parametrii descriptivi pe care i-ați calcula pentru a sumariza variabilele cantitative studiate dacă acestea nu urmează distribuţia normală?
3. Care este parametrul/parametrii descriptivi pe care i-ați calcula pentru a sumariza variabilele cantitative studiate dacă acestea urmează distribuţia normală?

Argumentați răspunsurile dvs.

3. Probabilitea și aplicațiile ei medicale

„Medicine is a science of uncertainty and an art of probability"
(William Osler)

Obiective educaționale:

- însuşirea noțiunilor de experiment aleator, eveniment şi spațiu fundamental
- asimilarea noțiunii de probabilitate
- calculul probabilității prin formulele de bază şi interpretarea rezultatelor
- identificarea evenimentelor independente prin prisma noțiunii de probabilitate
- aplicabilitatea legii lui Bayes
- identificarea aplicațiilor medicale ale teoremei lui Bayes

3.1 Conceptul de probabilitate

Expresia „probabil că...." o cunoaştem cu toții şi o folosim frecvent în limbajul curent atunci când evaluăm posibilitatea ca anumite situații (evenimente) să se realizeze sau nu. Indiferent de nivelul cunoştiințelor pe care le avem despre teoria probabilităților, estimăm adesea probabilități.

În analiza statistică a datelor medicale, conceptul de probabilitate este folosit atât pentru a reda aspectul aleator al unei caracteristici (variabile) pe o populație cu ajutorul distribuției de probabilitate cât şi pentru a formula concluziile inferenței statistice.

Să presupunem că un obiectiv clinic de interes ar fi evidențierea dacă subiecții de o anumită vârstă expuşi la un factor de risc (sedentarismul) au un risc mai mare de a face afecțiuni cardiovasculare decât cei care desfaşoară o activitate fizică regulată. Printr-un studiu observațional prospectiv (subiecții au fost urmăriți o perioadă de timp) dintr-un eşantion de 1000 de subiecți, 200 din cei sedentari și respectiv 50 din cei cu activitate fizică regulată au dezvoltat afecțiuni cardiovasculare, iar din cei cu o activitate fizică regulată 20 au dezvoltat afecțiuni cardiovasculare. Ne întrebăm în acest caz dacă avem suficiente dovezi pentru a afirma că există o diferență de risc. Formularea concluziilor în acest caz se va face în termeni de probabilități.

Pentru introducerea conceptului de probabilitate, precum şi a modului în care aceasta se calculează în anumite situații, este nevoie de cunoaşterea în prealabil a noțiunilor de experiment aleator, spațiu fundamental şi eveniment.

3.1.1 Experiment aleator, spațiu fundamental şi eveniment

În medicină sunt rare situațiile în care se întâlnesc experimentele deterministe, adică experimentele al căror rezultat poate fi ştiut a priori, cele mai frecvente fiind experimentele

aleatoare. Aşa cum rezultă din denumire, experimentul aleator presupune un anumit grad de hazard sau incertitudine.

Exemplul 3.1

Să presupunem că suntem interesaţi de următoarele investigaţii (teste) pe o anumită populaţie:

- determinarea colesterolului din sânge (LDL-colesterol sau HDL-colesterol)
- determinarea tensiunii arteriale sistolice (TAS mmHg)
- determinarea statusului (prezenţa/absenţa) mutaţiei V600E în gena BRAF
- determinarea oxidului nitric (FeNO ppb=părţi pe miliard) în aerul expirat în vederea diagnosticului şi monitorizării inflamaţiei pulmonare la pacienţii cu astm
- stabilirea prezenţei sau absenţei sedentarismului ca factor de risc pentru boală cardiovascuară, osteoporoză sau diabet zaharat
- determinarea gradului de fibroză hepatică cu ajutorul Fibrotestului la pacienţii cu hepatită cronică VHC
- determinarea statusului (prezent/absent) pentru anticorpi anti HCV şi testarea pentru ARN-ul virusului hepatic C la pacienţii susceptibili de infecţie cu virus hepatic C

Toate aceste investigaţii sunt exemple de **experimente aleatoare** pentru că presupun realizarea practică a condiţiilor unui criteriu de interes (măsurarea caracteristicii de interes: colesterol, mutaţia genetică, oxid nitric, sedentarism), iar rezultatul măsurătorii la un subiect ales întâmplător din colectivitatea de studiu nu poate fi cunoscut înaintea efectuării acesteia. De asemenea, măsurărarea caracteristicilor de interes la pacienţii colectivităţii conduce la rezultate diferite între pacienţi.

Definiţii:

Conform definiţiei uzuale, un **experiment** reprezintă realizarea practică a unui ansamblu de condiţii (criteriu de cercetare) atunci când aceste sunt aplicate unei colectivităţi (populaţie sau eşantion) care poate fi repetat în condiţii identice.

Experimentul aleator este acel experiment ce satisface următoarele condiţii:

- rezultatul aplicării lui asupra unui element al populaţiei nu poate fi prezis cu certitudine înaintea efectuării experimentului
- repetat în condiţii identice poate conduce la rezultate diferite.

Fiecare repetare a experimentului aleator se numeşte **probă**.

Exemple:

- în determinarea statusului (prezent/absent) mutaţiei V600E în gena BRAF, mulţimea tuturor rezultatelor posibile este: {0,1}, unde 0=absenţa şi 1=prezenţa mutaţiei
- în determinarea grupei sanguine, mulţimea tuturor rezultatelor posibile este: {A, B, AB, 0}
- pentru determinarea tensiunii arteriale sistolice putem considera că valorile acestei variabile sunt incluse în intervalul $[0, \infty)$

- în cazul determinării statusului(prezent/absent) pentru anticorpi anti HCV şi testarea pentru ARN-ul virusului hepatic C la pacienții susceptibili de infecție cu virus hepatic C, fiecare pacient este supus la două teste diagnostice, ambele teste cu rezultat de tip pozitiv=1/negativ=0, deci mulțimea tuturor rezultatelor posibile este: {(1,1), (1,0), (0,1), (0,0)}

Exemplele de mai sus evidențiază faptul că mulțimea tuturor rezultatelor posibile ale unui experiment poate fi o mulțime finită (număr de elemente cunoscut) sau infinită de elemente.

Mulțimea tuturor rezultatelor posibile ale unui experiment aleator se numeşte **spațiul fundamental (mulțimea cazurilor posibile)**, notat cu E în cadrul acestui capitol. Fiecare rezultat al experimentului este un element al lui E.

Spaţiul fundamental poate fi **finit** sau **infinit**.

Exemplul 3.2

Grupa sangvină. În cadrul experimentului de determinare a grupei sangvine, ne întrebăm dacă:

i. un subiect ales întâmplător poate să aibă grupa sangvină AB?
ii. un subiect ales întâmplător poate să aibă grupa sangvină C?

În exemplul dat, proba este reprezentată de determinarea grupei sangvine la un subiect, iar la punctul i) evenimentul care ne interesează este A: subiectul să aibă grupa sangvină AB. Este evident că acestui eveniment îi va corespunde o mulțime de rezultate sau cazuri favorabile (mulțimea de cazuri care realizează evenimentul A). Astfel, evenimentului A îi va corespunde submulțimea: {AB}. Pentru că evenimentul şi submulțimea din E asociată lui se determină reciproc, nu vom face diferență între acestea şi vom scrie A={AB}.

Evenimentul de la punctul ii) B={C} este un exemplu de eveniment care nu se realizează la nici o efectuare a experimentului şi este numit eveniment imposibil.

Un **eveniment** este o submulțime a spațiului fundamental (o submulțime de rezultate posibile).

Fiecare probă implică realizarea sau nerealizarea unui eveniment, astfel încât se poate spune că fiecarui eveniment îi corespunde o mulțime de cazuri (rezultate) favorabile.

Evenimentele se notează cu majuscule (A, B) şi pot fi reprezentate şi sub forma grafică, cu ajutorul diagramei Venn (Figura 3.1) în care dreptunghiul reprezintă spațiul fundamental E, iar cercurile reprezintă evenimente.

Oricărui experiment aleator i se pot asocia **trei tipuri de evenimente: posibil, imposibil şi evenimentul sigur** (cert). **Evenimentul aleator** (întâmplător) este evenimentul care poate să se producă sau nu. **Evenimentul imposibil** (notat cu $\varnothing$) este acel eveniment care nu se realizează niciodată la efectuarea experimentului aleator (nu conține nici un rezultat). De exemplu apariția cancerului de prostată la un subiect de gen feminin este un eveniment imposibil.

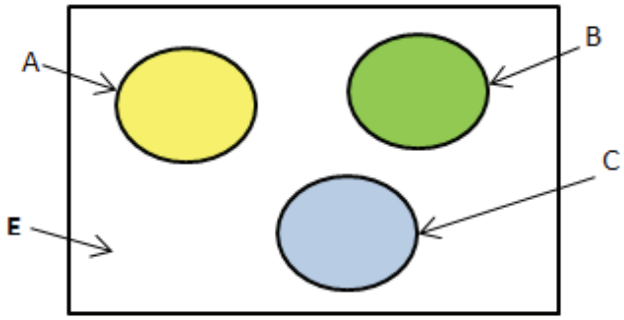


Fig. 3.1. Reprezentarea grafică a evenimentelor unui experiment aleator. A,B, C sunt evenimente, E reprezintă spațiul fundamental

Evenimentul sigur (notat cu E) este evenimentul care se realizează la fiecare efectuare a experimentului aleator, deoarece conține toate rezultatele posibile ale experimentului (este evenimentul compus alcătuit din toate evenimentele posibile).

3.1.2 Operații cu evenimente

În teoria probabilităților putem vorbi despre **eveniment elementar** (evenimentul ce conține un singur rezultat; acel eveniment ce nu se poate descompune în alte evenimente şi care se realizează printr-o singură probă) şi **eveniment compus** a cărui realizare depinde de cel puțin două evenimente elementare.

Există câteva tipuri particulare de evenimente compuse care formează pornind de la operațiile matematice între mulțimi, operații care în limbajul comun utilizează cuvinte cheie de tipul: *SAU, ŞI, NU, DACĂ*.

Eveniment compus prin disjuncție sau reuniune

Exemplul 3.3

Durata de supraviețuire. S-a observat timpul de supraviețuire definit ca timpul scurs de la diagnostic până la deces la pacienții cu melanom cutanat avansat.

Considerăm că:

i) A este evenimentul ca un pacient să aibă timpul de supraviețuire mai mic de 6 luni şi B evenimentul ca un pacient să aibă timpul de supraviețuire cuprins între 6 şi 12 luni.

Atunci evenimentul „A sau B" (notat $A \cup B$) este evenimentul ca un pacient cu melanom cutanat avansat, ales la întâmplare, să aibă timpul de supraviețuire mai mic de 12 luni.

ii) A={pacient ales aleator să aibă timp supraviețuire≥6 luni} şi B={pacient ales aleator să aibă timp supravietuire≤12 luni}

Atunci $A \cup B$={pacient ales aleator să aibă timp supraviețuire≥6 luni sau timp supravietuire≤12 luni sau timp de supraviețuire între 6 şi 12 luni}

Definiție

Evenimentul compus "A sau B" (notație A∪B), unde A şi B sunt două evenimente date, este evenimentul care are loc dacă cel puţin unul dintre evenimentele A sau B are loc.

Eveniment compus prin conjuncție sau intersecție

Exemplul 3.4

Experiment: măsurarea valorilor HDL-colesterol mg/dl Considerăm evenimentele A={pacient ales aleator să aibă HDL≥60 mg/dl} şi B={pacient ales aleator să aibă 40≤HDL≤100 mg/dl}

Atunci putem defini şi evenimentul A∩B={pacient ales aleator să aibă 60≤HDL≤100}={pacientul să aibă HDL-colesterol cuprins între 60 şi 100 mg/dl inclusiv} (Fig.3.2)

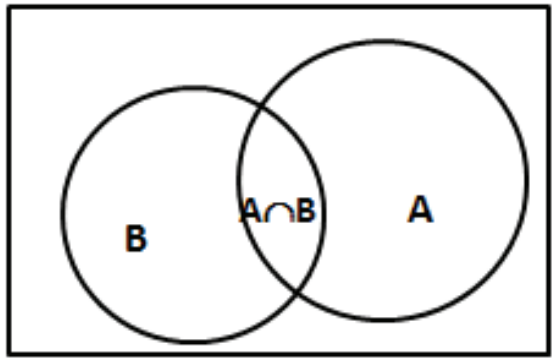


Fig 3.2. Reprezentarea grafică a evenimentului A∩B

Definiție:

Evenimentul compus "A şi B" (notație A∩B), unde A şi B sunt două evenimente date care pot avea loc simultan, este evenimentul care se realizează dacă evenimentele A şi B au loc.

Eveniment compus prin negație

Exemplul 3.5

Experiment: evaluarea durerii acute la copil prin Scala Termometru pentru Durere asociată cu chipuri simbolice, scală ce cuantifică intensitatea durerii cu un scor de la 0 la 10.

Dacă evenimentul A={subiectul să aibă durere moderată}={3,4}, atunci putem vorbi şi despre evenimentul complementar lui A care are loc atunci când A nu se produce: nonA={subiectul poate să nu aibă durere sau subiectul suferă de durere mică sau suferă de durere severă sau suferă de durere foarte severă sau suferă de cea mai mare durere posibilă}={0,1,2,5,6,7,8,9,10}

Exemplul 3.6

Experiment: determinarea nivelului anticorpilor anti-tTGA (anti-transglutaminază) ca metodă de screening pentru boala celiacă (intoleranță la gluten), considerând că o valoare tTGA≥100U/mL semnifică rezultat pozitiv.

Dacă evenimentul A={subiectul să aibă tTGA ≥100U/mL}, atunci putem vorbi şi despre evenimentul complementar lui A care are loc atunci când A nu se produce: nonA={ pacient ales aleator să aibă tTGA <100 U/mL}.

Definiție

Evenimentul compus nonA este evenimentul care se realizează dacă evenimentul A nu are loc.

În tabelul 3.1 sunt descrise simbolurile utilizate în teoria probabilităților şi corespondența acestora cu teoria mulțimilor.

Tabel 3.1. Corespondența dintre limbajul mulțimilor şi cel al evenimentelor

Terminologie mulțimi	Notație	Terminologie evenimente	Notație
Mulțime dată	A={a, b, c}	Eveniment dat	A={a, b, c}
Mulțimea totală	A=E	Evenimentul cert	E
Mulțimea vidă	∅	Evenimentul imposibil	∅
Mulțime cu un singur element	A={a}	Eveniment elementar	A={a}
Reuniunea lui A cu B	A∪B	Evenimentul A sau B	A∪B
Intersecția lui A cu B	A∩B	Evenimentul A şi B	A∩B
Complementara lui A	C_A	Opusul lui A	nonA

3.1.3 Principalele relații între evenimente

Între două evenimente există câteva relații fundamentale care pot determina realizarea evenimentelor pe care le implică: complementaritate, incluziune, incompatibilitate şi dependență.

Relația de complementaritate

Două evenimente sunt **complementare** dacă realizarea unuia dintre ele conduce la nerealizarea celuilalt. În această situație se încadrează evenimentele contrare (evenimentul A şi B=nonA).

Relația de incluziune

Exemplul 3.7

Experiment: măsurarea TAS mmHg; Considerăm evenimetele A={TAS≥120mmHg} şi B={TAS≥100 mmHg}.

Se observă că mulțimea rezultatelor favorabile lui B conține mulțimea rezultatelor favorabile lui A, aceasta însemnând că ori de câte ori evenimentul A are loc se realizează şi evenimentul B.

Definiție

Evenimentul A implică evenimentul B ($A \subset B$) dacă odată cu realizarea evenimentului A se realizează şi evenimentul B.

Relația de compatibilitate (sau incompatibilitate)

Exemplul 3.8

Experiment: măsurarea TAD mmHg

Considerăm evenimentele A={ pacient ales aleator să aibă 60≤TAD≤79 mmHg} şi B={ pacient ales aleator să aibă 100≤TAD≤110 mmHg}. Cele două evenimente nu pot avea loc simultan (un subiect ales întâmplător nu poate avea în acelaşi timp valori normale în privința TAD şi valori crescute ale TAD).

Experiment: determinarea numărului de noduli limfatici cu celule canceroase la pacienții cu neoplasm pulmonar avansat

Evenimentele A={pacientul să aibă cel mult 2 noduli limfatici pozitivi} şi B ={pacientul să aibă cel puțin 4 noduli limfatici pozitivi} sunt realizabile simultan? NU

Definiții

Două evenimente se numesc **compatibile** dacă pot să aibă loc (să se realizeze) în acelaşi timp.

Două evenimente sunt **incompatibile** sau **mutual exclusive** sau **disjuncte** dacă nu se pot realiza în acelaşi timp, cu alte cuvinte, dacă unul dintre ele se realizează este cert că celalalt nu poate avea loc.

De remarcat: Două evenimente complementare sunt întotdeauna incompatibile.

Relația de dependență (sau independență)

Exemplul 3.9

Indice de masă corporală. S-a determinat indicele de masă corporală (IMC=greutate/înălțime^2(kg/m^2)) la studenții unei facultăți. Evenimentul ca un student ales aleator să aibă IMC≥30 este independent de evenimentul ca un student aleator să aibă diabet?!

Realizarea evenimentului A=(subiectul să aibă IMC≥30} înseamnă de fapt că studentul este obez, deci realizarea evenimentului A aduce o informație ce ar putea influența realizarea evenimentului B (subiectul să aibă diabet).

Definiție

Două evenimente se numesc **evenimente dependente,** dacă realizarea unuia influențează (depinde de) realizarea celuilalt. Două evenimente sunt **independente** dacă realizarea unuia dintre ele nu depinde de faptul că evenimentul celălalt s-a produs sau nu.

De remarcat: Două evenimente incompatibile sunt dependente.

3.1.4 Definiția teoretică a probabilității

Conceptul de probabilitate este singurul în măsură să ofere o descriere adecvată a incertitudinii. Există două abordări în ceea ce priveşte definirea probabilității: obiectivă (cu definiția clasică, definiția frecvențială) şi subiectivă, ultima dintre acestea nefăcând obiectul acestui curs.

Exemplul 3.10

Aruncarea zarului. La experimentul de aruncare a unui zar ne întrebăm: care este şansa să apară față cu număr par? (probabilitatea de apariție a unei fețe cu număr par)?
Definim evenimentul A={număr par}={2,4,6}. Pentru că evenimentele elementare {2};{4};{6} sunt egal posibile, putem cosidera că şansa realizării fiecăruia este egală cu frecvența relativă:Pr({2})=Pr({4})=Pr({6})=1/6.

Pentru că evenimentul A este compus din 3 evenimente elementare, fiecare cu aceeaşi şansă de realizare, şansa producerii evenimentului A va fi deci de 3 ori mai mare: Pr(A)=3/6=1/2=0,5 (interpretare: un caz din 2 conduce la realizarea lui A sau altfel spus: evenimentul A se realizează în 50% din cazuri).

Definiție: probabilitatea unui eveniment A este egală cu raportul dintre numărul de rezultate favorabile evenimentului şi numărul total de rezultate posibile ale experimentului:

$$Pr(A) = \frac{Numărul\ de\ cazuri\ favorabile\ \ evenimentului\ A}{Numărul\ de\ cazuri\ posibile}$$

Probabilitatea calculată în acest mod este o probabilitate teoretică, *a priori* experimentului pentru că se poate calcula înaintea efectuării acestuia cunoscându-se numărul posibil de evenimente şi implică faptul că toate evenimentele elementare din spațiul fundamental au aceeaşi şansă de realizare (sunt echiprobabile).

3.1.5 Definiția frecvențială a probabilității

Cele mai multe situații implică evenimente a căror probabilitate nu se poate determina decât repetând experimentul (de exemplu probabilitatea ca un nou-născut să fie de gen feminin nu se poate determina observând repartiția în funcție de gen a copiilor unei singure familii, ci mai degrabă observând numărul total de naşteri vii pe o populație într-o anumită perioadă). În acest caz se apelează la definiția empirică conform căreia probabilitatea este definită ca o generalizare a noțiunii de frecvență relativă a unui eveniment.

Exemplul 3.11

Testarea medicală. Imunofluorescența este un test indirect de depistare a boreliozei (boala Lyme), rezultatul testului fiind de tipul pozitiv, negativ sau incert. S-a realizat testarea prin imunofluorescență la un grup de 2000 de subiecți dintre care 100 au avut un rezultat pozitiv la test. Care este şansa ca un subiect extras aleator din acest grup să aibă rezultat pozitiv la test: Pr(A)=?

Probabillitatea cerută va fi o aproximare a frecvenței relative asociată evenimentului A: Pr(A)≈100/2000=0,05 care se interpretează astfel: dacă un număr suficient de mare de subiecți au fost testați, aproximativ 5% dintre aceştia au obținut un rezultat pozitiv la test.

Definiții:

Dacă într-o serie de probe, evenimentul A s-a realizat de n_A ori atunci numărul n_A se numeşte **frecvența absolută a evenimentului A**, iar raportul n_A/n se numeşte **frecvența relativă a evenimentului A**.

Dacă un experiment se repetă de un număr suficient de mare de ori atunci **probabilitatea evenimentului A** se poate aproxima prin frecvența relativă a evenimentului (sau putem spune că probabilitatea este numărul către care tind frecvențele relative ale evenimentului dacă se repetă experimentul):

$$Pr(A) \approx \frac{n_A}{n} = \frac{Numărul\ de\ cazuri\ favorabile\ \ evenimentului\ A}{Numărul\ de\ probe\ (repetiții)\ experiment}$$

definită în acest mod, vorbim de o probabilitate **a posteriori** a experimentului, calculabilă doar după efectuarea lui (prin repetarea experimentului în condiții identice).

Aplicație: în medicina clinică, termenul de **prevalență a unei boli** reflectă noțiunea de probabilitate într-un context particular, prevalența fiind probabilitatea ca un subiect din populație să aibă boala şi se calculează ca raportul dintre numărul de subiecți care au efectiv boala şi volumul populației într-un anumit interval de timp și într-o regiune geografică.

3.1.6 Reguli de calcul a probabilităților

Exemplul 3.12

Tensiune crescută. Într-o populație de interes se cunoaşte că 60% din subiecți au avut valori TAS≤119 mmHg şi 10% dintre subiecți au avut valori ale TAS cuprinse între 120 şi 139 inclusiv. Care este probabilitatea ca un subiect ales la întâmplare să aibă valoarea TAS ≤139 mmHg?

Definim evenimentele A={subiectul să aibă TAS≤119 mmHg}, B={subiectul să aibă 120≤TAS≤139 mmHg} şi C={subiectul să aibă TAS≤139 mmHg}

Evenimentele A şi B nu pot avea loc simultan deci Pr(C=A∪B)=Pr(A) + Pr(B)=0,6+0,1=0,7.

Reguli de calcul cu probabilități:

- $0 \leq Pr(A) \leq 1$
- $Pr(E) = 1$
- Dacă A şi B sunt incompatibile atunci $Pr(A \cup B) = Pr(A) + Pr(B)$
- Dacă A_1, A_2, ..., A_n sunt evenimente incompatibile două câte două atunci:

$$Pr(A_1 \cup A_2 \cup \ldots \cup A_n) = Pr(A_1) + Pr(A_2) + \ldots + Pr(A_n)$$

- $Pr(\text{non } A) = 1 - Pr(A)$.
- Dacă $A \subseteq B$ atunci $Pr(A) \leq Pr(B)$
- Pentru orice evenimente A şi B are loc egalitatea:

$$Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$$

3.2 Evenimente independente

Exemplul 3.13

Boala celiacă. Să considerăm că toți pacienții care s-au prezentat la o clinică pentru susceptibilitate de boală celiacă au fost evaluați de medicii X şi Y. Medicul X a diagnosticat 20% din cazuri ca fiind pozitive, medicul Y determină 30% din cazuri ca fiind pozitive în timp ce 10% din cazuri sunt confirmate cu diagnostic pozitiv de ambii medici.

Considerăm evenimentele: A={pacient ales aleator să aibă diagnostic pozitiv dat de medicul X} şi B={pacient ales aleator să aibă diagnostic pozitiv dat de medicul Y}.

Evenimentele A şi B pot fi considerate independente sau dependente?

Pe baza definiției date în capitolul anterior, referitoare la independența evenimentelor, dar şi cunoscând că metodele folosite de cei doi medici pentru diagnosticarea bolii ar trebui să aibă un anumit grad de similaritate, evenimentele A şi B sunt dependente.

Dacă evenimentele A şi B ar fi independente în probabilitate, atunci se poate afirma că probabilitatea de producere simultană a acestora este egală cu produsul dintre probabilităţile lor. În cazul nostru putem observa că: $Pr(A)=0,20$, $Pr(B)=0,30$, $Pr(A \cap B)=0,10$, iar $Pr(A) \cdot Pr(B)=0,06$, deci $Pr(A \cap B) \neq Pr(A) \cdot Pr(B)$.

Definiții:

- Două evenimente A şi B sunt **independente** dacă şi numai dacă este satisfăcută condiția următoare:

$$Pr(A \cap B) = Pr(A) \cdot Pr(B)$$

- Două evenimente A şi B sunt **dependente** dacă dacă şi numai dacă este satisfăcută condiția următoare:

$$Pr(A \cap B) \neq Pr(A) \cdot Pr(B)$$

- Dacă A şi B sunt două **evenimente independente** atunci este satisfăcută relația:

$$Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A) \cdot Pr(B)$$

3.3 Probabilități condiționate

Exemplul 3.14

Helicobacter pylori. O procedură paraclinică (test) de detectare a infecției cu Helicobacter pylori este determinarea Antigenului Hpylori din materii fecale, rezultatul testului fiind de tipul pozitiv sau negativ. Considerăm evenimentele HPT+={pacient ales aleator să aibă rezultat pozitiv la testare} şi HP+={pacient ales aleator să aibă infecție cu Helicobacter pylori}. Să presupunem că a fost testat un eşantion reprezentativ de 1500 pacienţi, dintre care 1050 au obținut un rezultat pozitiv la test iar 450 au avut un rezultat negativ. Numărul subiecțiilor cu infecție Hpylori, respectiv subiecții fără infecție este descris în tabelul 3.2.

Tabelul 3.2. Repartiția pacienților în funcție de rezultatul testului şi prezența infecției

Infecția / Test	**HP+**	**HP-**	**Total**
HPT+	1000	60	1060
HPT-	200	240	440
Total	1200	300	**1500**

i) Care este probabilitatea ca un bolnav extras aleator să aibă rezultat pozitiv la test?

ii) Care este probabilitatea ca un subiect fără infecție, extras aleator să aibă rezultat negativ la test?

iii) Care este probabilitatea ca un subiect cu rezultat pozitiv la test să aibă într-adevăr infecție cu Hpylori?

iv) Care este probabilitatea ca un subiect cu rezultat negativ la test să nu aibă infecție cu Hpylori?

La punctul:

i) Pornind de la evenimentele date în exemplu HPT+={pacient ales aleator să aibă rezultat pozitiv a test} şi HP+={pacient ales aleator să aibă infecție cu Helicobacter pylori} putem vorbi şi despre evenimentele contrare: HPT-={pacient ales aleator să aibă rezultat negative la test} şi HP- ={pacient ales aleator să nu aibă infecție cu Helicobacter pylori}. Probabilitatea ca un bolnav extras aleator să aibă rezultat pozitiv la test poate fi aproximată prin proporţia bolnavilor pe care testul este capabil să-i identifice şi este raportul 800/1200=0.83. Conform definiției avem:

$$\Pr(HPT+ \mid HP+) = \frac{\Pr(HPT+ \cap HP+)}{\Pr(HP+)} \approx \frac{\frac{1000}{1500}}{\frac{1200}{1500}} = \frac{1000}{1200} = 0{,}83$$

ii) Probabilitatea ca un subiect fără infecție, extras aleator să aibă rezultat negativ la test poate fi aproximat de proporţia subiecțiilor fără infecție Hpylori confirmați ca fiind în afara stării de boală în urma înregistrării rezultatelor negative ale testului şi este raportul 240/300=0,80. Pe de altă parte, utilizând definiția probabilității condiționate avem:

$$\Pr(HPT\text{-}\,|\,HP\text{-}) = \frac{\Pr(HPT\text{-} \cap HP-)}{\Pr(HP\text{-})} \approx \frac{\frac{240}{1500}}{\frac{300}{1500}} = \frac{240}{300} = 0{,}80$$

iii) Probabilitatea ca un subiect cu rezultat pozitiv la test să aibă intr-adevăr infecție cu Hpylori este data de proporţia cazurilor cu infecție Hpylori din totalul rezultatelor pozitive ale testului şi este raportul 1000/1060=0,94:

$$\Pr(HP+\,|\,HPT+) = \frac{\Pr(HPT+ \cap HP+)}{\Pr(HPT+)} \approx \frac{\frac{1000}{1500}}{\frac{1060}{1500}} = \frac{1000}{1060} = 0{,}94$$

iv) Probabilitatea ca un subiect cu rezultat negativ la test să nu aibă infecție cu Hpylori este proporția cazurilor fără infecție Hpylori din totalul rezulatelor negative ale testului: $\Pr(HP\text{-}\,|\,HPT\text{-}) = \frac{\Pr(HPT\text{-} \cap HP\text{-})}{\Pr(HPT\text{-})} \approx \frac{\frac{240}{1500}}{\frac{440}{1500}} = \frac{240}{440} = 0{,}55$

Exemplul 3.15

Boala celiacă. Să considerăm că toți pacienții care s-au prezentat la o clinică pentru susceptibilitate de boală celiacă au fost evaluați de medicii X şi Y. Medicul X a diagnosticat 20% din cazuri ca fiind pozitive, medicul Y determină 30% din cazuri ca fiind pozitive. Din totalul pacienților examinați 10% din cazuri sunt confirmate cu diagnostic pozitiv de ambii medici, iar 20% din cazuri au diagnostic negativ dat de medicul X şi diagnostic pozitiv dat de medicul Y.

Considerăm evenimentele: A={medicul X să pună un diagnostic pozitiv la un pacient aleator} şi B={ medicul Y pune un diagnostic pozitiv la un pacient aleator }.

i) Care este probabilitatea ca medicul Y să pună un diagnostic pozitiv la un pacient ales aleator ştiind că medicul X a pus un diagnostic pozitiv?

ii) Care este probabilitatea ca medicul Y să pună un diagnostic pozitiv la un pacient ales aleator ştiind că medicul X a pus un diagnostic negativ?

La punctul:

i) Probabilitatea cerută se notează cu Pr(B/A) şi se calculează astfel:

$$\Pr(B/A) = \frac{\Pr(A \cap B)}{\Pr(A)} = \frac{0{,}10}{0{,}20} = 0{,}50$$

Interpretare: medicul Y va pune un diagnostic pozitiv în 50% din cazurile cu diagnostic pozitiv dat de medicul X.

ii) $$\Pr(B/nonA) = \frac{\Pr(nonA \cap B)}{\Pr(nonA)} = \frac{\Pr(nonA \cap B)}{1\text{-}\Pr(A)} = \frac{0{,}20}{0{,}80} = 0{,}25$$

Interpretare: medicul Y va pune un diagnostic pozitiv în 25% din cazurile cu diagnostic negativ dat de medicul X.

iii) În acest caz vorbim de riscul relativ (RR) al lui B dat de evenimentul A definit ca raport între două probabilități:

$$RR = \frac{\Pr(B/A)}{\Pr(B/nonA)} = \frac{0{,}50}{0{,}25} = 2$$

Interpretare: medicul Y este 2 ori mai probabil să pună un diagnostic pozitiv la un pacient când medicul X pune un diagnostic pozitiv față de situația când medicul X pune un diagnostic negativ.

Definiții:

Fiind date două evenimente arbitrare A şi B, **probabilitatea lui B condiționată de către A** este probabilitatea de a se realiza evenimentul B dacă în prealabil s-a realizat evenimentul A:

$$\Pr(B/A) = \frac{\Pr(A \cap B)}{\Pr(A)}$$

Formula probabilității totale:

$$\Pr(B) = \Pr(B/A) \cdot \Pr(A) + \Pr(B/nonA) \cdot \Pr(nonA)$$

În medicina clinică probabilitățile condiționate îşi găsesc aplicabilitatea într-un context special, cel al evaluării performantelor unui test diagnostic, indicatorii cunoscuți sub numele de sensibilitate (Se), specificitate (Sp), valoare predictivă pozitivă (VPP), valoare predictivă negativă (VPN). Dacă considerăm evenimentele T+={pacient aleator să aibă rezultat pozitiv la test diagnostic} şi M+={pacient aleator să aibă boala M} atunci:

$$Se = \Pr(T+/M+)$$
$$Sp = \Pr(T-/M-)$$
$$VPP = \Pr(M+/T+)$$
$$VPN = \Pr(M-/\mathrm{T}-)$$

O aplicație a probabilităților condiționate în medicina clinică este **riscul relativ (RR)** privit ca o măsură a relație dintre un factor de expunere F şi o boală M. Dacă considerăm evenimentele F+={pacient aleator să fie expus factorului} şi M+={pacient aleator să aibă boala M} atunci riscul relativ este raportul dintre două probabilități condiționate (probabilitatea de a face boala pentru cei expuşi şi probabilitatea de a face boala pentru cei neexpuşi):

$$RR = \frac{\Pr(B/A)}{\Pr(B/nonA)} = \frac{\Pr(M+/F+)}{\Pr(M+/F-)}$$

3.4 Legea lui Bayes și aplicațiile ei

Exemplul 3.16

Helicobacter pylori (Hpylori). Se ştie că 80% dintr-o populaţie de subiecți cu infecție Hpylori şi 10% dintre subiecții fără infecție sunt clasificați drept pozitivi de testul Antigen Hpylori. Știind că prevalența infecției cu Hpylori este de 74%, ne interesează să aflăm valoarea predictivă pozitivă şi valoarea predictivă negativă a testului antigen Hpylori.

Considerăm evenimentele:

HPT+={pacient ales aleator să aibă rezultat pozitiv a test}
HP+={pacient ales aleator să aibă infecție cu Helicobacter pylori}
HPT- ={pacient ales aleator să aibă rezultat negativ la test}
HP- ={pacient ales aleator să nu aibă infecție cu Helicobacter pylori}.
Se poate determina Se=Pr(HPT+/HP+)=0.80 şi Sp=Pr(HPT-/HP-)=1-0.10=0.90.

Conform formulei lui Bayes:

$$\Pr(B/A) = \frac{\Pr(A/B)\cdot \Pr(B)}{\Pr(A/B)\cdot \Pr(B) + \Pr(A/nonB)\cdot \Pr(nonB)}$$

Putem determina

$$VPP = \Pr(HP+/HPT+) =$$
$$= \frac{\Pr(HPT+/HP+)\cdot \Pr(HP+)}{\Pr(HPT+/HP+)\cdot \Pr(HP+) + \Pr(HPT+/HP-)\cdot \Pr(HP-)}$$
$$= \frac{0{,}80\cdot 0{,}74}{0{,}80\cdot 0{,}74 + (1-0{,}90)\cdot(1-0{,}74)} = 0{,}96$$

$$VPN = \Pr(HP-/HPT-) =$$
$$= \frac{\Pr(HPT-/HP-)\cdot \Pr(HP-)}{\Pr(HPT-/HP-)\cdot \Pr(HP-) + \Pr(HPT-/HP+)\cdot \Pr(HP+)}$$
$$= \frac{0{,}90\cdot(1-0{,}74)}{0{,}90\cdot(1-0{,}74) + (1-0{,}80)\cdot(0{,}74)} = 0{,}62$$

Definiție:

Pe baza formulei de calcul a probabilităților condiționate se deduce formula lui Bays:

$$\Pr(B/A) = \frac{\Pr(A/B)\cdot \Pr(B)}{\Pr(A/B)\cdot \Pr(B) + \Pr(A/nonB)\cdot \Pr(nonB)}$$

Aplicațiile formulei lui Bayes se reflectă în calculul VPP şi VPN dacă se cunoaşte prevalența (p) a unei boli (în anumite tipuri de studii medicale aceste valori nu se pot calcula direct din tabelul de contingență dar se pot deduce astfel):

$$VPP = \frac{Se \cdot p}{Se \cdot p + (1 - Sp) \cdot (1 - p)}$$

$$VPN = \frac{Sp \cdot (1 - p)}{Sp \cdot (1 - p) + p \cdot (1 - Se)}$$

3.5 Exerciții propuse

1. Dacă considerăm o familie cu doi copii aleasă aleator dintr-o populație şi definim evenimentele: A={mama are gripă}, B={tatăl are gripă}, C={primul copil are gripă}, D={al doilea copil are gripă}, E={cel puţin un copil are gripă}, F={cel puţin un părinte are gripă}, G={cel puţin o persoană are gripă}.

- Ce înseamnă A∪B?
- Ce înseamnă A∩B?
- C si D sunt incompatibile?
- Ce înseamnă C∪E?
- Ce înseamnă C∩E?
- Exprimaţi F în termenii A, B, C şi D
- Exprimaţi G în termenii E şi F.

2. Frecvențele grupelor sanguine pe populația României se prezintă astfel: 33% grupa 0, 43% grupa A, 16% grupa B şi 8% grupa AB. Probabilitatea ca un subiect ales aleator din populație să aibă grupa 0 sau A este:

A. 0,55
B. 0,76
C. 0,43
D. 0,70
E. 0,20

3. Frecvențele grupelor sanguine pe populația României se prezintă astfel: 33% grupa 0, 43% grupa A, 16% grupa B şi 8% grupa AB. Probabilitatea ca un subiect ales aleator din populație să nu aibă grupa 0 este:

A. 0,55
B. 0,70
C. 0,67
D. 0,72
E. 0,20

4. Frecvențele grupelor de vârstă a persoanelor peste 60 de ani pe populația României se prezintă astfel:

Grupa de vârstă (ani)	Frecvența (%)
≤19	21
20-29	13
30-39	15
40-49	14
50-59	14
60-69	11
≥70	12

Probabilitatea ca un subiect ales aleator din populație să fie până în 39 ani este:
A. 0,15
B. 0,70
C. 0,67
D. 0,49
E. 0,22

5. Prespunem că are loc un episod de epidemie de gripă într-o populație, astfel că în 10% dintre familii mama are gripă, în 15% dintre familii tata are gripă iar în 3% dintre familii ambii părinți au gripă. Dacă extragem aleator o familie din populație şi considerăm evenimentele A={mama are gripă} şi B={tata are gripă} atunci:
A. evenimentele A şi B sunt independente
B. evenimentele A şi B sunt dependente
C. evenimentele A şi B sunt incompatibile
D. evenimentele A şi B sunt compatibile
E. evenimentele A şi B sunt imposibile

6. Într-o populație adultă de 1300 subiecți, 500 subiecți s-au vaccinat împotriva gripei. In cursul unei epidemii de gripă, 10% dintre subiecți s-au îmbolnăvit de gripă, iar 3% dintre cei vaccinați au avut gripă.
i. Se alege aleator un subiect din populația de referință şi se consideră evenimentele: A={subiectul a fost vaccinat}, B={subiectul a avut gripă}, C={subiectul a fost vaccinat și a avut gripă}. Probabilitatea ca subiectul extras aleator să fi avut gripă sau să fi fost vaccinat este:
A. 0,470
B. 0,200
C. 0,067
D. 0,490
E. 0,202

ii. Se selectează în mod aleator unul dintre subiecții nevaccinați. Probabilitatea ca acesta să aibă gripă este:
A. 0,14
B. 0,20
C. 0,17
D. 0,39
E. 0,22

7. Din 300 de pacienți cu insuficiență renală cronică au existat 200 de subiecți care au consumat în exces medicamente metabolizate prin rinichi. Mai mult, un procent de 25% dintr-un eşantion reprezentativ de 1500 de subiecți din populația de referință sunt subiecți care au consumat în exces medicamente metabolizate prin rinichi. De câte ori este mai probabil ca un subiect care a consumat în exces medicamente să aibă insuficiență renală cronică față de un subiect fără consum excesiv de medicamente metabolizate prin rinichi?
A. 15
B. 6
C. 2
D. 3,15
E. 0,22

8. O firmă producătoare de medicamente face un studiu privind apariția unor posibile reacții adverse asociate cu un medicament. Pentru aceasta, selectează un eșantion de subiecți sănătoși dintre care: 85% sunt până în 49 de ani, iar 15% au fost peste 50 de ani. 18% dintre cei până în 49 de ani și 15% dintre cei peste 50 de ani au observat apariția unor efecte secundare. Se alege în mod aleator un subiect care a participat la acest studiu. Considerăm evenimentele:
A={subiectul are între 18-49 ani}
B={subiectul are peste 50 ani}
R={subiectul a observat apariția unor reacții adverse}
Probabilitatea de a alege un subiect care nu a avut reacții secundare este:
A. 0,82
B. 0,59
C. 0,17
D. 0,49
E. 0,60

9. În farmacii s-au introdus aparate automate de determinare a tensiunii arteriale. Un astfel de aparat clasifică 70% dintre hipertensivi şi 13% din normotensivi ca având hipertensiune arterială. Dacă 30% din populatia adultă are HTA atunci VPP şi VPN ale acestui aparat sunt:
A. VPP=0,55 şi VPN=0,75
B. VPP= 0,70 şi VPN=0,87
C. VPP=0,65 şi VPN=0,90
D. VPP=0,87 şi VPN=0,70
E. Nu se pot determina

10. Într-un studiu de evaluare a performanței testului de toleranță la glucoza pentru diagnosticul diabetului, printr-o culegere de tip eşantion reprezentativ (fiecare subiect a avut aceeaşi sansă de a fi ales în eşantion) dintre cei 500 de diabetici 400 au fost depistați de test iar din cei 100 de subiecti sănătoşi 18 au fost incorect clasificați ca fiind diabetici. Sensibilitatea şi respectiv specificitatea testului sunt:
A. Se=0,95 şi Sp=0,85
B. Se= 0,77 şi Sp=0,87
C. Se=0,99 şi Sp=0,82
D. Se=0,67 şi Sp=0,57
E. Nu se pot determina

11. Prespunem că 92% dintre femeile însărcinate care își fac un test de diagnosticare a toxoplasmozei obțin un rezultat pozitiv al acestuia în timp ce 94% dintre femeile care nu sunt însărcinate obțin un rezultat negativ al testului. Prevalența toxoplasmozei este de 25%. Prespunând că o femeie însărcinată extrasă aleator din populația de referință are rezultat pozitiv, care este probabilitatea ca ea să aibă efectiv boala?

A. 0,84
B. 0,50
C. 0,92
D. 0,35
E. Date insuficiente pentru a se putea calcula

4. Studiul evenimentelor repetitive

Obiective educaționale. După parcurgerea acestui capitol, studenții:

- vor putea să determine distribuția de probabilitate a unei variabile aleatoare finite pe baza unui scenariu dat;
- vor putea să determine speranța matematică, varianța și abaterea standard asociate unei variabile aleatoare finite;
- vor putea identifica condițiile de aplicare ale legii binomiale și în aceste condiții vor putea calcula probabilitatea diferitelor evenimente asociate;
- vor putea identifica caracteristicile distribuției normale și vor putea calcula diferite valori numerice asociate acestei distribuții;
- vor putea descrie principiile teoremei limitei centrale;
- vor putea enumera principalii estimatori fără bias utilizați în inferența statistică;
- vor putea calcula intervalul de încredere asociat unei medii și frecvențe.

4.1 Variabile aleatoare

Din punct de vedere al studiului probabilităților evenimentele sunt rezultatul unor experimente. Fenomenele naturale complexe au multiple posibilități de obținere a unuia sau a mai multor rezultate iar modelarea matematică cantitativă a rezultatelor posibile și studiul lor probabilistic se face cu ajutorul variabilelor aleatoare. Caracterul aleator al acestor variabile este generat de diferențele care apar între membrii diferitelor eșantioane, la fiecare repetare a unui experiment.

Denumirea de variabilă aleatoare nu implică faptul că valorile variabilei (caracteristicile) sunt întâmplătoare, ci faptul că experimentul este atât de complex încât nu îi putem anticipa rezultatul; de exemplu nu putem calcula cu precizie greutatea la maturitate pentru un nou născut.

Exemplu 4.1

Se dorește determinarea frecvenței numărului de episoade de gripă într-o comunitate. În acest scop, un număr de 100 de adulți au fost chestionați legat de numărul de episoade de gripă din ultimul an. 65 dintre aceștia au declarat că au suferit un episod de gripă, 30 două episoade, iar 5 persoane au declarat că au suferit 3 episoade de gripă. Considerând E mulțimea persoanelor chestionate, funcția f:E → {1, 2, 3}, funcție care asociază fiecărui membru al eșantionului numărul de episoade de gripă pe care l-a declarat, este variabila aleatoare asociată acestui proces de culegere de date.

Pentru un alt eșantion de 100 de adulți chestionați relativ la numărul de episoade de gripă din ultimul an, 26 dintre aceștia au declarat că nu au suferit de gripă, 70 au declarat un episod de gripă, iar 4 persoane au declarat ca au suferit 3 episoade de gripă. În acest caz, considerând E mulțimea persoanelor chestionate, funcția f:E → {0, 1, ,3}, funcție care asociază fiecărui membru al eșantionului numărul de episoade de gripă pe care l-a declarat, este variabila aleatoare asociată acestui proces de culegere de date.

Exemplu 4.2

Se dorește estimarea nivelului de trai al unei comunități exprimat prin venitul mediu pe familie. În acest scop, s-au studiat 50 de familii din respectiva comunitate. Considerând E mulțimea formată din familiile luate în studiu, funcția f, f:E→R, funcție care asociază fiecărei familii suma veniturilor membrilor familiei raportată la numărul de membri ai familiei, este variabila aleatoare asociată acestui proces de culegere de date.

Variabilele aleatoare pot fi de trei tipuri: finite, discrete și continue, acest lucru corespunzând numărului de valori în care acestea pot să se manifeste, atunci când experimentul s-ar repeta (ipotetic) la infinit.

Mulțimea de valori pe care o variabilă aleatoare o poate lua se numește spațiu fundamental.

Variabila aleatoare descrisă în exemplul 4.1 este de tip discret, finit, deoarece mulțimea posibilă de valori este {0, 1, 2, ... , n}, unde n este un număr natural mai mic de 366 (presupunând că un episod de gripă trebuie să dureze minim o zi, iar anul calendaristic are maxim 366 zile) în timp ce variabila aleatoare descrisă în exemplul 2 este de tip continuu – mulțimea de valori pe care această variabilă poate să o ia este nenumărabilă.

Pentru o variabilă aleatoare finită, asociată unui proces de colectare de date, valorile posibile, pe care această variabilă le poate lua, formează o serie statistică. Astfel, pentru fiecare valoare se poate determina frecvența relativă a acesteia. Pentru primul eșantion din exemplul 1, valoarea 1 are frecvența relativă 65/100, valoarea 2 are frecvența relativă 30/100, iar valoarea 3 are frecvența relativă 5/100. Aceste valori corespund probabilității ca luând la întâmplare o persoană din eșantion, ea să fi suferit una, două respectiv trei episoade de gripă.

Având o variabilă aleatoare finită X, cu mulțimea de valori $\{x_1 , x_2 , ... , x_n\}$, mulțimea de probabilități $\{Pr(x_1), Pr(x_2), ... , Pr(x_n)\}$ se numește distribuția de probabilitate asociată variabilei aleatoare X. Distribuția de probabilitate a unei variabile aleatoare X se mai notează prin:

$$X: \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ Pr(x_1) & Pr(x_2) & \dots & Pr(x_n) \end{pmatrix}$$

Interpretarea distribuției de probabilitate este următoarea: valoarea x_1 poate fi întâlnită cu probabilitatea $Pr(x_1)$, valoarea x_2 poate fi întâlnită cu probabilitatea $Pr(x_2)$, etc. Pentru orice variabilă aleatoare $Pr(x_1) + Pr(x_2) + ... + Pr(x_n) = 1$.

Distribuția de probabilitate face legătura dintre valorile unei variabile aleatoare și probabilitățile lor de apariție. Distribuția de probabilitate se poate prezenta sub formă de listă, tabel sau grafic.

Pentru primul eșantion prezentat în exemplul 4.1, distribuția de probabilitate va fi

$$X: \begin{pmatrix} 1 & 2 & 3 \\ 65/100 & 30/100 & 5/100 \end{pmatrix}$$

iar pentru al doilea eșantion, distribuția de probabilitate va fi

$$X: \begin{pmatrix} 0 & 1 & 3 \\ 26/100 & 70/100 & 4/100 \end{pmatrix}$$

Speranța matematică a unei variabile aleatoare finite X, numită și media variabilei aleatoare finite, se definește astfel:

$$\bar{X} = \sum_{i=1}^{n} x_i * p(x_i)$$

Varianța variabilei aleatoare X, având speranța matematică $\bar{X}$ se notează cu S^2 și se calculează în modul următor:

$$S^2 = \sum_{i=1}^{n} (x_i - \bar{X})^2 * p(x_i)$$

Abaterea standard a variabilei aleatoare X se notează cu S și se calculează S= $\sqrt{S^2}$

Funcția de probabilitate, definită ca funcția F: R → R, definită prin

F(k) = Probabilitate (X ≤ k) = $\sum_{x_i \leq k} p(x_i)$.

Exemplu 4.3

Se dorește estimarea costului tratamentului gripei pe un an într-o comunitate. Distribuția de probabilitate reprezentând numărul de episoade de gripă într-un an este

$$X: \begin{pmatrix} 0 & 1 & 2 & 3 \\ 0{,}25 & 0{,}60 & 0{,}1 & 0.05 \end{pmatrix}.$$

Care este probabilitatea ca luând la întâmplare o persoană, aceasta să nu fi făcut decât maxim două episoade de gripă? Știind că costul tratamentului unui episod de gripă este de 200 lei și că comunitatea are 5000 de membri, să se calculeze intervalul în care se așteaptă să se situeze costul tratamentului gripei pentru această comunitate.

Probabilitatea căutată este dată de valoarea funcției de repartiție F(2) = Pr(X ≤ 2) = 0,25+0,6+0,1 = 0,95

Intervalul dorit este dat de formula:

(Nr. așteptat de episoade/persoană ± abatere standard) * cost/episod* nr. persoane

unde numărul așteptat de episoade este speranța matematică a variabilei aleatoare X, abaterea standard este abaterea standard a variabilei aleatoare X, cost/episod=200 lei, iar numărul persoanelor este 5000.

$$\bar{X} = 0 * 0{,}25 + 1 * 0{,}6 + 2 * 0{,}1 + 3 * 0{,}05 = 0{,}95$$

$S^2 = (0\text{-}0{,}95)^2*0{,}25 + (1\text{-}0{,}95)^2*0{,}6 + (2\text{-}0{,}95)^2*0{,}1 + (3\text{-}0{,}95)^2*0{,}05 \approx 0{,}55$

$$S=\sqrt{0{,}55} \approx 0{,}74$$

Deci costul așteptat al tratamentului gripei pe un an se situează între (0,95 – 0,74) * 200 * 5000 și (0,95 + 0,74) * 200* 5000, adică [210000 lei, 1690000 lei]

Noţiunile şi proprietăţile prezentate anterior pentru variabilele aleatoare finite se pot introduce în mod analog pentru cazul variabilelor aleatoare discrete având o mulţime infinită de valori, sau pentru cazul variabilelor aleatoare continue, prin înlocuirea sumei finite cu una infinită, precum se poate observa în figura 6.1, studiul modului de calcul al valorilor asociate acestor tipuri de distribuții nefiind obiectul acestui curs.

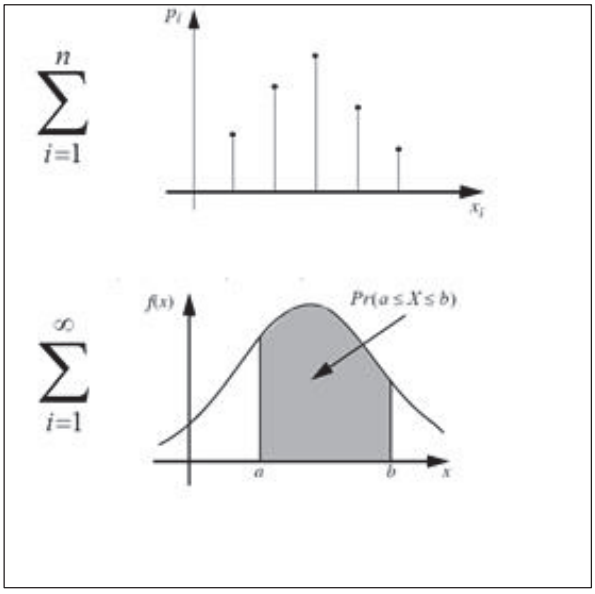


Fig. 6.1 Reprezentare grafică pentru distribuția de probabilitate discretă (sus) sau continuă (jos)

4.2 Cele mai importante distribuții de probabilitate și aplicațiile lor

În practică, foarte des se încearcă încadrarea distribuțiilor de probabilitate asociate diferitelor experimente în modele cunoscute în funcție de modelul matematic ales care pot fi grupate în distribuții discrete: binomială , Poisson și continue: normală, Student.

Una dintre cele mai utilizate distribuții de probabilitate asociate variabilelor aleatoare finite este legea binomială, lege care are la bază următorul model:

- Experimentul constă din efectuarea unei încercări elementare de un număr dat și finit de ori – n
- Experimentul poate genera doar două rezultate, cu probabilitățile p, respectiv q=1-p constante de la o încercare la alta
- Cele n încercări elementare sunt independente.

Dacă variabila aleatoare poate lua valorile {0, 1, 2, ... , n} atunci $Pr(X=k)= C_n^k p^k q^{n-k}$ iar speranța matematică și varianța au valorile n*p, respectiv n*p*q.

Exemplu 4.4

Un vaccin provoacă reacție adversă (febră) cu o probabilitate de 0,3. Care este probabilitatea ca vaccinând 5 copii aleși la întâmplare, la 2 dintre ei să apară reacție adversă?

Condițiile de lucru:

- Este vorba despre un experiment (vaccinare) care se repetă de un număr finit de ori (5)
- De fiecare dată experimentul poate produce doar următoarele rezultate – „Reacție adversă", cu probabilitatea p = 0,3 și „Fără reacție adversă", cu probabilitatea q=1-p=0,7. Aceste probabilități rămân constante pentru fiecare dintre cele 5 vaccinări
- Cele cinci repetări ale experimentului sunt independente

Probabilitatea căutată „din 5 copii la 2 să apară reacție adversă" este:

$$Pr(X=2) = C_5^2 * 0,3^2 * 0,7^{5-2} = \frac{5!}{2!*(5-2)!} * 0,3^2 * 0,7^3 = \frac{1*2*3*4*5}{(1*2)*(1*2*3)} * 0,09 * 0,343 = 0.3087$$

Distribuția normală este una dintre cele mai întâlnite distribuții ale unei variabile aleatoare continue din natură, iar reprezentarea sa grafică poartă și denumirea de „curba lui Gauss" sau „clopotul lui Gauss". Această reprezentare grafică este prezentată în figura 6.2.

Variabila aleatoare normală este o variabilă aleatoare continuă, definită la nivelul populației, cu o funcție de probabilitate care depinde de doi parametri (media în populație notată μ respectiv deviația standard în populație, notată σ) ce are formula:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

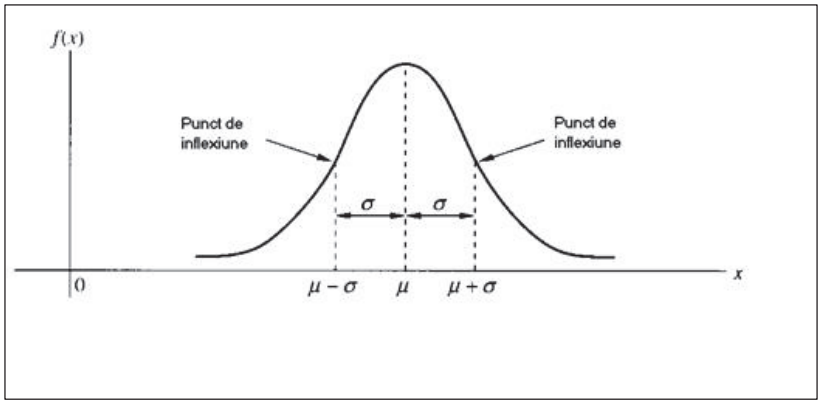


Fig. 6.2 Reprezentare grafică a distribuției normale

Formulările „datele sunt normal distribuite" sau „datele urmează legea normală" indică faptul că în respectiva situație predomină valorile grupate în jurul mediei aritmetice, valorile extreme întâlnindu-se cu o probabilitate mai redusă.

Distribuția normală are următoarele proprietăți:

- Este simetrică față de media aritmetică (media este egală cu mediana);
- Are ca punct de maxim media aritmetică (media aritmetică este egală cu modulul);
- În intervalul medie±abatere standard se regăsesc aproximativ 68% din observaţii;
- În intervalul medie±2*abatere standard se regăsesc aproximativ 95% din observaţii;
- În intervalul medie±3*abatere standard se regăsesc aproximativ 99.9% din observaţii.

Exemplu 4.5

Cunoscând faptul că, în condiții normale, media lungimii nou născuților băieți este de 51 cm, cu o abatere standard de 3 centimetri, care este probabilitatea ca un nou născut băiat să aibă lungimea în intervalul 54 cm-57 cm?

Rezolvare: În condiții normale 68% din nou născuți vor avea lungimea în intervalul [51-3cm, 51+3cm], iar 95% din nou născuți vor avea lungimea în intervalul [51-2*3cm, 51+2*3cm] adică [48 cm, 54 cm], respectiv [45 cm, 57 cm]. Deci în interiorul intervalului [45 cm, 57 cm], dar în exteriorul intervalului [48 cm , 54 cm] se regăsesc în condiții normale 95%-68% = 27% din observații. Datorită faptului că distribuția normală este simetrică, putem concluziona că în intervalul [54 cm, 57 cm] se vor regăsi $\frac{27}{2}\%$ din observații, deci probabilitatea căutată este 0,135.

4.3 Distribuții de eșantionare

Inferența statistică permite obținerea de informații referitoare la populație, plecând de la studiul unei submulțimi ale acesteia (eșantion). Acesta este un proces inductiv, care permite cercetătorilor să extrapoleze concluziile referitoare la populație plecând de la concluziile obținute dintr-un eșantion extras din respectiva populație.

Noțiunea de distribuție de eșantionare este o noțiune teoretică, care permite trecerea de la distribuția de probabilitate observată într-un eșantion (finită) la distribuția în populație (necunoscută).

Exemplu 4.6

Într-o universitate, se dorește evaluarea capacității de efort fizic al studenților. În acest scop, s-a selectat aleator un grup de 50 de studenți care au efectuat o serie de teste, capacitatea de efort fiind cuantificată printr-un scor individual (variabilă numerică).

Pentru membrii acestui eșantion, studiind distribuția de probabilitate a variabilei aleatoare generate de această culegere de date, cercetătorii pot să răspundă la diferite întrebări de interes – care este capacitatea de efort așteptată, între ce limite se situează aceasta, care este procentul de studenți care au capacitatea de efort peste (sub) un anumit scor dat, etc. Această distribuție de probabilitate se numește distribuția eșantionului. Prin studiul distribuției eșantionului se obțin informații referitoare la forma distribuției, precum și referitor la tendințele de centralitate și dispersie la nivelul eșantionului.

Pentru a răspunde la aceleași întrebări la nivelul populației, cercetătorii ar trebui să studieze o altă distribuție – distribuția populației. Datorită motivelor obiective, această distribuție nu poate fi studiată, ea fiind rezultatul culegerii de date de la toți membrii populației. Nu există certitudinea că distribuția eșantionului studiat este măcar asemănătoare cu distribuția populației.

Dacă s-ar repeta acest experiment, extrăgând din populație un alt eșantion de 50 de studenți se vor obține alte rezultate. Teoretic, s-ar putea extrage toate eșantioanele posibile de 50 de studenți din univeristate, obținându-se, astfel, măsurile de centralitate și dispersie pentru toate eșantioanele posibil de extras din respectiva populație. Acest lucru este posibil doar în teorie, deoarece numărul de eșantioane de talie fixată care se pot extrage dintr-o populație depășește cu mult talia populației. Spre exemplificare, pentru o universitate cu 1000 de studenti, numărul de eșantioane de talie 50 care se poate extrage este egal cu C^{50}_{1000} , număr care are în componență 85 de cifre! Variabila aleatoare (teoretică) care ar fi generată de parametrul de interes (de exemplu media aritmetică) pentru toate eșantioanele de talie fixată ce pot fi extrase din populație va avea distribuția de probabilitate numită distribuția de eșantionare.

Teorema limitei centrale face legătura între distribuția de eșantionare și distribuția populației.

Esența teoremei limitei centrale este faptul că valorile obținute pe baza eșantionării repetate pentru o talie fixată, suficient de mare, sunt modelate de o variabilă aleatoare normal distribuită, având ca parametrii media μ a populației și ca abatere standard raportul $\frac{s}{\sqrt{n}}$, distribuție ce permite astfel estimarea a cât de probabilă este obținerea respectivei valori a mediei într-o populație.

4.4 Estimarea parametrilor statistici

Din motive cunoscute, nu se pot utiliza populații pentru a realiza studii statistice. Pe de altă parte, rezultatele studiilor trebuie să reflecte caracteristicile populației. De obicei, pentru studiul într-o populaţie P a parametrilor unei caracteristici oarecare (cantitative sau calitative), este necesar să se urmeze procedeul:

- Se extrage un eşantion reprezentativ al acestei populaţii.
- Prin mijloacele statisticii descriptive se descrie distribuţia de probabilitate pe eşantionul extras la etapa 1, fiindcă talia acestuia permite o investigare exhaustivă a sa. Astfel, se determină frecvenţa observată, dacă este vorba de o caracteristică calitativă, sau se calculează media şi varianţa, în cazul unei caracteristici cantitative.
- Prin mijloacele statisticii inferenţiale sau inductive se extind la întreaga populaţie rezultatele observate pe eşantion. Adica, pornind de la parametrii observaţi (frecvenţa, media, varianţa, etc) pe eşantion se încearcă să se estimeze parametri necunoscuți ai întregii populaţii.

Această extindere la nivel de populație se realizează cu ajutorul estimatorilor. Principalii estimatori statistici utilizați sunt: frecvența – pentru variabilele calitative, și media aritmetică și varianța de eșantionare pentru variablele cantitative.

Valorile estimatorilor pe datele de cercetare (pe eșantioane) vor fi numite estimări punctuale.

Un estimator este fără bias – adică corect – dacă el poate genera o distribuție de eșantionare a cărei speranță matematică să fie egală cu valoarea reală în populație a parametrului estimat.

Media aritmetică și frecvența sunt estimatori fără bias. Varianța descriptivă, în schimb nu este.

La fel ca și varianța descriptivă, varianța de eșantionare (varianța) se notează de asemenea cu S^2 și se calculează conform formulei:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1},$$

unde $\bar{X}$ reprezintă media aritmetică a seriei statistice asociate eșantionului și n este talia eșantionului. Varianța este un estimator fără bias.

4.5 Intervalul de incredere

Un estimator este cu atât mai eficace cu cât varianța lui este mai mică, deci precizia unui estimator depinde de mărimea variantei sale. Prin procesul de estimare punctuală, se vor obține valori tributare fluctuațiilor de eșantionare, valori care pot fi la mare distanță de valoarea adevărată a parametrului care trebuie să fie estimat. De aceea, este recomandat să se facă această estimare nu ca o singură valoare, ci ca un interval, interval în care, cu o probabilitate ridicată, valoarea parametrului cercetat să se regăsească. Acest interval, numit și interval de încredere, este un interval mărginit, care include estimarea punctuală a parametrului studiat. Cu cât intervalul este mai larg, probabilitatea ca să conțină valoarea adevărată (necunoscută) a parametrului în populație va crește. Mărimea încrederii este dată de probabilitatea ca valoarea respectivă să se regăsească în respectivul interval.

În ceea ce urmează, se va prezenta modul în care se utilizează intervalele de încredere pentru estimarea mediei și frecvenței într-o populație.

Pentru estimarea unei medii (necunoscute) în populație, este necesar să se urmeze următorii pași:

- Din populaţia P se extrage eşantionul E reprezentativ de talie n;
- In eşantionul E pentru variabila X se observă o medie m şi o varianţă S^2;
- Se încearcă să se estimeze valoarea necunoscută a mediei în populație μ cu ajutorul lui m şi S^2 observate, calculându-se un interval în care să putem afirma cu o probabilitate ridicată că va conține valoarea μ.

În cazul intervalelor cu talia mai mare decât 30, intervalul de încredere se calculează conform formulei:

$$\left[m - Z_\alpha * \frac{S}{\sqrt{n-1}} ; m + Z_\alpha * \frac{S}{\sqrt{n-1}}\right], \text{ unde}$$

- m reprezintă media aritmetică,
- S reprezintă abaterea standard de eșantionare,
- n reprezintă talia eșantionului,
- Z_α reprezintă o constantă (cunoscută), depinzând de gradul de încredere cu care se face estimarea.

Cel mai utilizat grad de încredere din studiile medicale este 95%, acestui grad corespunzându-i constanta $Z_{95\%}$ = 1,96. În acest caz, intervalul de încredere cu o precizie de 95% este: $\left[m - 1{,}96 * \frac{S}{\sqrt{n-1}} ; m + 1{,}96 * \frac{S}{\sqrt{n-1}}\right]$.

În cazul intervalelor mici (talia mai mică decât 30) intervalul de încredere se calculează conform formulei

$$\left[m - T_\alpha * \frac{S}{\sqrt{n-1}} ; m + T_\alpha * \frac{S}{\sqrt{n-1}}\right], \text{ unde}$$

- m reprezintă media aritmetică,
- S reprezintă deviația standard,
- n reprezintă talia eșantionului,

- T_α reprezintă o constantă (cunoscută), depinzând atât de gradul de încredere cu care se face estimarea cât și de mărimea eșantionului.

Exemplu 4.7

Într-un eşantion de 101 de nou născuţi, cunoaştem media greutăţii la naştere: 3400g şi abaterea standard: 142. Să se determine cu o precizie de 95% intervalul de încredere generat de acest eșantion.

Rezolvare: Intervalul de încredere generat de acest eșantion este:

$\left[3400 - 1{,}96 * \frac{142}{\sqrt{101-1}} ; 3400 + 1{,}96 * \frac{142}{\sqrt{101-1}}\right]$, cu aproximație [3372; 3428].

Deci, putem afirma cu o încredere de 95 % că media greutății la naștere în populația din care a fost extras eșantionul se găsește în intervalul [3372; 3428], existând o probabilitate de 5% ca media în populație să fie în afara acestui interval.

Pentru estimarea frecvenței necunoscute p în populație se urmează următorul raționament:

- Din populaţia P se extrage eşantionul E reprezentativ de talie n;
- În eşantionul E pentru variabila X se observă o frecvenţă relativă f;
- Se încearcă să se estimeze valoarea necunoscută a lui p cu ajutorul lui f observat.

În cazul eșantioanelor mari (f*n> 10 și (1-f)*n>10), intervalul de încredere este:

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

unde Z_α este o constantă (cunoscută), depinzând de gradul de încredere cu care se face estimarea.

Cel mai utilizat grad de încredere utilizat în studiile medicale este 95%, acestui grad corespunzându-i constanta $Z_{95\%} = 1{,}96$.

Exemplu 4.8

Se realizează un studiu pentru a se estima prevalența hipertensiunii arteriale (HTA) într-o comunitate. Pe un eșantion de 100 persoane, s-a observat o prevalență de 10%. Între ce limite este de așteptat ca HTA să se regăsească în comunitate.

Rezolvare: Răspunsul la întrebarea de cercetare este dat de intervalul de încredere cu precizie de 95% generat de prevalența observată în eșantion. Deci, prevalența în comunitate va fi în intervalul:

$\left[0{,}1 - 1{,}96 * \sqrt{\frac{0{,}1(1-0{,}1)}{100}}; 0{,}1 + 1{,}96 * \sqrt{\frac{0{,}1(1-0{,}1)}{100}}\right]$, adică aproximativ [0,04; 0,16]

Dacă se ridică gradul de încredere, constanta (Z,T) crește și ea, atrăgând după sine un interval de încredere mai larg. Pentru același eșantion, intervalul de încredere generat cu o încredere de 97% îl va include pe intervalul de încredere generat cu o încredere de 95%.

Crescând numărul de persoane în eșantion, precizia estimării este de așteptat să crească.

Pentru estimarea unei variabile cantitative, un eșantion cu abatere standard mare va genera un interval de încredere larg, deci o precizie redusă a estimării mediei.

4.6 Exerciții propuse

1. Având următoarea distribuție de probabilitate

$$X : \begin{pmatrix} 90 & 120 & 150 & 180 \\ 5/35 & 11/35 & 8/35 & 11/35 \end{pmatrix},$$

speranța matematică a variabilei aleatoare X este (aproximativ):

A. 120.5
B. 141.4
C. 150.6
D. 99.3
E. 127.6

2. Pe un eșantion de 65 de persoane, s-a măsurat greutatea (kg) și s-a observat greutatea medie 75 cu o abatere standard 16. Pentru un grad de încredere de 95%, intervalul de încredere este (aproximativ):

A. [59; 91]
B. [43; 107]
C. [71; 79]
D. [50; 100]
E. [74; 76]

3. Într-un eșantion de 81 persoane s-a observat prevalența unei boli, aceasta având valoare de 35%. Cu un risc de eroare de 5%, în populația din care a fost extras eșantionul, prevalența este cuprinsă în intervalul:

A. [30%; 40%]
B. [24.6%; 45,4%]
C. [32.4%; 37,6%]
D. [27.4%; 42,6%]
E. [10%, 60%]

4. Ce se întâmplă cu amplitudinea intervalului de încredere generat pe baza unui eșantion atunci când gradul de încredere crește de la 95% la 97%

A. Amplitudinea nouă este mai mică decât amplitudinea inițială
B. Amplitudinea nouă este mai mare decât amplitudinea inițială
C. Amplitudinea nouă este cu 2% diferită de amplitudinea inițială
D. Amplitudinea nouă este la fel cu amplitudinea inițială
E. Amplitudinea se va modifica, dar nu se poate spune în ce direcție

5. Testarea ipotezelor statistice

5.1 Formularea de noi ipoteze

Cercetarea medicală caută răspunsuri la întrebări privind starea de sănătate și factorii implicați în apariția și evoluția problemelor de sănătate.

Unele întrebări pot fi analizate sau testate statistic pentru a afla un răspuns, altele nu.

Întrebările care nu pot fi analizate sau testate statistic pentru a afla un răspuns sunt cele de tipul „În ce mod percep pacienții o anumită problemă de sănătate?" sau „Care sunt problemele și soluțiile sugerate de către comunitate în sistemul actual de sănătate?". Asemenea întrebări își pot găsi răspunsuri prin realizarea unor studii calitative, fără utilizarea unor teste sau analize statistice cantitative.

Întrebările care pot fi analizate sau testate statistic sunt formulate de obicei după următoarele modele generice:

1. „Există o corelație între anumiți parametri biologici?"
2. „Există o legătură semnificativă între expunerea la anumiți factori și apariția unui anumit efect?"
3. „Există o diferență semnificativă între anumite grupuri în ceea ce privește prezența anumitor caracteristici?"

Este important ca aceste întrebări să fie **formulate clar și concret**, cu precizarea atentă a parametrilor, factorilor, efectelor, grupurilor sau caracteristicilor studiate, astfel încât să permită apoi **testarea existenței unor posibile corelații** (primul tip de întrebare), respectiv formularea și **testarea unor posibile ipoteze** (întrebările de tipul 2 sau 3).

În aceeași măsură, este importantă diferențierea dintre întrebarea / ipoteza de cercetare și ipoteza statistică, în sensul în care fiecare ipoteză de cercetare poate fi separată într-una sau mai multe ipoteze statistice. Diferența imediată, observabilă, între cele două tipuri de ipoteze este că cele de cercetare se referă la situații medicale în timp ce ipotezele statistice sunt strict matematice și reprezintă transpunerea în termeni matematici (statistici) a unei părți sau a întregii idei de cercetare.

Ipoteza H_1, numită **ipoteza alternativă**, reprezintă răspunsul așteptat de către medicul cercetător atunci când decide să caute un răspuns la ipoteza formulată. Însă această ipoteză nu poate fi testată direct.

Orice testare statistică se realizează de fapt asupra unei **ipoteze nule** (notată H_0) care contrazice ipoteza studiului (notată H_1), prin măsurarea dovezilor existente împotriva acesteia, iar rezultatul testării ne va permite fie „să respingem ipoteza nulă și să considerăm temporar adevărată ipoteza alternativă", fie „să nu putem respinge ipoteza nulă", fără însă

a o putea accepta, întrucât lipsa dovezilor ce susțin respingerea ipotezei nule nu este o dovadă suficientă pentru a o putea considera validă.

Subliniem importanța evitării termenului de „acceptare", chiar și în cazul ipotezei alternative, întrucât acest termen are conotații absolute: o idee acceptată nu mai poate fi pusă în discuție; întreaga cercetare științifică are însă la bază verificarea de ipoteze „mai bune" decât cele anterioare și nu demonstrarea unui adevăr absolut. Ipoteza anterioară (nulă) poate sau nu poate fi respinsă în determinismul cercetării științifice, iar ipoteza formulată de cercetător (alternativă) poate fi sau nu validată (confirmată / păstrată până la apariția unor eventuale dovezi contrare).

Exemplul 5.1.

Vrem să verificăm întrebarea clinică "Există o diferență semnificativă între copiii respiratori orali și copiii respiratori nazali în ceea ce privește dezvoltarea transversală a maxilarului la vârsta de 12 ani ?"

În acest sens putem formula următoarele două răspunsuri (ipoteze), formulate în termeni medicali / clinici:

- Nu există o diferență semnificativă între copiii respiratori orali și copiii respiratori nazali în ceea ce privește dezvoltarea transversală a maxilarului (sub forma distanței bizigomatice măsurate în milimetri) la vârsta de 12 ani. – ipoteza nulă clinică

- Există o diferență semnificativă între copiii respiratori orali și copiii respiratori nazali în ceea ce privește dezvoltarea transversală a maxilarului (sub forma distanței bizigomatice măsurate în milimetri) la vârsta de 12 ani. – ipoteza alternativă clinică

Aceste ipoteze clinice pot fi formulate și în termeni statistici - **ipotezele statistice** (notate cu H_0 și H_1):

- H_0: Nu există o diferență semnificativă statistic între *media* dimensiunii transversale a maxilarului copiilor respiratori orali (sub forma distanței bizigomatice măsurate în milimetri) față de cea a copiilor respiratori nazali, la vârsta de 12 ani. – ipoteza nulă statistică

- H_1: Există o diferență semnificativă statistic între *media* dimensiunii transversale a maxilarului copiilor respiratori orali (sub forma distanței bizigomatice măsurate în milimetri) față de cea a copiilor respiratori nazali, la vârsta de 12 ani. – ipoteza alternativă statistică

Odată formulată în mod clar și concret, ipoteza nulă statistică poate fi testată prin aplicarea unui test statistic.

5.2 Paşii unui test statistic

Aplicarea unui test statistic presupune parcurgerea următorilor paşi:

1. **Formularea ipotezelor**
2. **Alegerea unui nivel prag de semnificație statistică**
3. **Alegerea testului statistic**
4. **Calcularea statisticii testului și a probabilității acesteia**
5. **Compararea statisticii calculate** a testului **cu valoarea critică** a statisticii testului SAU **compararea probabilității statisticii testului cu nivelul de semnificație** statistică ales la primul pas
6. **Luarea deciziei**

1. Formularea ipotezelor

Ipoteza nulă (H_0) și cea alternativă (H_1), trebuie formulate clar și concret, cu precizarea atentă a parametrilor, factorilor, efectelor, grupurilor sau caracteristicilor comparate, după cum am văzut în subcapitolul anterior.

2. Alegerea nivelului prag de semnificație statistică

Numit și nivel critic alfa (α), acest nivel prag reprezintă riscul de eroare pe care suntem dispuși să îl acceptăm în respingerea ipotezei nule.

În mod convențional, alegerea acestui prag se realizează frecvent la un nivel α=0,05, ceea ce corespunde unui risc asumat de a greși atunci când respingem ipoteza nulă, de cel mult 5 la sută.

Alegerea poate fi însă și mai precaută, ca de exemplu α=0,01, ceea ce corespunde unui risc asumat de a greși atunci când respingem ipoteza nulă, de cel mult 1 la sută.

3. Alegerea testului statistic corect trebuie realizată încă din etapa de planificare a studiului, în cadrul unui protocol riguros al cercetării, care prevede metodele de testare statistică a ipotezelor studiului, în funcție de tipul și caracteristicile variabilelor comparate pentru testarea acestor ipoteze, numărul de subiecți și natura datelor.

4. Calcularea statisticii testului și a probabilității acestei statistici se realizează prin algoritmi specifici fiecărui test statistic. Acești algoritmi vor fi prezentați în capitolele următoare (6, 7 și 8). Calculul acestei statistici și a probabilității observării unei asemenea statistici (valoarea p) sunt mult ușurate prin utilizarea programelor de analiză statistică computerizată.

5. Compararea statisticii calculate a testului **cu valoarea critică** a statisticii testului, găsită în tabele sau afișată de programul de analiză statistică, pe baza căreia se construiește un **interval critic**, numit și **interval de respingere** a ipotezei nule. Intervalul critic include valoarea critică și valorile mai extreme decât această valoare, care sunt progresiv mai puțin probabile, atunci când nu există o diferență semnificativă statistic între grupurile comparate. Statistica calculată a testului poate să aparțină sau nu acestui interval critic.

SAU: Compararea probabilității statisticii testului (p) **cu nivelul de semnificație** statistică ales la al doilea pas (α). Probabilitatea p poate fi mai mică, egală, sau mai mare decât nivelul de semnificație statistică α.

Dacă pentru ipoteza medicală testată nu se cunoaște direcționalitatea diferenței investigate atunci probabilitatea statisticii testului (p) luată în considerare trebuie să fie cea **bilaterală (two-tail p)**, fiind vorba despre o ipoteză în care nu se cunoaște cu certitudine semnul eventualei diferențe testate. Probabilitatea unilaterală (one-tail p) este utilizată în situațiile în care diferența testată nu poate fi decât fie mai mare, fie mai mică decât diferența stipulată prin ipoteza nulă.

Observație

Cunoașterea direcționalității favorizează obținerea unei semnificații statistice: înscrierea probabilității unilaterale (one-tail p) sub nivelul de semnificație statistică este mai ușor de obținut decât în cazul probabilității bilaterale (two-tail p).

Din acest motiv, atunci când nu există o direcționalitate a diferenței investigate, sau atunci când nu suntem siguri cu privire la existența unui anumit semn al acestei diferențe (ceea ce se întâmplă frecvent în cazul ipotezelor medicale), este mai sigură formularea unei ipoteze bilaterale și luarea în considerare a valorii bilaterale a lui p (two-tail p), pentru a reduce riscul de a respinge ipoteza nulă în mod eronat.

6. Luarea deciziei de a respinge sau nu ipoteza nulă poate fi realizată în baza oricăreia dintre cele două comparații anterioare:

- Dacă statistica calculată a testului aparține intervalului critic, ceea ce înseamnă că și valoarea probabilității acestei statistici este mai mică decât nivelul prag de semnificație statistică ($p<\alpha$), atunci **se respinge ipoteza nulă și se consideră că ipoteza alternativă este mai bună și poate fi păstrată**.
- Dacă statistica calculată a testului nu aparține intervalului critic, ceea ce înseamnă că și valoarea probabilității acestei statistici este mai mare sau egală cu nivelul prag de semnificație statistică ($p\geq\alpha$), atunci **nu se poate respinge ipoteza nulă**.

În mod uzual, luarea deciziei se realizează în funcție de **valoarea p (p-value)**, numită şi nivel de semnificaţie observat. Această valoare reprezintă probabilitatea de apariție a statisticii calculate a testului, sub premisa ipotezei nule.

Ipoteza nulă este respinsă dacă valoarea p este mai mică decât 0,05 (sau decât nivelul prag de semnificație ales).

Valoarea lui p ne indică nivelul de semnificație obținut în urma testării, fără a necesita repetarea testului statistic pentru diferite niveluri alese ale lui α:

- dacă $p > 0{,}05$ rezultatele sunt considerate **nesemnificative** statistic;
- dacă $0{,}05 < p \leq 0{,}1$ există o oarecare **tendinţă spre semnificaţie** statistică;
- dacă $0{,}01 < p \leq 0{,}05$ rezultatele sunt **semnificative** statistic;
- dacă $0{,}001 < p \leq 0{,}01$ rezultatele sunt **înalt semnificative** statistic;
- dacă $p \leq 0{,}001$ rezultatele sunt **foarte înalt semnificative** statistic.

Observație

De multe ori studiile corect desfășurate, dar cu un număr limitat de subiecți (sub cel necesar), conduc la valori p cuprinse între 0,05 și 0,1. Acesta este cazul pentru multe teze de licență ale studenților de la medicină și de aceea, în ciuda faptului că rezultatul nu este semnificativ statistic, comisia de susținere apreciază faptul că există tendința de obținere a unei semnificații statistice pe care candidatul este bine să o menționeze în concluziile lucrării.

Această terminologie se referă la rezultatele fiecărei testări în mod separat, și nu este corectă compararea între ele a rezultatelor unor testări, folosind termeni precum *"mai semnificativ decât ..."*.

Valoarea p **nu** reprezintă probabilitatea ca ipoteza nulă să fie adevărată, iar o valoare mică a lui p **nu** înseamnă că există o probabilitate mică pentru ca ipoteza nulă să fie adevărată. De asemenea, ipoteza testată nu poate fi inversată: de exemplu, pentru un p = 0,01 nu putem spune că avem o probabilitate de 0,99 ca diferenţa să existe. Așadar, valoarea p nu este un cuantificator al valorii de adevăr asociate ipotezei de lucru, ci se utilizează numai pentru a lua decizia de a respinge sau nu ipoteza nulă.

Pentru întrebarea clinică din exemplul 5.1, putem aplica un test statistic prin parcurgerea următorilor pași:

1. Formulăm ipotezele statistice:
    - H_0: Nu există o diferență semnificativă statistic între *media* dimensiunii transversale a maxilarului copiilor respiratori orali (sub forma distanței bizigomatice măsurate în milimetri) față de cea a copiilor respiratori nazali, la vârsta de 12 ani. – ipoteza nulă statistică
    - H_1: Există o diferență semnificativă statistic între *media* dimensiunii transversale a maxilarului copiilor respiratori orali (sub forma distanței bizigomatice măsurate în milimetri) față de cea a copiilor respiratori nazali, la vârsta de 12 ani. – ipoteza alternativă statistică
2. Alegem pragul de semnificație statistică la nivelul uzual: α=0,05.
3. Întrucât ne așteptăm să existe o distribuție normală (Gaussiană) a dimensiunii transversale a maxilarului în populația de copii studiată, alegem testul t – Student pentru eșantioane independente ca test adecvat pentru viitoarea comparație a

mediei dimensiunii transversale a maxilarului copiilor respiratori orali cu media dimensiunii transversale a maxilarului copiilor respiratori nazali.

4. Măsurăm și înregistrăm dimensiunea transversală a maxilarului (definită în acest studiu ca distanța bizigomatică măsurată în milimetri) a fiecărui copil dintr-un grup de copii respiratori orali care au fost urmăriți după criterii bine definite până la vârsta de 12 ani, precum și a fiecărui copil dintr-un grup de copii respiratori nazali, care a fost urmărit în paralel cu primul grup.
5. Urmând algoritmul testului, sau cu ajutorul unui program de analiză statistică, calculăm statistica t a testului Student ales anterior și probabilitatea (two-tail p) a acestei statistici și obținem: t = -9,52 și p = 1,04 x 10^{-14}.
6. Comparăm statistica calculată a testului (t = -9,52) cu valoarea critică a statisticii testului Student găsită în tabel sau afișată de programul de analiză statistică (t critical two-tail = 1,99) și constatăm că statistica calculată aparține intervalului critic $(-\infty ; -1,99] \cup [+1,99 ; +\infty)$.
   SAU: Comparăm probabilitatea calculată a statisticii testului (p = 1,04 x 10^{-14}) cu nivelul de semnificație statistică ales la primul pas (α=0,05) și constatăm că p<0,05, așadar un rezultat semnificativ statistic. De fapt, valoarea lui p este mult mai mică decât 0,05, ceea ce se notează ca p<0,001, pentru a sublinia că este vorba de un rezultat înalt semnificativ statistic.
7. Luarea deciziei:
   Întrucât statistica calculată a testului aparține intervalului critic respingem ipoteza nulă și validăm valoarea de adevăr a ipotezei alternative.
   SAU: Întrucât p<0,05, respingem ipoteza nulă și validăm valoare de adevăr a ipotezei alternative.

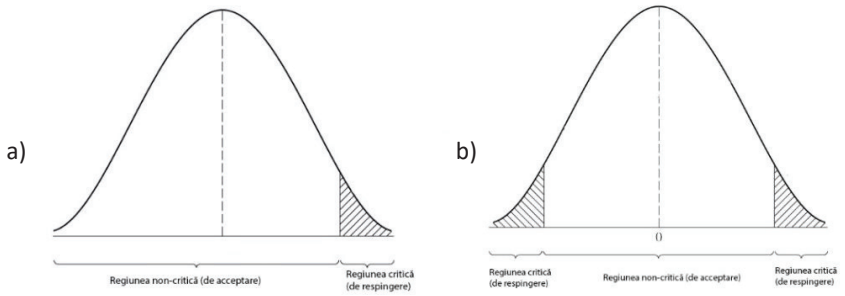


Figura 5.1. Regiunea (regiunile critice) pentru un test statistic aplicat unilateral (a) sau bilateral (b)

Așadar, la vârsta de 12 ani, există o diferență semnificativă statistic între media dimensiunii transversale a maxilarului copiilor respiratori orali (sub forma distanței bizigomatice măsurate în milimetri) față de cea a copiilor respiratori nazali.

Dacă valoarea calculată a parametrului statistic cade în regiunea critică a distribuției (regiunea de respingere) atunci ipoteza nulă se respinge, iar dacă valoarea este în regiunea non-critică (de acceptare) ipoteza nulă nu poate fi respinsă (figura 5.1).

5.3 Erori în aplicarea unui test statistic

Aplicarea unui test statistic poate fi asemănată cu un proces judiciar, în care ipoteza nulă este supusă judecății noastre cu ajutorul testului statistic.

Ca și în cazul unui proces judiciar, decizia pe care o luăm după aplicarea unui test statistic poate fi corectă sau greșită.

Decizia este una **corectă** fie atunci când prezumția de nevinovăție H_0 este respinsă în mod corect, iar un inculpat vinovat este condamnat, fie atunci când, în mod corect, prezumția de nevinovăție H_0 nu poate fi respinsă (din lipsă de probe), iar un inculpat nevinovat este eliberat în mod corect.

Decizia este una **greșită** fie atunci când prezumția de nevinovăție H_0 este respinsă în mod incorect, iar un inculpat nevinovat este condamnat din greșeală, fie atunci când, în mod incorect, prezumția de nevinovăție H_0 nu poate fi respinsă, iar un inculpat vinovat este eliberat din greșeală.

Există așadar două tipuri de erori care pot apărea în aplicarea unui test statistic:

- **Eroarea de tipul I**, numită și **eroare de tip α**, apare atunci când respingem din greșeală ipoteza nulă H_0 („prezumția de nevinovăție"), și considerăm din greșeală ipoteza alternativă H_1 ca adevărată („condamnarea din greșeală a unui nevinovat"). Riscul de a comite o eroare de tip I este stabilit prin nivelul de semnificație statistică ales (5%, când α=0,05).

- **Eroarea de tipul II**, numită și **eroare de tip β**, reprezintă nerespingerea ipotezei nule când în realitate ipoteza alternativă (H_1) este adevărată ("inculpatul este vinovat, dar este eliberat din lipsă de dovezi"). Riscul de a comite o eroare de tip II depinde de puterea testului aplicat, determinată la rândul ei de dimensiunea eșantionului studiat, pragul de semnificație, variabilitatea datelor și alte aspecte. Convențional, un risc acceptabil de a comite o eroare de tip II este de cel mult 20%, ceea ce corespunde unei puteri a testului de cel puțin 80% (1-β=1-0,2=0,8).

Pentru facilitarea înțelegerii celor două tipuri de erori, acestea pot fi asimilate unui rezultat fals pozitiv (eroarea de tip I), respectiv fals negativ (eroarea de tip II), conform tabelului 5.1.

Eroarea de tipul I apare în special din cauza unor erori de eșantionare, iar consecințele ei sunt de obicei mai grave decât cele ale erorii de tip II, întrucât poate conduce la acceptarea ca eficiente sau sigure a unor tratamente care în realitate sunt ineficiente sau nesigure.

Tabelul 5.1 Tipuri de erori în aplicarea unui test statistic

		ADEVĂRUL	
		H_0 este falsă	**H_0 este adevărată**
DECIZIA în urma aplicării testului	**H_0 este respinsă (rezultat pozitiv)**	Decizie corectă (real pozitiv)	Eroare de tip I (fals pozitiv)
	H_0 nu poate fi respinsă (rezultat negativ)	Eroare de tip II (fals negativ)	Decizie corectă (real negativ)

5.4 Puterea testului

Puterea unui test statistic reprezintă probabilitatea ca testul să respingă în mod corect ipoteza nulă (H_0), atunci când ipoteza alternativă (H_1) este adevărată.

Cu alte cuvinte, puterea testului măsoară capacitatea acestuia de a evita o eroare de tip II.

Puterea testului se notează cu 1-β, unde β reprezintă riscul de eroare de tip II.

Pentru a pune în evidență **un efect așteptat de o anumită mărime** (o anumită diferență așteptată între grupurile comparate), dimensiunea (talia) eșantionului studiat trebuie să asigure o putere a testului suficient de mare, convențional de cel puțin 0,8 (i.e. 80%), ceea ce corespunde unui risc de a comite o eroare de tip II (β) de cel mult 0,2 (i.e. 20%).

În asigurarea puterii unui test statistic, talia (dimensiunea) eșantionului poate fi privită ca jucând rolul obiectivului unui microscop. Cu cât efectul așteptat (pe care dorim să-l punem în evidență ca fiind o diferență semnificativă statistic între grupurile comparate) este mai mic, cu atât talia eșantionului necesar pentru a putea sesiza acel efect va fi mai mare. Și reciproc, cu cât diferența așteptată între două grupuri este mai mare, cu atât este suficient și un eșantion mai mic pentru ca testul să aibă puterea să sesizeze acea diferență ca fiind semnificativă statistic.

Dacă o anumită diferență nu este sesizată de către un test ca fiind semnificativă statistic, s-ar putea ca această diferență să fie prea mică pentru a putea fi sesizată la puterea oferită de o anumită talie a eșantionului. **Prin mărirea dimensiunii eșantionului studiat, testul va deveni mai puternic** și ar putea eventual sesiza această diferență ca fiind semnificativă. De aceea, ipoteza nulă nu va fi niciodată acceptată atunci când rezultatul unui test nu este semnificativ statistic, ci vom decide doar că, la puterea actuală a testului statistic, nu putem respinge ipoteza nulă.

O analiză **post hoc** (după analiza datelor) a puterii unui test statistic poate fi utilizată pentru a calcula puterea atinsă într-un anumit studiu, însă, pentru a fi utilă în planificarea unui studiu, analiza de putere trebuie condusă întotdeauna **a priori**, adică înaintea culegerii și analizei datelor.

5.5 Calculul dimensiunii eșantionului

Pentru a calcula dimensiunea (talia) minimă a eșantionului, necesară pentru detectarea unui efect (diferență) de o anumită mărime, analiza de putere trebuie realizată **în etapa de planificare a unui studiu**.

Așadar, încă din etapa de planificare a unui studiu este necesar să cunoaștem sau să putem estima **mărimea efectului studiat**, pe baza diferențelor (efectelor) observate în studii anterioare.

În lipsa unor rezultate anterioare publicate în literatură, mărimea efectului așteptat trebuie estimată prin realizarea unui **studiu pilot**, pe un eșantion de cel puțin 30 de subiecți.

Există metode de calcul, fie manual, fie prin programe computerizate de analiză de putere, a dimensiunii minime a eșantionului necesară pentru a atinge o anumită putere dorită (de cel puțin 80%, însă de dorit >90-95%) în detectarea unui efect de o anumită dimensiune estimată.

În calculul dimensiunii eșantionului intervine diferența minimă interesantă a efectului studiat (diferenței), variabilitatea datelor (prin deviația standard), tipul testului (unilateral, bilateral), puterea dorită, raportul dintre numărul de subiecți dintr-un grup față de numărul de subiecți din grupul comparat (1:1 sau 1:2), pragul de semnificație statistică (0,05).

6. Normalitatea datelor

Obiective educaționale. După parcurgerea acestui capitol, studenții:

- Vor înțelege conceptul de normalitate a datelor
- Vor înțelege utilitatea evaluării normalității - condiție de aplicare a unor teste statistice frecvent utilizate
- Vor cunoaște principalele metode de evaluare a normalității, cu ajutorul graficelor, indicatorilor de statistică descriptivă și a testelor statistice de normalitate - și interpretarea acestora

6.1 Utilitatea evaluării normalității datelor

Metodele statistice care utilizează inferența (teste statistice, intervale de încredere, ...), se pot folosi dacă sunt îndeplinite anumite condiții. Întotdeauna înainte de utilizarea unui test statistic sau pentru orice analiză statistică, trebuie căutate care sunt condițiile de aplicare (în engleză: assumptions). O astfel de condiție este normalitatea datelor cantitative. Condițiile trebuie verificate și în caz că sunt satisfăcute, testele se pot utiliza. Dacă nu sunt satisfăcute condițiile de aplicare, se caută alte metode statistice care nu necesită condiții atât de stricte.

În cazul datelor cantitative, una din condițiile de aplicare pentru multe teste statistice este normalitatea datelor. În statistică se spune că datele sunt normal distribuite, sau că urmează o distribuție normală dacă distribuția de probabilitate a unui set de date este asemănătoare cu distribuția normală (gausiană – curba lui Gauss). În caz contrar, spunem că datele nu sunt normal distribuite, sau că nu urmează o distribuție normală. În domeniul medical, dentar, biologic, date care nu au o distribuție normală sunt frecvente. Faptul că datele nu urmează o distribuție normală nu e ceva negativ, ci, pur și simplu, trebuie ținut cont de acest aspect când descriem și analizăm aceste date.

6.2 Evaluarea normalității datelor

Literatura de specielitate descrie diverse metode de a realiza evaluarea normalității datelor, fiecare cu avantaje și dezavantaje. Nu există un consens între statisticieni care e varianta perfectă pentru a evalua normalitatea.

Ceea ce interesează nu este de fapt distribuția datelor din eșantion, ci distribuția datelor din populația din care a fost extras eșantionul. Întrucât nu știm în mod uzual distribuția datelor în populație, frecvent evaluăm normalitatea datelor pentru variabilele din eșantion și presupunem că în populație distribuția e asemănătoare (evident această variantă este supusă riscului de eroare). Niciuna dintre metode nu garantează că vom afla corect dacă distribuția reală în populație e normală sau nu. Din distribuții perfect normale în populație

putem extrage eșantioane a căror distribuție sugerează deviere marcată de la normalitate și viceversa.

Pentru evaluarea normalității se pot folosi metode grafice (cele mai bune), metode utilizând indicatori de statistică descriptivă (mai puțin bune), respectiv teste statistice care evaluează normalitatea (considerate de unii statisticieni puțin folositoare).

6.2.1 Evaluarea grafică

Grafic putem evalua normalitatea cu ajutorul histogramelor, graficelor cutie cu mustăți, graficelor de quantile, dar și prin alte diagrame.

Graficul cutie cu mustăți (în engleză boxplot sau box and wiskers) sugerează normalitatea în cazul în care liniile corespunzătoare cuartilei 3 și 1 sunt aproximativ simetric distanțate față de linia medianei, respectiv lungimile mustăților par similare (figura 6.1 a). Celelalte situații sugerează date care nu au distribuție normală. Astfel, în cazul graficului vertical, dacă mustața superioară (care pleacă de la cuartila 3) este mai mare decât cea inferioară (care pleacă de la cuartila 1), este un indiciu de date care nu au distribuție normală cu asimetrie la dreapta (figura 6.1. b), în caz contrar – mustața inferioară e mai mare decât cea superioară – este sugerată o asimetrie la stânga (figura 6.1.c). Asimetria la dreapta este sugerată și de o distanța între cuartila 3 și mediană mai mare decât distanța între cuartila 1 și mediană (figura 6.1. b). Asimetrie la stânga în mod similar e sugerată și de o distanța între cuartila 3 și mediană mai mică decât distanța între cuartila 1 și mediană (figura 6.1.c).

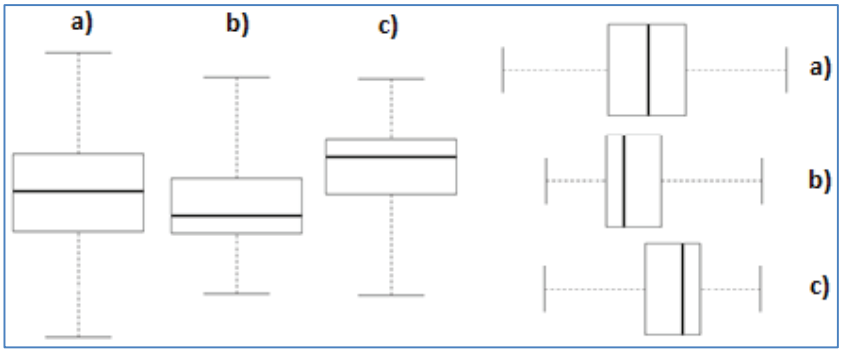


Fig. 6.1 Evaluarea normalității datelor utilizând graficul cutie cu mustăți. a) sugerează date normal distribuite, b) și c) sugerează date care nu urmează o distribuție normală, b) sugerază asimetrie la dreapta, c) sugerează asimetrie la stânga.

Histograma, în caz că pare aproximativ simetrică, cu formă de clopot sugerează normalitatea (figura 6.2.a). În caz contrar, este sugerată lipsa normalității. Astfel, în situația în care există o ”coadă” spre dreapta (valori mai distanțate de centru în partea dreaptă mai mult decât în partea stângă), putem presupune o asimetrie la dreapta (figura 6.2.b). În

situația în care există o ”coadă” spre stânga (valori mai distanțate de centru în partea stângă mai mult decât în partea dreaptă), putem presupune o asimetrie la stânga (figura 6.2.c).

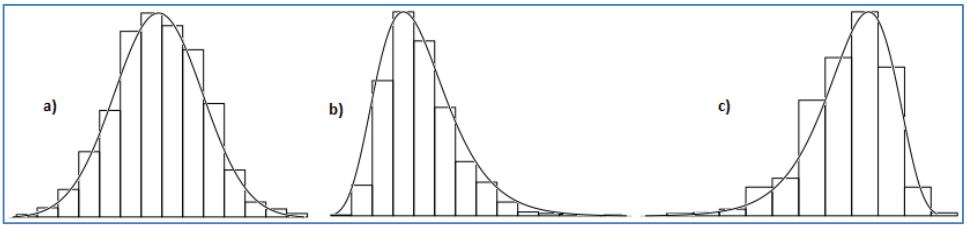


Fig. 6.2 Evaluarea normalității datelor utilizând histograma. a) sugerează date normal distribuite, b) și c) sugerează date care nu urmează o distribuție normală, b) sugerază asimetrie la dreapta, c) sugerează asimetrie la stânga.

Graficul cuantilă-cuantilă, permite cel mai bine evaluarea normalității întrucât observăm toate valorile din seria de date (reprezentate prin puncte sau cercuri). Linia diagonală reprezintă o distribuție normală. Cu cât punctele sunt mai apropiate de această linie, putem presupune o distribuție normală (figura 6.3.a), respectiv o tendință a punctelor de a se îndepărta de linia diagonală sugerează absența unei distribuții normale (figura 6.3.b).

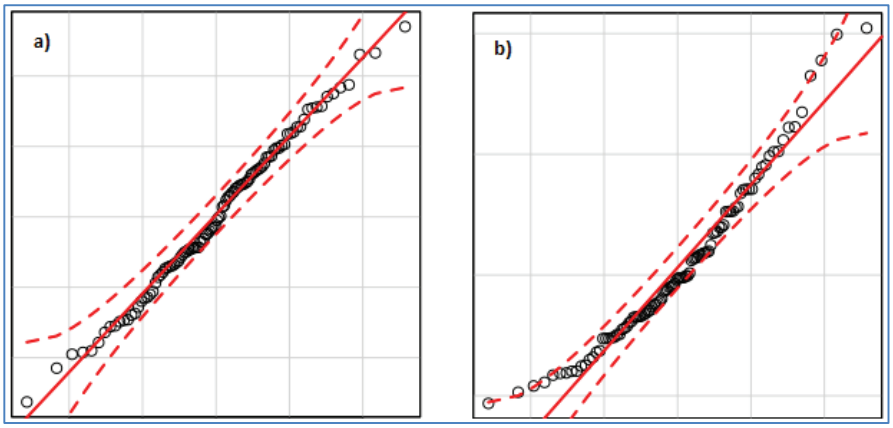


Fig. 6.3 Evaluarea normalității datelor utilizând graficul cuantilă-cuantilă. a) date normal distribuite, b) date care nu urmează o distribuție normală

6.2.2 Evaluarea prin parametrii de statistică descriptivă

Evaluarea normalității se poate face cu **indicatori de statistică descriptivă**, precum media, modulul, mediana, coeficientul de asimetrie (în engleză skewness) și coeficientul de boltire (în engleză kurtosis).

Media, modulul și mediana au valori apropiate, sau chiar identice în cazul distribuției normale. E greu de evaluat ce este apropiat sau nu, întucât depinde de caracteristica

studiată, respectiv există întotdeauna și o doză de subiectivism. O metodă simplă care ne sugerează probleme de normalitate a datelor este compararea mediei cu deviația standard: în general media este mai mare decât deviația standard în eșantioanele normal distribuite.

Coeficientul de asimetrie apropiat de 0 sugerează o distribuție simetrică. Vom considera o valoare în intervalul (-1, 1) ca sugestie de normalitate (desigur această alegere este una empirică, mai degrabă cu scop didactic). O valoare în afara intervalului o vom considera ca absență a unei distribuții normale. Interpretarea valorilor asimetriei se face în conformitate cu cele specificate în subcapitolul 2.6.1.

Excesul de boltire apropiat de 0 sugerează o distribuție având o formă a curbei similară distribuției normale. Vom considera o valoare în intervalul (-1, 1) ca sugestie de normalitate. O valoare în afara intervalului o vom considera ca absență a unei distribuții normale. Valorile pozitive sugerează o distribuție cu vârf mai ascuțit și mai înalt decât al unei distribuții normale, iar valorile negative sugerează o distribuție cu vârf aplatizat și mai mic decât al unei distribuții normale.

Vom considera că o serie de date are distribuție normală, dacă atât coeficientul de asimetrie cât **și** cel de boltire au valori în intervalul (-1, 1). În cazul în care oricare din cei doi coeficienți au valori în afara intervalului, vom considera că seria de date nu are o distribuție normală. Spre exemplificare, într-un studiu pentru variabila vârstă, coeficientul de asimetrie a fost de -0,3, iar coeficientul de boltire 0,6 – vom interpreta că datele par normal distribuite, cu ușoară asimetrie la stânga și cu un vârf mai ascuțit. Dacă în același studiu observăm nivelul insulinei în sânge, cu o asimetrie de 2,3 și o boltire de -0,4, vom interpreta că datele nu par normal distribuite, cu o asimetrie la dreapta și o ușoară aplatizare a vârfului.

6.2.3 Evaluarea prin teste statistice

O altă metodă de a evalua normalitatea este utilizarea **testelor statistice pentru evaluarea normalității**. Există numeroase teste specifice pentru normalitate: Shapiro-Wilk (folosit mai ales pentru eșantioane având sub 50 de subiecți), Kolmogorov Smirnov, D' Agostino-Pearson dar chiar și testul Chi-pătrat poate fi folosit în anumite condiții pentru a o testa.

Testele specifice compară distribuția observată cu una teoretică normal distribuită. Interpretarea pentru toate aceste teste se face similar. Ipoteza nulă spune că distribuția este normală, ipoteza alternativă spune că distribuția nu este normal distribuită (de fapt, corect este să spunem că ipoteza nulă reprezintă situația în care nu există diferență statistic semnificativă între distribuția datelor și distribuția normală, repectiv ipoteza alternativă reprezintă situația în care există diferență statistic semnificativă între distribuția datelor și distribuția normală). Dacă valoarea lui p a unui test pentru evaluarea normalității este sub 0,05, atunci se consideră că datele nu urmează o distribuție normală.

Dacă într-un studiu aplicăm testul de normalitate Shapiro-Wilk pentru variabila vârstă și obținem valoarea p mai mare de 0,05, avem o sugestie de normalitate a acestei variabile.

Folosind același test pentru variabila colesterol total, presupunând că am obținut valoarea lui p mai mică de 0,05, avem o sugestie că datele nu sunt normal distribuite.

Observație

Problema acestor teste, este că atunci când avem cel mai mult nevoie să știm dacă datele sunt sau nu normal distribuite (pentru eșantioane mici), aceste teste nu au putere suficientă să ne ajute cu un răspuns valid. Când avem mai puțină nevoie de aceste teste (pentru eșantioane mari), testele de normalitate sunt prea puternice și pot sesiza minine abateri de la normalitate, care nu deranjează analizele de care avem nevoie.

6.3 Teste parametrice/neparametrice

Cum se preciza mai sus, normalitatea datelor reprezintă o condiție de aplicare pentru numeroase teste statistice – și anume pentru unele teste parametrice.

Un test statistic parametric este un test al cărui aplicare presupune că populația din care a fost extras eșantionul studiat urmează o distribuție de probabilitate care depinde de un set fix de parametri. Spre exemplu, testul Z se bazează pe distribuția normală, care depinde de 2 parametri: medie și deviația standard, testul student se bazează pe distribuția t, care depinde de un parametru: numărul de grade de libertate.

Printre testele statistice cele mai folosite în domeniul medical referitor la comparații între grupuri, pentru date de tip cantitativ, alegerea între teste parametrice și neparametrice echivalente se face în funcție de prezența normalității datelor. Astfel, dacă datele sunt normal distribuite se va merge în principal pe teste parametrice, în caz contrar se vor utiliza teste neparametrice echivalente (Tabel 6.1). Desigur normalitatea datelor nu e unicul criteriu, în capitolele referitoare la teste statistice sunt precizate mai multe detalii privind criteriile de aplicare.

Tabel 6.1. Alegerea între teste parametrice și neparametrice echivalente în funcție de normalitatea datelor, pentru comparații între grupuri, pentru variabile cantitative.

Teste parametrice	Teste neparametrice echivalente
Student (t) pentru eșantioane independente	Mann–Whitney U Mann–Whitney–Wilcoxon Wilcoxon–Mann–Whitney Wilcoxon rank-sum Wilcoxon pentru eșantioane independente
Student (t) pentru eșantioane pereche	Wilcoxon signed-rank Wilcoxon pentru eșantioane dependente
ANOVA	Kruskal-Wallis

6.4 Exerciții propuse

1. Corpul osului femur a 40 de pacienți a fost măsurat. Următoarele statistici au fost calculate: mediana=45 cm, percentila 25=43,2 cm, percentila 75=49,4 cm, minimul=39,1, maximul=56,2. Care din următoarele răspunsuri sunt corecte:
A. Datele nu par să fie normal distribuite
B. Un grafic utilizabil pentru a evalua normalitatea este graficul cutie cu mustăți
C. Datele par să aibă o asimetrie la stânga
D. Un grafic foarte bun pentru a evalua normalitatea este graficul pie (sectorial)
E. Distribuția datelor pare să aibă o coadă spre dreapta

2. Greutatea a 40 de pacienți a fost măsurată. Următoarele statistici au fost calculate: media=78 kg, mediana=77,8 kg, coeficientul de asimetrie=0,18, coeficientul de boltire=0,7, coeficientul de variație=0,08, deviația standard descriptivă=6 kg. Care din următoarele răspunsuri sunt corecte:
A. Datele par să fie normal distribuite
B. Un grafic utilizabil pentru a evalua normalitatea este histograma
C. Datele par să aibă o asimetrie la stânga
D. Datele par să aibă o anumită boltire
E. Corelația este slabă spre absentă

3. Greutatea a 40 de pacienți a fost măsurată. Următoarele statistici au fost calculate: deviația standard descriptivă=6 kg, coeficientul de asimetrie= - 0,7, coeficientul de boltire= - 0,9, testul Shapiro-Wilk a oferit o valoare p de 0,02. Care din următoarele răspunsuri sunt corecte:
A. Datele nu par să fie normal distribuite
B. Un grafic utilizabil pentru a evalua normalitatea este graficul cuantilă-cuantilă
C. Datele par să aibă o asimetrie la stânga
D. Corelația este negativă
E. Distribuția datelor pare aplatizată

7. Compararea variabilelor cantitative

Obiective educaționale:

- Studentul va fi capabil să aleagă testul potrivit pentru compararea a două medii sau mai multe medii.
- Să aplice testele statistice Z, t, Anova pentru compararea mediilor, testele Fisher, Levene, Bartlett pentru compararea variațelor, testele neparametrice Mann-Whitney, Wilcoxon, Kruskall-Wallis, Friedman, testul de mediană.

7.1 Concepte de bază

Exemplul 7.1. (compararea a două medii).

Pacienții cu diabet de tip 2 au unele caracteristici diferite față de pacienți care nu au această afecțiune. În primul scenariu propus au fost comparați 64 de pacienți consecutivi cu diabet, numiţi în continuare grup caz, cu 20 de subiecți fără diabet de tip 2 selectați aleator, numiți în continuare grup control. Pentru descrierea celor două loturi, autorii au înregistrat și analizat diverși parametri precum vârsta, indicele de masă corporală (IMC) sau tensiunea arterială sistolică (TAS) și diastolică (TAD). Scopul studiului a fost de a investiga relația dintre adiponectină (un marker al inflamației) și diverși markeri ai stresului oxidativ. Acest scop se poate atinge după analizarea loturile studiate, loturi care ar fi de dorit să fie asemănătoare din punctul de vedere a factorilor ce ar putea influența adiponectina sau markerii de stres oxidativ. Dacă în final se observă o relație între adiponectină și markerii de stres oxidativ, aceasta s-ar putea datora parametrilor care diferă între cele două grupuri de pacienți. Din această cauză, analiza se începe cu compararea caracteristicilor grupurilor studiate. În studiu este raportată media vârstei la grupul caz 65,61±11,13 ani și media vârstei la grupul control 62,40±12,04 ani, media IMC la grupul caz 31,76±5,82 kg/m2 și, respectiv, la grupul control 25,04±7,33 kg/m2. În ce privește TAS, este raportată mediana și intervalul quartilic la grupul caz 130 (120 - 140) mmHg, respectiv, 120 (116,25 – 133,75) mmHg, analog pentru TAD la grupul caz 80 (70 - 80) mmHg, respectiv, la grupul control 80 (70 – 80) mmHg.

Exemplul 7.2 (compararea a două medii)

În acest scenariu se compară două grupuri de pacienți a câte 21 de copii fiecare, primul format din copii între 3-5 ani cu multiple carii (>5) și respectiv copiii fără nici o carie. Scopul studiului a fost de a determina diferențele dintre cele două grupuri din punctul de vedere al diverșilor factori salivari. Autorii au fost interesați de nivelul pH-ului salivar care a fost raportat la primul grup 6,45±0,5, respectiv la al doilea grup 7,15±0,31.

Exemplul 7.3 (măsurători repetate)

Obiectivul studiului a fost de a evalua adiponectina în dinamică și evoluția bolii cronice renale pe un grup de 56 pacienți timp de un an. Adiponectina a fost testată la momentul inițial și la finalul studiunlui iar pentru evaluarea afectării renale, unul dintre parametrii testați a fost raportul albumină creatinină din urină.

S-au avut în vedere următoarele întrebări:

Întrebare 1: Au existat diferențe între adiponectina initială 8,61±1,27 μg/ml și finală 10,87±1,26 μg/ml?

Întrebare 2: Au existat diferențe între raportul de albumină creatinină din urină (UACR) inițial 81,58±26,42 mg/g și final 69,07±20,90 mg/g?

Exemplul 7.4 (testări repetate multiple)

S-a dorit studierea combinată a nivelului interleukinelor (IL)-1β, IL-2, IL-4, IL-5, etc la pacienții cu septicemie, pentru observarea dinamicii acestora. Au fost introduși în studiu 30 de pacienți cu septicemie. Au fost prelevate 7 probe de sânge de la fiecare pacient, la interval de o zi. Din aceste probe au fost determinate interleukinele respective și studiată dinamica lor. Valorile obținute (mediana per lot) pentru IL-1β au fost M1=2,94; M2=2,93; M3=1,96; M4=1,64; M5=1,65; M6=1,76; M7=1,35.

Exemplul 7.5 (mai multe medii).

Au fost studiate diverse biomateriale compozite folosite la restaurarea dentară în vederea comparării durabilității. Pentru aeasta s-a studiat in vitro mărimea interfeței de legătură între material și dentină la 6 grupuri a câte 20 de dinți (dinții au fost randomizați în cele 6 grupuri). În acest scenariu pentru materialele Gradia (mărimea medie ale interfeței de legătură între material și dentină: 2190,85±2124,73 μm), Barodent (2265,95±1943,15 μm), Excite DSC (563,10±883,19 μm), Optibond – FL (480,20±843,53 μm) vor fi comparate mărimile medii per grup ale interfeței de legătură între material și dentină.

Exemplul 7.6 (mai multe medii).

S-a dorit evaluarea legăturii dintre nivelurile unor metale serice: Fe, Cu, Zn, Cd și Pb și existența anemiei. Au fost studiați 256 de copii selectați aleator dintr-un centru industrial din Turcia. Copiii au fost împărțiți în 4 grupuri: 23 de copii cu anemie și deficiență de fier (IDA), 36 de copii cu deficiență de fier fără anemie (ID), 18 copii numai cu anemie (OA) și 179 de copii indemni (control). Valorile hemoglobinei (HB) au fost raportate în studiu pentru descrierea loturilor: grup IDA 9,91±0.31 g/dl, grup ID 12,64±0.15 g/dl, grup OA 10,45±0.23 g/dl, grup control 13,30±0.10 g/dl.

7.2 Introducere. Noțiunea de comparare.

Acest capitol va descrie metode ce aparțin inferenței statistice. Mai precis, va răspunde la întrebări de genul „Există diferență între IMC-ul diabeticilor de tip 2 și a persoanelor sănătoase?" (exemplul 7.1), „Există diferență între nivelul pH-ului salivar la copiii care prezintă multe carii față de cei care nu au carii?" (exemplul 7.2). Întrebările din exemplul 7.3 sunt diferite față de cele din primele două exemple din acest capitol. În exercițiul 7.3, diferența se raportează nu între două grupuri diferite de pacienți, ci în cadrul aceluiași grup, comparând caracteristica studiată între două momente de timp. Pacienții la momentul inițial reprezintă un grup care va fi comparat cu pacienții la momentul final considerat al doilea grup. Din punct de vedere statistic avem două grupuri, deși sunt formate din aceeași indivizi. Dacă indivizii din primele două exemple sunt diferiți, indivizii aparținând grupurilor din exemplul 7.3 sunt aceeași, grupul al doilea (cel testat final) depinde de grupul testat inițial. Acest tip de grupuri se vor numi grupuri dependente sau grupuri perechi. Grupurile din exemplele 7.1 sau 7.2 se vor numi grupuri independente.

Variabilele testate în exemplul 7.7, fumatul (Da/Nu) şi amputaţia (Da/Nu) sunt variabile calitative. Pentru a studia asocierea lor se vor compara tabele de frecvenţe cu metode prezentate în capitolul următor. Acestea nu pot fi comparate cu testele statistice prezentate în acest capitol, teste folosite pentru a compara medii sau mediane.

Dacă dorim să comparăm o caracteristică cantitativă, între două grupuri de subiecți, ca în exercițiul 7.1, un grup de pacienți cu diabet de tip 2 (n_1=64) și un altul de control (n_2=20). Caracteristica urmărită (de ex. IMC vezi exercițiul 7.1) va fi măsurată pe fiecare individ în parte. Vor rezulta astfel 64 de valori ale IMC ($X_1(i), i = \overline{1,64}$), câte una pentru fiecare pacient din grupul cu diabet și 20 de valori ale IMC ($X_2(j), j = \overline{1,20}$), câte una pentru fiecare subiect din grupul control. Vom calcula media și deviația standard pentru fiecare grup în parte. Acestea vor fi notate cu $\bar{X}_1, s_1,$ respectiv $\bar{X}_2, s_2$.

De remarcat că vom porni de la întrebarea: Sunt cele două grupuri de subiecți asemănătoare din punctul de vedere al IMC-ului? Dacă rezultatul va fi unul negativ, atunci vom trage concluzia că cele două grupuri aparțin unor populații diferite, P_1 populația cu diabet de tip 2, respectiv P_2 populația indemnă de această boală (Dacă extragem două eșantioane din aceeași populație ne așteptăm să aibă medii apropiate). Sau, cu alte cuvinte, din punctul de vedere al IMC-ului, populația diabeticilor de tip 2 este diferită statistic de populația indemnă de boală. Cu cât diferențele dintre medii sunt mai mari, cu atât vom putea să afirmăm mai sigur că populațiile diferă statistic din punct de vedere al IMC-ului.

Încă din acest moment este important de reținut:

- Diferențele aritmetice observate nu sunt întotdeauna și diferențe statistic validate.
- Diferențele chiar validate statistic trebuie să aibă și semnificație clinică (de exemplu scăderea în medie a colesterolului cu 2% în urma unui tratament pe o perioadă de 5 ani chiar dacă este semnificativă statistic nu are și relevanță pentru pacienți).

Dacă rezultatul va fi unul pozitiv şi eşantioanele sunt suficient de mari, atunci vom trage concluzia că cele două grupuri aparţin aceleiaşi populaţii, deci că populaţia P_1 cu diabet de tip 2, nu diferă de populaţia P_2, indemnă de această boală. Sau, cu alte cuvinte, din punctul de vedere al IMC-ului, populaţia diabeticilor de tip 2 nu este diferită de populaţia indemnă de boală sau diferenţele dintre mediile observate se datorează întâmplării. Dar mai există o posibilitate: aceea de a selecta grupurile eronat, astfel încât să nu reprezinte populaţia din care le-am extras , fie acestea să fie prea mici, fie să nu prezinte aceleaşi caracteristici ca şi populaţia ţintă. Atunci diferenţele dintre medii s-ar putea datora selecţiei inadecvate. De aceea se face selecţie aleatoare şi se calculează volumul eşantioanelor, diminuând în acest mod posibilitatea de eroare.

Aşadar, dacă dorim să comparăm o variabilă cantitativă, vom selecta aleator un număr de n_1, respectiv n_2 subiecţi din populaţiile P_1 şi P_2, având mediile (necunoscute μ_1, respectiv μ_2), pentru care se vor determina mediile $\bar{X}_1$şi $\bar{X}_2$; varianţele S_1^2 şi S_2^2.

În exemplul 7.2 comparaţia se face între grupul de copii cu multiple carii (aparţinând populaţiei P_1) și grupul de copii care nu au carii (populaţia P_2). Caracteristica studiată este nivelul pH-ului salivar. Se măsoară nivelul pH-ului la fiecare copil intrat în studiu. Rezultă astfel seria de 21 de valori ale pH-ului ($X_1(i), i = 1,2, \dots ,21$), câte una pentru fiecare copil din grupul cu carii multiple și seria de 21 de valori ale pH-ului ($X_2(j), j = 1,2, \dots ,21$), câte una pentru fiecare subiect din grupul fără carii. Vom calcula media și deviaţia standard pentru fiecare grup în parte. Acestea vor fi notate cu $\bar{X}_1, S_1,$ respectiv $\bar{X}_2, S_2$.

Pentru comparare se poate alege metoda intervalelor de încredere sau metoda testelor statistice.

În cazul exemplului 7.1 pentru grupul caz, populaţia ţintă a fost cea de diabetici de tip 2. Populaţia accesibilă a fost cea venită la control în clinică. Pentru ca eșantionul extras să fie reprezentativ pentru populaţie, acesta trebuie extras aleator din populaţie, astfel încât orice individ să aibă șanse egale de a fi selectat. Cum vom proceda? Ar trebui să cunoaștem toată populaţia cu diabet de tip 2 arondată la o clinică (populaţia accesibilă) și din aceasta, selecţia se va face aleator, selectând numărul de fișă. Calculul numărului de indivizi necesari (mărimea eșantionului) pentru a detecta mărimea așteptată a caracteristicii s-a discutat în capitolele anterioare.

În acest capitol vom prezenta cele mai utilizate teste statistice pentru compararea grupurilor în cazul unor caracteristici cantitative: testele t, Anova și cerinţele lor. Ele vor fi completate de testele neparametrice, teste de mediane, Mann-Whitney, Wilcoxon, Kruskal-Wallis, Friedman, utilizate când nu sunt îndeplinite cerinţele testelor parametrice. Vom prezenta și testele Z, utilizate în mai mică măsură în studiile medicale. Vom prezenta și testele de verificare a cerinţelor: pentru testarea varianţelor sau omogenităţii: Fisher, Levene și Bartlett.

Ordinea logică ne cere să discutăm pentru început unele aspecte legate de luarea deciziei de existenţă sau non existenţă a diferenţei între două medii.

Să luăm drept model exemplul 7.1: „Există diferenţă între vârsta diabeticilor de tip 2 și a persoanelor din grupul de control?”

Presupunem că medile vâstelor subiecţilor din grupul cu diabet, respectiv control sunt $\bar{X}_1 = 65$ ani, $\bar{X}_2 = 35$ ani, vezi figura 7.1. Aşadar, în acest prim exemplu diferenţele dintre medii sunt mari, concluzia este aparent clară: există o diferenţă aritmetică între cele două medii, diferenţă care trebuie validată statistic.

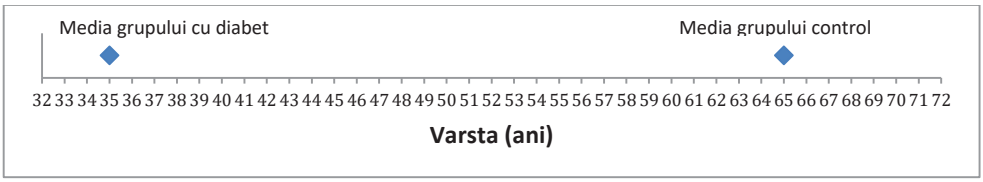


Fig. 7.1. Media aritmetică a vârstei grupului pacienţilor cu diabet de tip 2 şi a grupului control.

Dacă $\bar{X}_1 = 65$ ani, $\bar{X}_2 = 63$ ani, vezi figura 7.2. Diferenţele dintre medii sunt mici, concluzia: faptul că nu există diferenţă între cele două medii pare evident, dar aceasta trebuie validată statistic.

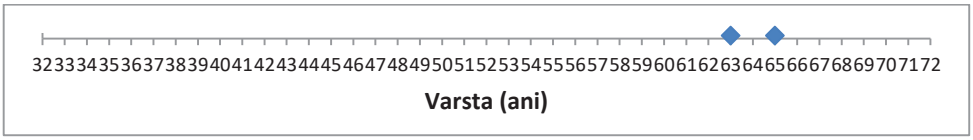
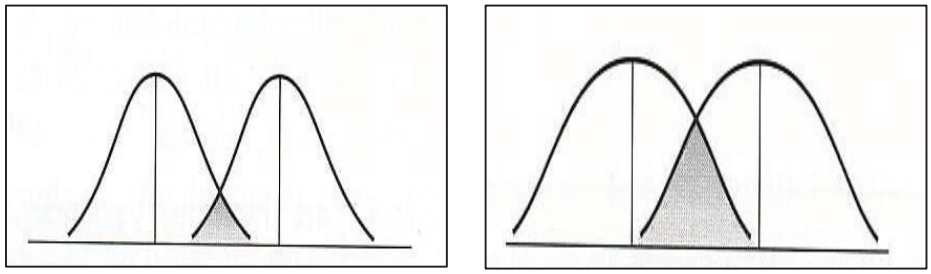


Fig. 7.2. Media aritmetică a vârstei grupului pacienţilor cu diabet de tip 2 şi a grupului control.

Validarea statistică ne va spune dacă rezultatul observat este sau nu datorat întâmplării. Diferenţele mai mici sau mai mari validate statistic pot fi interpretate clinic. Cele fără semnificaţie statistică sunt rezultatul întâmplării şi nu pot fi folosite clinic.

Ce putem spune din punct de vedere statistic în cazurile în care diferenţa nu este clară? Ne punem întrebarea: ce diferenţă ar trebui să existe ca să putem spune că aceasta nu se datorează întâmplării?

Ca să ne dăm seama, trebuie să ne uităm la datele observate, mai precis la distribuţia acestora. În distribuţiile din figura 7.3 a şi b mediile sunt aceleaşi. Însă diferenţe clare între medii există numai în figura 7.3.a, unde deviaţia standard (datele variază mai puţin) este mai mică.



a. b.

Fig. 7.3. Compararea a două distribuţii a. Deviaţiile standard mici; b. Deviaţiile standard mari

Dacă selecția este aleatoare, atunci variația datelor din eșantion oglindește variația din populație. O variație mare în date (indicată de o varianță sau o deviație standard mare) înseamnă indivizi foarte diferiți, neomogeni. În aceste cazuri avem nevoie de diferențe mai mari între medii ca să putem spune că există diferențe semnificative statistic.

Pe de altă parte, în luarea deciziei ne influențează și numărul de indivizi din eșantion. Eșantioanele mari ne asigură reprezentativitate pentru populație. Cu cât eșantionul este mai mare, cu atât estimarea mediei populației este mai precisă.

În concluzie, răspunsul la întrebarea: „Există diferență între media caracteristicii studiate în grupul A față de media grupului B?" depinde de:

- diferența dintre medii,
- mărimea deviației standard,
- mărimea eșantionului,
- tipul testului (unidirecțional, bidirecțional),
- pragul de semnificație ales.

Pentru compararea a două medii există mai multe teste statistice. Parametrii testelor respective vor ține cont de cele de mai sus, în general diferența dintre medii se va împărți la diferența dintre erorile standard (eroarea standard se calculează cu ajutorul deviației standard și a numărului de indivizi). Fiecare test statistic se poate aplica numai în anumite condiții. Aplicarea unui test fără verificarea întrunirii condițiilor este una dintre sursele de erori în cercetarea medicală. În cele ce urmează vom prezenta pe rând testele statistice cu cerințele lor și vom exemplifica modul de aplicare al acestor teste.

7.3 Teste parametrice

7.3.1 Testul Z pentru medii în cazul eșantioanelor independente

Condiţiile de aplicare ale testului Z:

- Eșantioanele sunt independente, populațiile din care se extrag eșantioanele sunt distribuite normal și au deviațiile standard σ_1 și σ_2 cunoscute sau
- Eșantioanele sunt mari ($n_1 \geq 30$ și $n_2 \geq 30$) și independente.

Datorită obligativității distribuției normale, înainte de aplicarea testului se face verificarea distribuției normale pe fiecare eșantion în parte (distribuţiile pot să difere între eşantioane, deoarece este posibil ca să aparţină la două populaţii diferite).

Exemplu 7.7.

Se dorește să se studieze în ce măsură consumul de zahăr este diferit la copiii cu un număr mai mare de carii față de cei care nu prezintă această afecțiune dentară. Se înrolează în studiu prin selecție aleatoare 10 copii care au un număr ridicat de carii și 10 copii care nu au carii. Se urmărește consumul de zahăr timp de o săptămână. Media consumului de zahăr la copiii cu carii a fost $\bar{X}_1 = 114$ g şi media consumului de zahăr la copiii fără carii a fost $\bar{X}_2 = 108$ g, deviațiile standard din populație sunt cunoscute σ_1=10 şi σ_2=8. Datele sunt distribuite normal pe ambele eşantioane.

Se formulează:

Ipoteza nulă H_0: „media consumului de zahăr a copiilor cu un număr ridicat de carii și media consumului de zahăr a copiilor fără carii nu diferă statistic semnificativ" $(\mu_1 - \mu_2 = 0)$, unde μ_1, μ_2 sunt mediile consumului de zahăr în cele două populații, copiii cu număr ridicat de carii, respectiv copiii fără carii.

Ipoteza alternativă H_1 în cazul testului bilateral: „media consumului de zahăr a copiilor cu un număr ridicat de carii și media consumului de zahăr a copiilor fără carii diferă statistic semnificativ" $(\mu_1 - \mu_2 \neq 0)$.

Ipoteza alternativă H_1 în cazul testului unilateral: „media consumului de zahăr a copiilor cu un număr ridicat de carii este statistic semnificativ mai mare decât media consumului de zahăr a copiilor fără carii" $(\mu_1 > \mu_2)$ sau „media consumului de zahăr a copiilor cu un număr ridicat de carii este statistic semnificativ mai mică decât media consumului de zahăr a copiilor fără carii" $(\mu_1 < \mu_2)$.

Parametrul testului Z este:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} - \frac{\sigma_2^2}{n_2}}}.$$

Alegem pragul de semnificație α=0,05.

Regiunea critică în cazul distribuției Z este: $(-\infty; -Z_\alpha) \cup (Z_\alpha; +\infty)$. Pentru nivelul de semnificație ales Z_α=1,96.

Rezolvare: Calculăm parametrul testului presupunând că H_0 este adevărată $(\mu_1 - \mu_2 = 0)$:

$$Z = \frac{114 - 108 - 0}{\sqrt{\frac{10^2}{10} - \frac{8^2}{10}}} = \frac{6}{\sqrt{10 - 6{,}4}} = \frac{6}{1{,}90} = 3{,}2$$

Deoarece $Z \in (-\infty; -1{,}96) \cup (1{,}96; +\infty)$, putem respingem ipoteza nulă și să luăm în considerare ipoteza alternativă H_1.

Pentru Z=3,2 probabilitatea ca ipoteza nulă să fie adevărată este foarte redusă, p=0,0006. Acestă probabilitate este mai mică decât pragul de semnificație ales α=0,05. Din această cauză putem respinge ipoteza nulă și lua în considerare ipoteza alternativă. În acest exemplu, concluzia testului Z bilateral va fi că media consumului de zahăr a copiilor cu un număr ridicat de carii și media consumului de zahăr a copiilor fără carii diferă semnificativ statistic, adică numărul mare de carii la copii s-ar putea datora consumului mare de zahăr. Dacă repetăm studiul în 95% dintre posibilităţi vom avea acelaşi rezultat: media consumului de zahăr a copiilor cu un număr ridicat de carii vva fi mai mare decât media consumului de zahăr a copiilor fără carii.

Testul Z nu este foarte utilizat în literatura de specialitate, din cauză că varianțele populațiilor comparate trebuie să fie cunoscute, cerință greu de îndeplinit.

7.3.2 Testul t pentru eșantioane independente

În locul testului Z se folosește testul t, care are o formulă asemănătoare. Testul t poate fi folosit în dacă eșantioanele au observații independente și valorile sunt normal distribuite. Testul t poate fi folosit și la comparații de eșantiona mai mici de 30 de observații iar pentru eșantioanele mari valoarea parametrului său tinde spre Z.

Datorită necesității existenței distribuției normale înainte de aplicarea testului se face verificarea distribuției normale pe fiecare eșantion în parte.

Deoarece din punct din evedere matematic se folosesc formulte diferite în funcție de egalitatea sau inegalitatea varianțelor între eșantioanele comparate înainte de aplicarea testului t este nevoie să testăm varianțele. Acest lucru se realizează cu un test de varianțe (capitolul 7.4 din acest curs va trata testele pentru varianțe). În funcție de rezultat, se optează pentru una din cele două variante ale testului t: pentru varianțe egale sau pentru varianțe inegale.

Testul t pentru eșantioane independente în cazul varianțelor egale

Formula parametrului t se calculează folosind formula parametrului Z în care deviațiile standard sunt egale și, deviația standard σ care nu este cunoscută se aproximează cu s, adică combinând deviațiile standard observate pe cele două eșantioane. Deviația standard s poate prezenta erori de măsurare, datorate selecției. Totuși, cu cât eșantioanele sunt mai mari, cu atât este mai probabil ca deviația standard *s* să aproximeze mai bine deviația standard din populație. Eroarea de aproximare depinde așadar de volumul eșantioanelor n_1, respectiv n_2, care se vor numi grade de libertate ale testului t. Gradele de libertate se notează cu df și se calculează cu formula $df=n_1-n_2-2$.

Cu cât gradele de libertate sunt mai mari, cu atât distribuția parametrului testului t se apropie de distribuția normală. De reținut: nu vorbim despre o singură distribuție t, ci de mai

multe: pentru fiecare df avem o distribuție t foarte asemănătoare distribuției normale când df se apropie și depășește 30 (figura 7.4).

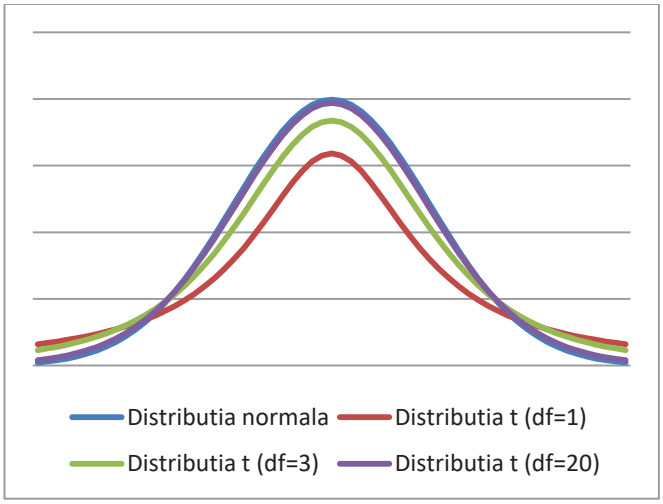


Fig. 7.4. Distribuția normală versus diverse distributii t

Pașii testului sunt aceeași ca și în cazul testului Z: se formulează ipotezele (identice cu cele prezentate în cazul testului Z), se alege pragul de semnificație α=0,05, regiunea critică depinde de gradele de libertate și este: $(-\infty; -t_\alpha) \cup (t_\alpha; +\infty)$, se calculează parametrul testului, se trage concluzia.

În cazul testului t, folosirea unei aplicații pentru prelucrarea statistică a datelor este justificată.

Exemplul 7.8. (test t bilateral pentru eșantioane independente în cazul variațelor egale)

Se dorește urmărirea variațiilor de greutate în rândul adolescentelor. Pentru aceasta, se aleg două grupuri de adolescente: un grup la care se va aplica o intervenție timp de 6 luni și un grup control, unde nu se va aplica nici o intervenție (grup martor). Pentru prevenția obezității la adolescente, intervenția utilizată constă într-o aplicație pe telefonul mobil, care adresează utilizatorului o întrebare/zi despre stilul de viață sănătos, întrebare cu răspunsuri multiple și afișeaza un clasament al cunoștințelor membrilor grupului de intervenție. După 6 luni se măsoară diferența de greutate față de momentul inițial și se compară cu cea a grupului control.

Începem aplicarea testului statistic prin formularea ipotezelor:

Ipoteza nulă H_0: Media diferenței greutății după 6 luni la adolescentele din grupul intervenție ***nu diferă*** statistic semnificativ față de media diferenței greutății după 6 luni la adolescentele din grupul control.

Ipoteza alternativă H_1: Media diferenței greutății după 6 luni la adolescentele din grupul intervenție ***diferă*** statistic semnificativ față de media diferenței greutății după 6 luni la adolescentele in grupul control.

Rezultatele aplicării testului t în Excel sunt date în tabelul 7.1.

Fie că judecăm după parametrul testului, fie că judecăm după probabilitatea p ca ipoteza nulă să fie adevărată, concluzia va fi aceeași: se respinge ipoteza nulă și se validează ipoteza alternativă: Media diferenței greutății după 6 luni la adolescentele din grupul de intervenție diferă semnificativ față de media diferenței greutății după 6 luni la adolescentele din grupul martor, sau, cu alte cuvinte: intervenția aplicată a avut succes în prevenția obezității la adolescente în intervalul testat de 6 luni (test bilateral).

Observație. Deviațiile standard nu sunt egale, dar în urma aplicării testului F a rezultat că ele nu diferă semnificativ, ceea ce ne-a plasat în cazul 1. În prealabil a fost verificată și normalitatea datelor.

Tabel 7.1. Rezultatele aplicării testului t pentru exemplul 7.8.

	Grup interventie	*Grup martor*
Medie	2,95	0,63
Varianță	2,55	2,21
Număr de observații	29	29
df	56	
t	5,72	
P(T<=t) unilateral	0,0000002	
t critic unilateral	1,67	
P(T<=t) bilateral	0,0000004	
t critic bilateral	2,00	

Testul t pentru eșantioane independente în cazul variațelor inegale

Dacă variațele celor două grupuri nu sunt egale atunci există posibilitatea utilizării testului t în cazul în care variațele nu sunt egale. Testul respectiv presupune o estimare a deviațiilor standard σ_1, σ_2 cu ajutorul S_1 si S_2.

Exemplul 7.9. (test t unilateral pentru eșantioane independente în cazul variațelor inegale)

Se dorește compararea IMC-ului unui grup de diabetici de tip 2 (T2DM) cu IMC-ul unui grup de indivizi care nu au această afecțiune (vezi exemplul 7.1). Ambele grupuri au fost selectate aleator. Datele fictive au fost astfel simulate ca variațele să difere semnificativ (testul F pentru variațe).

Începem aplicarea testului statistic prin formularea ipotezelor:

Ipoteza nulă H_0: Media IMC a pacienților din grupul cu T2DM ***nu diferă*** semnificativ de media IMC a indivizilor din grupul control.

Ipoteza alternativă H_1: Media IMC a pacienților din grupul cu T2DM este ***statistic semnificativ mai mare*** decât media IMC a indivizilor din grupul control.

Rezultatele aplicării testului t în Excel sunt date în tabelul 7.2.

Concluzia: Deoarece p<0,05 se respinge ipoteza nulă și putem concluziona că media IMC a pacienților din grupul cu T2DM este statistic statistic semnificativ mai mare decât media IMC a indivizilor din grupul control (test unilateral).

Observație. Deviațiile standard nu sunt egale, în urma aplicării testului F a rezultat că ele diferă semnificativ statistic, ceea ce ne-a plasat în cazul 2. A fost verificată și normalitatea distribuției datelor.

Tabel 7.2. Rezultatele aplicării testului t pentru exemplul 7.9.

	Grup T2DM	*Grup control*
Medie	34,51	24,19
Varianță	60,30	0,65
Număr de observații	20	15
df	20	
t	5,90	
P(T<=t) unilateral	0,000004	
t critic unilateral	1,72	
P(T<=t) bilateral	0,000009	
t critic bilateral	2,09	

7.4 Testarea variațelor

Dacă dorim să utilizăm testul t în cazul eșantioanelor independente trebuie să testăm egalitatea variațelor. Pentru testarea variațelor există mai multe teste care se pot utiliza: testul F (Fisher), testul Levene, testul Bartlett sau testul Brown-Forsythe. Vom exemplifica mai jos testul F.

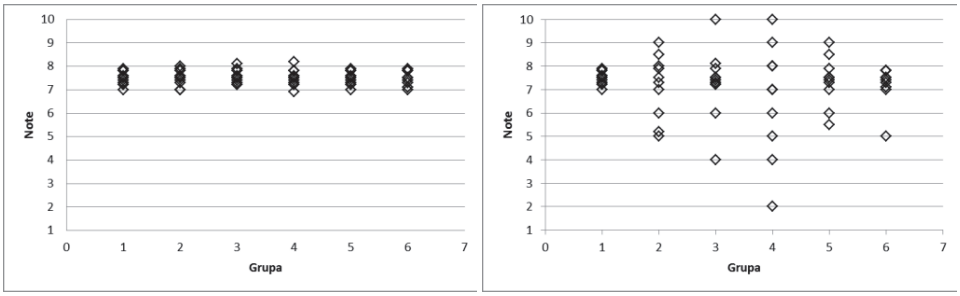
Testul F se poate utiliza numai în cazul eșantioanelor independente (ca în exemplul 7.1) dacă datele celor două eșantioane urmează distribuția normală.

Exemplu 7.10

La un concurs studențesc participă studenți din fiecare dintre cele 6 grupe ale aceleiași serii. Notele obținute de către studenți la acest concurs sunt prezentate în figura 7.5. (în figura 7.5.a. notele obținute sunt omogene, iar în figura 7.5.b. notele sunt eterogene).

De ce este important să putem testa variațele? Omogenitatea nu înseamnă dispersie mică a datelor, nici dispersie asemănătoare. Omogenitatea înseamnă că datele sunt grupate

în jurul mediei (multe în jurul mediei, puține în rest). Dacă se face un studiu asupra unui tratament generic, acesta ar trebui să fie echivalent ca efect cu medicamentul inițial după care a fost copiat. Media singură nu poate să demonstreze echivalența. Aceste tratamente, ca să fie echivalente ar fi necesar ca, atât media, cât și varianța să fie egale.



a. b.

Figura 7.5. Notele studentilor pe grupe. a. Note omogene pe grupe b. Note eterogene pe grupe.

Dacă luăm exemplul 7.8 de mai sus și aplicăm într-un program statistic testul F vom obține rezultatele din tabelul 7.3. în acest caz al testării egalității variațelor ipotezele vor fi formulate astfel:

Ipoteza nulă H_0: **Varianța** scăderii în greutate în 6 luni a adolescenților din grupul intervenție ***nu diferă*** semnificativ statistic față de **varianța** scăderii în greutate în 6 luni a adolescenților din grupul control.

Ipoteza alternativă H_1: Varianța scăderii în greutate în 6 luni a adolescenților din grupul intervenție ***diferă*** semnificativ statistic față de **varianța** scăderii în greutate în 6 luni a adolescenților din grupul control.

Probabilitatea p ca ipoteza nulă să fie adevărată este > 0,05, deci experimentul arată că nu avem suficiente evidenţe pentru a respinge ipoteza nulă: varianța scăderii în greutate în 6 luni a adolescenților din grupul intervenție nu diferă semnificativ de varianța scăderii în greutate în 6 luni a adolescenților din grupul control, adică putem considera cele două variațe egale.

Tabel 7.3. Rezultatele aplicării testului F pentru exemplul 7.8.

	Grup interventie	*Grup control*
Medie	2,95	0,63
Varianță	2,55	2,21
Număr de observații	29	29
Df	28	28
F	1,15	
P(F<=f) unilateral	0,35	
F critic unilateral	1,88	

Observație. Distribuția F urmează o distribuție hi-pătrat cu toate valorile pozitive, nesimetrică și care tinde la 0 spre +∞ (figura 7.6).

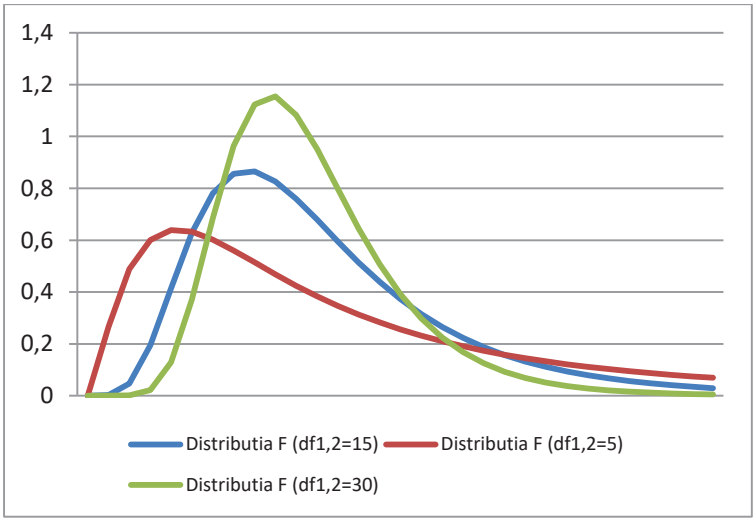


Fig.7.6. Distribuția F pentru diverse grade de libertate

Pentru cel de-al treilea exemplu rezultatele aplicării testului F se găsesc în tabelul 7.4. Acestea au fost discutate în paragraful anterior.

Tabel 7.4. Rezultatele aplicării testului F pentru exemplul 7.9.

	Grup T2DM	*Grup control*
Medie	34,51	24,19
Variață	60,30	0,65
Număr de observații	20	15
Df	19	14
F	92,64	
P(F<=f) unilateral	0,00000000002	
F critic unilateral	2,40	

7.5 Compararea a două medii în cazul eșantioanelor dependente

După cum am arătat în exemplul 7.3 eșăntioanele testate nu sunt întotdeauna independente și în acest caz compararea mediilor provenite din date dependente necesită un alt tip de test statistic.

În trecut, cercetarea medicală a folosit pentru eșantioane dependente gemenii monozigoți dar constituirea unui lot suficient de mare este mult prea dificilă și astfel de studii nu mai apar în literatura științifică modernă.

De multe ori putem recunoaște studiile cu eșantioane dependente deoarece acestea urmăresc evoluția unor parametri pentru un singur lot de pacienți pentru un interval de timp; de exemplu alterarea unor valori biologice ca o consecința a fumatului, a bolilor cardiovasculare sau a aplicării unui tratament. Astfel, apar studiile în care testăm același pacient (pacienți) de mai multe ori, la intervale de timp. Acestea sunt studii cu date dependente, perechi (momentul inițial și final). Se realizează măsurători inițiale (de ex. adiponectina în exemplul 7.3) pentru toți pacienții. La finalul experimentului se efectuează din nou măsurători. Valoarea măsurată inițial are un corespondent în valoarea măsurată final pe același pacient. Aceste date se numesc date perechi.

Un alt model de studiu cu date dependente este cel în care se alege grupul control, astfel încât să fie asemănător cu grupul de cazuri, mai precis, fiecare pacient din grupul caz are un corespondent în grupul control. Corespondența se face pe baza unor caracteristici care sunt identice, cum ar fi genul, vârsta, situația socială, nivelul de educație, mediul de proveniență, etc.

Un alt model de studiu cu date dependente este cel în care același pacient este consultat de două ori de către medici diferiți. Valoarea măsurată de primul observator (medic) are corespondent în valoarea măsurată de al doilea observator (vorbim de același pacient).

Caracteristici ale eșantioanelor dependente (perechi):

- Cele două grupuri au același număr de indivizi.
- Fiecare valoare din primul grup are o valoare corespondentă în al doilea grup.

Dacă avem aceste condiții întrunite, va trebui să alegem pentru prelucrarea datelor teste statistice care compară parametrii proveniți din eșantioane dependente. Aceste teste vor fi descrise în continuare.

7.5.1 Testul t pentru eșantioane perechi

Exemplu 7.11.

Există supoziția că utilizarea algocalminului contribuie la apariția agranulocitozei. Pentru verificarea acestei supoziții într-un studiu se extrage din populație un eșantion de indivizi tratați cu algocalmin în doză crescută,i se dozează numărul neutrofilelor la internarea pacientului și la externarea pacientului. Din punct de vedere statistic întrebarea este: există diferențe semnificative între neutrofilele inițiale și cele finale la pacienții urmăriți? Datele pentru 5 pacienți sunt prezentate în Tabelul 7.5.

Media aritmetică a neutrofilelor la internare a fost 3,6. La externare a fost 2,90. Cele două deviații standard au fost egale 1,85. Dacă ar fi să comparăm mediile, am pleca de la ipoteza nulă presupunând că este adevărată: media neutrofilelor la internare nu diferă semnificativ statistic de media neutrofilelor la externare. Această afirmație este totuna cu: Diferența dintre media la internare și la externare nu diferă semnificativ statistic de 0. Dacă facem media diferențelor între valorile la internare și externare pentru fiecare pacient (vezi Tabel 7.5 coloana 3) ne dă aceeași valoare ca și diferența mediilor.

Testul t pentru eșantioane perechi compară media diferențelor cu 0. Parametrul testului folosește ca estimare a varianței diferențelor din populație varianța diferențelor de pe eșantion (în acest caz 1,30).

Tabel 7.5. Neutrofilele la internare și externare în cazul pacienților tratați cu algocalmin

	Neutrofile (x1000/mm^3)		
	Internare	Externare	Diferență (Internare-Externare)
	1	2	3
Pacient 1	3	0,5	2,5
Pacient 2	1,5	2	-0,5
Pacient 3	5	3,5	1,5
Pacient 4	2,5	3	-0,5
Pacient 5	6	5,5	0,5
Media aritmetică	3,60	2,90	0,70
Deviația standard	1,85	1,85	1,30

Pentru aplicarea corectă a testuuil t pentru eșantioane perechi diferența valorilor pereche (rezultate de pe două eșantioane dependente) trebuie să fie normal distribuităiar perechile să fie independente între ele.

Ipoteza nulă H_0: Media diferențelor valorilor pereche nu diferă semnificativ de 0.

Ipoteza alternativă H_1 (test bilateral): Media diferențelor valorilor pereche diferă semnificativ de 0.

Ipoteza alternativă H_1 (test unilateral): Media diferențelor valorilor pereche este semnificativ statistic mai mare decât 0.

Ipoteza alternativă H_1 (test unilateral): Media diferențelor valorilor pereche este semnificativ statistic mai mică decât 0.

Rezultatele testului t pentru exemplul de mai sus este dat în tabelul 7.6.

Concluzia: Deoarece $p>0,05$ nu am reușit să respingem ipoteza nulă: media diferențelor valorilor pereche nu diferă semnificativ statistic de 0.

Observație. Verificarea normalității trebuie făcută pe valorile rezultate prin diferența dintre internare și externare (coloana intitulată 3 în tabelul 7.5).

Chiar dacă tabelul prezintă cele două eșantioane (internare și externare) separat, în cadrul calculelor se ține cont de valorile de pe coloana 3 din tabelul 7.5, deci de media diferențelor și de deviația standard a diferențelor.

Tabel 7.6. Rezultatele aplicării testului t pentru exemplul 7.11.

	Internare	*Externare*
Media	3,60	2,90
Varianță	3,43	3,43
Număr de observații	5	5
Coeficient de corelație Pearson	0,75	
df	4	
t	1,20	
P(T<=t) unilateral	0,15	
t criticl unilateral	2,13	
P(T<=t) bilateral	0,30	
t critic bilateral	2,78	

7.6 Comparații multiple - testul ANOVA

Până în acest moment am discutat numai cazul comparării a două medii rezultate de pe două eșantioane. Deseori este nevoie de compararea mai multor eșantioane ca în exemplele 7.5 sau 7.6. Este foarte des întâlnit design-ul în care cercetătorul își pune întrebarea dacă un tratament este mai eficient decât altele, ceea ce este cazul exemplului 7.5 sau poate fi vorba despre o condiție medicală cu mai multe categorii (ex. categorii de vârstă) ceea ce este cazul exemplului 7.6.

Exemplu 7.12.

S-a dorit să se evalueze asocierea dintre polimorfismul genei care codifică receptorul vitaminei D (RVD) și obezitate. Polimorfismul RVD are trei genotipuri F/F, f/f și F/f. După selecția aleatoare a unui lot de subiecți și efectuarea genotipării, aceștia au fost împărțiți în trei loturi distincte, pentru fiecare genotip câte un lot. A fost măsurat IMC-ul subiecților. Ca să răspundem la întrebarea studiului trebuie să comparăm mediile IMC-ului subiecților aparținând celor trei grupuri diferite.

Ca să comparăm mediile se pot utiliza teste pentru compararea a două medii, astfel ar trebui comparate pe rând grupul FF cu grupul ff, grupul FF cu grupul Ff și, respectiv grupul ff cu grupul Ff, în total 3 comparații. Rezultatele posibile și interpretarea lor sunt date în tabelul 7.7.

Interpretarea rezultatelor celor trei teste este anevoioasă, după cum se poate observa în tabelul 7.7. Dacă avem în vedere alte studii, unde se compară 4 grupuri (5 comparații), 5 grupuri (10 comparații), ..., 10 grupuri (45 de comparații) etc., interpretarea și analiza devine din ce în ce mai complexă.

Tabel 7.7. Rezultate posibile ale comparațiilor IMC-ului între diversele genotipuri ale polimorfismului RVD: grupul FF cu grupul ff, grupul FF cu grupul Ff și, respectiv grupul ff cu grupul Ff

Comparații între grupul FF cu grupul ff	Comparații între grupul FF cu grupul Ff	Comparații între grupul ff cu grupul Ff	Concluzie
p<0,05	p<0,05	p<0,05	Există diferențe semnificative între loturi
p<0,05	p<0,05	p≥0,05	Există diferențe semnificative între loturi
p<0,05	p≥0,05	p≥0,05	Există diferențe semnificative între loturi
p<0,05	p≥0,05	p<0,05	Există diferențe semnificative între loturi
p≥0,05	p<0,05	p<0,05	Există diferențe semnificative între loturi
p≥0,05	p<0,05	p≥0,05	Există diferențe semnificative între loturi
p≥0,05	p≥0,05	p≥0,05	Nu există diferențe semnificative între loturi
p≥0,05	p≥0,05	p<0,05	Există diferențe semnificative între loturi

Soluția la această problemă de complexitate ridicată este aplicarea testului Anova.
Se dau mediile $\bar{X}_1$, $\bar{X}_2$, …, $\bar{X}_k$ observate pe k eşantioane având n_1, n_2, …, n_k subiecți și S_1, S_2, …, S_k varianţele observate.
Întrebare: Există diferențe statistic semnificative între aceste medii?
Ipoteza nulă H_0: Nu există diferențe semnificative statistic între medii.
Ipoteza alternativă H_1: Există diferențe statistic semnificative între medii.

Cerințele testului Anova:

- Eșantioane independente, distribuția datelor normală, deviațiile standard egale.
- Eșantioane independente, deviațiile standard egale, $n_1=n_2=, \ldots , =n_k$.

Dacă p≥0,05 atunci nu reușim să respingem ipoteza nulă: între mediile testate nu putem spune că există difențe semnificative statistic, deci toate eșantioanele respective aparțin aceleiași populații și în cazul exemplului dat nu putem spune că există legătură între polimorfismul testat și obezitate.

Dacă p<0,05 atunci putem respinge ipoteza nulă și să validăm ipoteza alternativă: între mediile testate există difență semnificativă statistic. Dacă ne situăm în acest caz știm că există diferențe, dar nu știm în care dintre cazurile prezentate în tabelul 7.7 ne situăm (pentru 3 grupuri avem 7 situații posibile).

7.6.1 Analiza post-hoc

Dacă dorim să continuăm analiza pentru a afla care dintre cele 7 situații posibile prezentate în tabelul 7.7 o prezintă datele analizate, atunci va trebui să analizăm eșantioanele două câte două, analiză ce se numește post-hoc.

După cum se poate observa în tabelul 7.7, ca rezultat al acestor comparații vom avea 3 probabilități. În fiecare dintre testări am presupus că eroarea nu trebuie să depășească pragul de 5%. Dar, deoarece avem trei teste aplicate, această eroare se combină și depășește

pragul de 5%. Pentru ca să se preîntâmpine această situație se aplică o corecție asupra acestor probabilități rezultate din analiza post-hoc. Menționăm aici câteva corecții posibile: Bonferroni, Tukey, Dunett, Scheffe, Student-Newman-Keuls, etc. De exemplu corecția Bonferroni se poate realiza foarte simplu: se multiplică fiecare probabilitate rezultată din analiza post-hoc cu numărul total de comparații. în cazul exemplului din tabelul 7.7, ar trebui înmulțită fiecare probabilitate cu 3.

7.6.2 Comparații multiple în cazul eșantioanelor dependente

Și în cazul eșantioanelor dependente, dacă dorim să comparăm eșantioane multiple vom realiza acest lucru cu un test statistic. Exemplu de studiu cu testări repetate avem dat mai sus în exemplul 7.4. În respectivul studiu s-au executat măsurători repetate multiple, în fiecare zi, ale parametrilor urmăriți. Ca să comparăm aceste testări se folosește o procedură numită Anova pentru măsurători repetate, procedură ce implică folosirea mai multor teste și corecții precum: testul de sfericitate (pentru asumpții), testul Mauchly, corecțiile Greenhouse - Geisser sau Huynh – Feldt. Dar aceste teste nu fac obiectul acestui curs, ele necesitând cunoștințe complexe.

7.7 Teste neparametrice

Este de preferat întotdeauna folosirea unui test parametric deoarece acesta este mai puternic decât un test neparametric, în cazul în care distribuția datelor este conformă cu prezumpțiile testului parametric. Unii autori recomandă chiar proiectarea studiilor cu eșantioane egale, astfel încât dacă vor fi abateri de la distribuția normală să se poată utiliza totuși testele t sau Anova la care cerința normalității poate fi eludată, teoretic, în cazul eșantioanelor cu număr egal de indivizi.

Un alt dezavantaj al testelor neparametrice este că nu ne dau informații exacte asupra mărimii diferenței, deoarece aceasta se calculează între ranguri.

7.7.1 Testul Mann-Whitney U

Dacă eșantioanele nu îndeplinesc cerințele enunțate anterior pentru testul t sau cum spune A. Field, atunci când „datele nu arată bine de loc", când s-a întâmplat ceva catastrofal cu ele și „lucrurile nu au mers bine", se poate utiliza testul Mann-Whitney U cunoscut și cu alte denumiri precum Wilcoxon rank-sum test, Mann-Whitney-Wilcoxon. A nu se încurca cu testul Wilcoxon signed-rank test (vezi paragraful următor). Nu este neapărată utilizarea acestui test. Datele se pot normaliza prin transformări, de exemplu prin logaritmare. După logaritmare datele medicale trebuie prezentate utilizând media geometrică. Datele

normalizate devin greu de interpretat din punct de vedere medical. De exemplu o TAS de 140 mmHg devine prin logaritmare 4,94, ceea ce este prea puțin intuitiv. De aceea, se preferă utilizarea testului Mann-Whitney și prezentarea medianelor și a intervalului quartilic (vezi exemplul 7.1).

Cerințele testului neparametric Mann-Whitney sunt: eșantioanele să fie independente, datele în interiorul fiecărui eșantion să fie independente, datele să fie ordinale sau cantitative. Testul Mann-Whitney compară media rangurilor datelor în locul mediilor acestora.

Exemplu 7.13

Să presupunem că avem TAS de la 5 sportivi și TAS de la 5 indivizi sedentari: 120, 130, 180, 130, 115, respectiv 220, 140, 110, 120, 130. Aceste valori se ordonează crescător: 110, 115, 120, 120, 130, 130, 130, 140, 180, 220.

Cea mai mică valoare 110 primește rangul 1, a doua 115 primește rangul 2, a treia și a patra fiind egale primesc același rang (media rangurilor lor), adică 3,5. A 5-a, a 6-a și a 7-a sunt egale și au rangul 6, 140 are rangul 8, 180 are rangul 9 și 220 are rangul 10. Se face media rangurilor pe fiecare lot în parte: (3,5+6+9+6+2)/5=5,3 respectiv 10+8+1+3,5+6=5,7. Analog se calculează și deviațiile standard. Apoi se utilizează testul t pentru compararea mediilor rangurilor (ca metodă aproximativă, sau se utilizează alte tehnici pentru a avea rezultatul exact).

Dacă se găsesc diferențe semnificative între ranguri înseamnă că unul dintre eșantioane este caracterizat prin valori stochastic mai mari ale caracteristicii testate, față de valorile celuilalt grup. În ambele cazuri se poate trage concluzia că nu există, sau respectiv, există, diferențe ale caracteristicii testate între cele două eșantioane. Dacă nu se găsesc diferențe semnificative între ranguri înseamnă că valorile dintre cele două eșantioane sunt amestecate aproximativ uniform.

7.7.2 Testul Wilcoxon pentru eșantioane dependente

În cazul în care cerințele testului t nu sunt îndeplinite se poate utiliza testul neparametric Wilcoxon pentru eșantioane dependente (Wilcoxon signed-rank test).

Cerințele acestui test sunt ca eșantioanele să fie dependente sau perechi, observațiile în cadrul eșantioanelor să fie independente, datele să fie cantitative sau ordinale.

Diferențelor li se atribuie ranguri exact ca în cazul testului Mann-Whitney. Sunt comparate mai apoi rangurile diferențelor negative cu cele pozitive. Aceste ranguri ar trebui să nu difere semnificativ în cazul în care ipoteza nulă este adevărată.

7.7.3 Testul Kruskal-Wallis pentru eșantioane independente multiple

Testul Kruskal-Wallis este echivalentul unui test Anova pentru ranguri. Cerințele necesare pentru aplicarea testului Kruskal-Wallis sunt ca eșantioanele să fie independente, observațiile în cadrul eșantioanelor să fie independente, datele să fie cantitative sau ordinale.

În cazul respingerii ipotezei nule, rezultatul ne indică faptul că avem diferențe între eșantioane ca ordine a valorilor, dar nu vom ști care sunt diferite. Este necesară o analiză post-hoc, ca și în cazul testului Anova. Și în acest caz va fi nevoie de corecții. Se pot aplica corecții precum Bonferroni, Dunn, etc.

7.7.4 Testul Friedman pentru eșantioane dependente multiple

Testul Friedman este echivalentul unui test Anova pentru măsurători repetate pentru ranguri. Cerințele necesare pentru aplicarea testului Friedman sunt ca eșantioanele să fie dependente sau perechi, observațiile în cadrul eșantioanelor să fie independente, datele să fie cantitative sau ordinale.

În cazul respingerii ipotezei nule este necesară o analiză post-hoc, ca și în cazul testului Anova. Și în acest caz va fi nevoie de corecții.

7.7.5 Compararea medianelor

O alternativă a testului t pentru eșantioane independente îl constituie testul Moods pentru compararea a două mediane. Testul pentru mediane testează ipoteza nulă H_0: Nu există diferență semnificativă statistic între medianele celor două eșantioane. În timp ce ipoteza alternativă este H_1: există diferență statistic semnificativă între medianele celor două eșantioane. Parametrul testului se calculează ținând cont de distanța între fiecare observație și mediana eșantionului respectiv.

Acest test nu se folosește prea des, mulți cercetători preferând testul Mann-Whitney, deoarece este mai puternic. Prin „puternic” se înțelege faptul că avem nevoie de diferențe mai mari între valori pentru a trage concluzia că ele sunt semnificative. Testul Mann-Whitney poate fi utilizat pentru a compara mediane doar dacă forma distribuției grupurilor comparate este similară.

7.8 Sinteza metodei de alegere a testelor

Pentru a avea o imagine globală a metodei de alegere a testelor de comparare a variabilelor cantitative este utilă reprezentarea din figura 7.7.

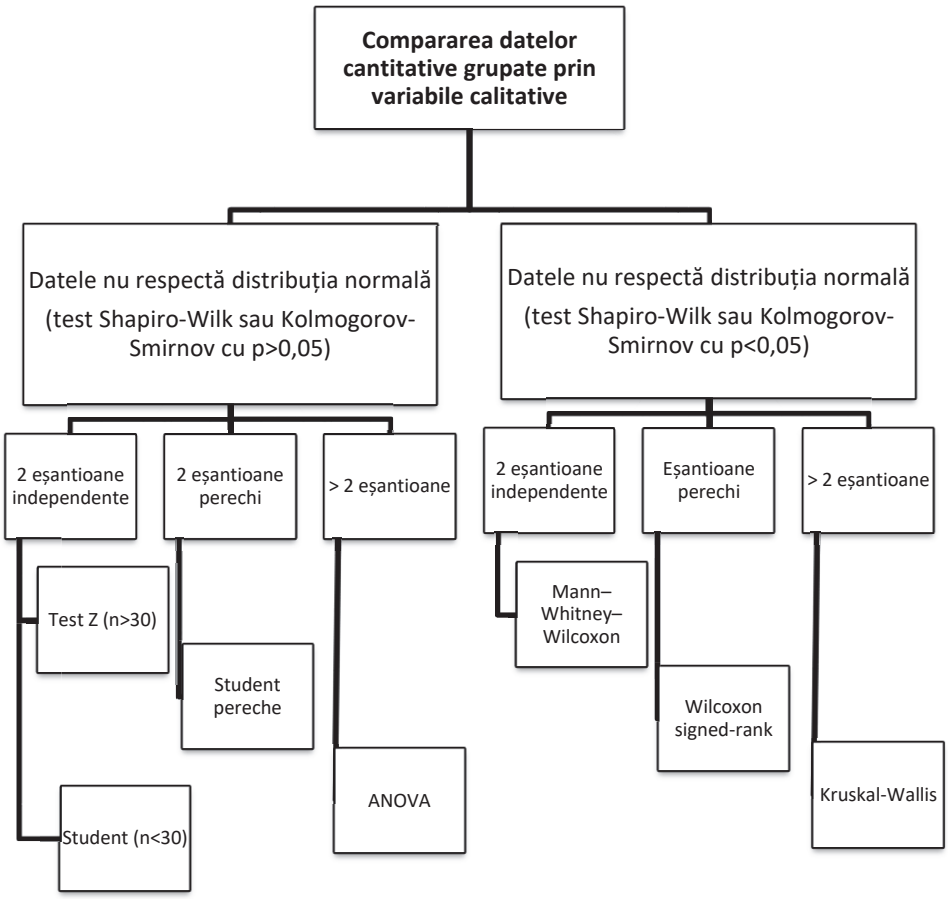


Figura 7.7. Alegerea testelor parametrice/non-parametrice

În cazul datelor normal distribuite selectarea metodei statistice va ține seama de figura 7.8 cu observația că testul Student pentru eșantioane independente poate fi aplicat pentru eșantioane și mai mici de 30 iar când dimensiunea eșantionului este peste 30 parametrul său statistic va fi apropiat de al testului Z, și , ca urmare, poate să îl înlocuiască cu succes.

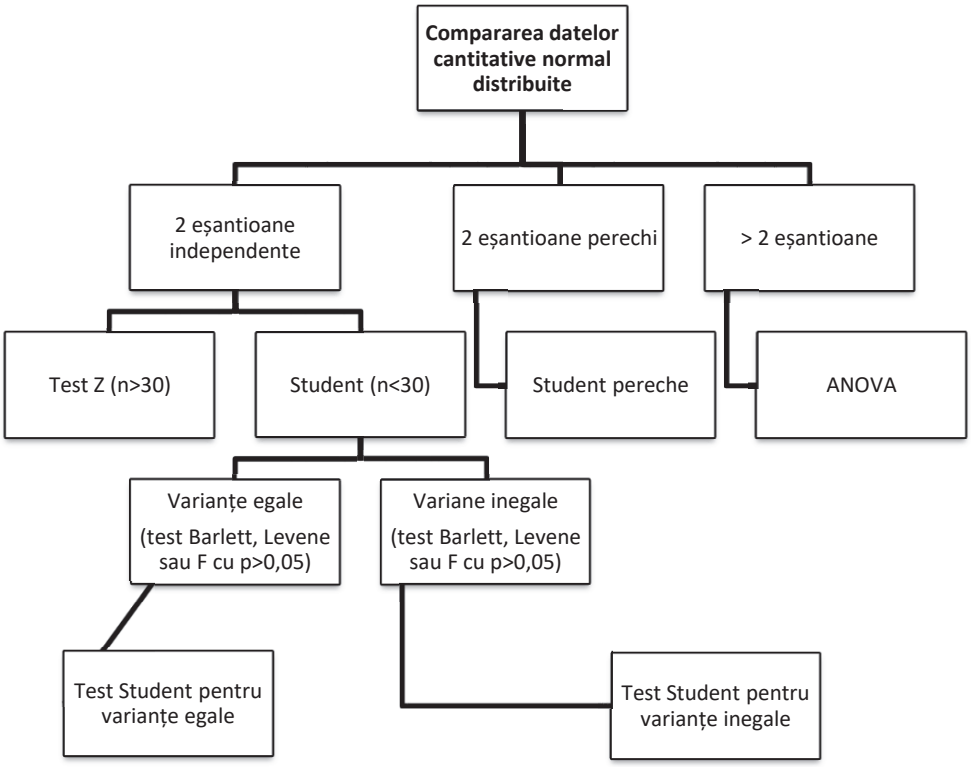


Figura 7.8. Alegerea testelor statistice pentru date normal distribuite

7.9 Exerciții propuse

1. Rezultatele măsurării presiunii inspiratorii pentru 10 pacienţi cu BPOC (boala pulmonară obstructivă cronică) și pentru 10 pacienți indemni de boală (sănătoși) sunt prezentate în tabelul de mai jos:

Număr	Presiunea inspiratorie (cm H_2O) la pacienții cu BPOC	Presiunea inspiratorie (cm H_2O) la subiecții indemni de boală
1	**54,8**	**74,8**
2	**52,0**	**82,0**
3	**63,3**	**63,3**
4	**34,2**	**64,2**
5	**40,3**	**90,3**
6	**36,3**	**46,3**
7	**19,3**	**69,3**
8	**24,6**	**74,6**
9	**46,6**	**86,6**
10	**43,6**	**82,6**

i. Calculaţi media şi deviaţia standard pentru presiunea inspiratorie separat pentru cele două grupuri
ii. Formulați ipoteza nulă și cea alternativă în cazul testului bilateral
iii. Alegeți testul statistic potrivit pentru testarea ipotezei nule
iv. Cu ce formulă de calcul se pot calcula intervalele de încredere de 95% pentru media presiunii inspiratorii la grupul cu BPOC și la cei indemni de boală.
v. Care este efectul asupra intervalului de încredere dacă n, numărul de pacienţi, creşte la 40 în fiecare grup?
vi. Cu ce formulă se calculează intervalul de încredere dacă n, numărul de pacienţi, creşte la 40 în fiecare grup?

2. Cel mai bun test statistic pentru compararea valorilor unui parametru biologic cantitativ obţinute la copii și la adulți este:
A. testul t pentru eşantioane perechi
B. testul Z
C. testul Hi pătrat
D. testul t pentru grupuri independente
E. testul Levene.

3. Media tensiunii arteriale sistolice TAS la un eşantion reprezentativ de 75 de persoane adulte din firma A este de 145 mmHg cu o deviaţie standard de 5. Pentru un eşantion reprezentativ de 95 de persoane din firma B media este de 125 mmHg, iar deviaţia standard de 6. Există deosebiri semnificative statistic între TAS la persoanele adulte din cele două firme? (Utilizaţi $t_{\alpha,168} \cong 1,66$ şi pragul de semnificaţie $\alpha=0,05$) Prezentaţi decizia luată în urma aplicării testului.

4. Să presupunem că pentru 5 persoane s-a măsurat greutatea înainte de efectuarea unui efort (maraton) şi după efort şi s-au obţinut observaţiile:

Greutatea inițială	70	83	79	92	85
Greutatea finală	67	82	75	88	83

Calculaţi media şi abaterea standard pentru diferenţele dintre greutatea înainte şi după efort. Există o deosebire semnificativă statistic între greutatea înainte şi după efort? Aplicaţi testul statistic corespunzător pentru această comparaţie. Interpretaţi rezultatele testului.

5. Dacă dorim să comparăm două variante de pe două eșantioane independente vom utiliza:

A. Test Levene
B. Test Bartlet
C. Test Fisher pentru variante
D. Testul Fisher exact
E. Testul Anova

6. Dacă dorim să comparăm statistic două medii de pe două eșantioane independente cu 20 de indivizi fiecare, vom utiliza mai întâi:

A. Testul Student
B. Testul Kolmogorov-Smirnov
C. Testul Shapiro-Wilk
D. Un test pentru compararea distribuțiilor cu distribuția normală
E. Un test pentru compararea variantelor

8. Compararea distribuțiilor variabilelor calitative

Obiective educaționale. După parcurgerea acestui capitol, studenții:

- Vor putea să formuleze ipotezele statistice pentru date calitative
- Vor putea realiza tabele de contingență observate și teoretice
- Vor interpreta corect concluzia testului Chi pătrat

8.1 Teste de tip Chi-pătrat

Așa cum variabilele cantitative au teste specifice care sunt folosite în funcție de condițiile care sunt îndeplinite, și pentru variabilele calitative avem teste care se bazează pe compararea a două distribuții (frecvențe).

Testul Chi-pătrat (Hi-pătrat), unul dintre cele mai des folosite și citate teste din literatura biomedicală, evaluează relațiile între două sau mai multe variabile calitative. Pentru a aplica acest test, este necesar să se calculeze statistica asociată testului, χ^2 , care urmează o distribuție chi-pătrat. Termenii de Hi sau Chi sunt echivalenți și sunt reprezentarea fonetică a semnului prezentat în acest paragraf din alfabetul grec.

8.1.1 Testul Chi-pătrat de independenţă

Cum se poate testa dacă două variabile calitative sunt independente?

Exemplul 8.1

Presupunem ca dorim să verificăm dacă există o asociere între fumat (F) și apariția unei anumite boli (B). Considerăm ca avem un eșantion de 400 de indivizi împărțiți astfel: 160 prezintă boala B (240 nu prezintă boala) și 130 sunt fumători F (270 nu sunt fumători).

Ambele variabile sunt calitative dicotomiale, deoarece o persoană fie fumează, fie nu fumează, iar o boala este prezentă sau absentă. Repartiţia indivizilor în funcţie de cele două criterii se prezintă în următorul tabel de contingenţă (2×2):

Fumat	Boala		
	B bolnavi	$\bar{B}$ sănătoși	Total
F (fumători)	80	50	130
$\bar{F}$ (nefumători)	80	190	270
Total	160	240	400

Se caută să se stabilească dacă fumatul influenţează apariţia bolii B sau dacă apariţia acesteia este independentă de fumat.

Tabelul de contingenţă de mai sus se numeşte tabel de contingenţă observat, iar frecvenţele pe care le conţine se numesc frecvenţe observate.

Principiul testului Chi-pătrat

La fel ca în cazul tuturor testărilor ipotezelor statistice, testul χ^2 se bazează pe un raţionament de reducere la absurd.

Astfel, dacă se va testa ipoteza de independenţă între cele două caractere F şi B, atunci se poate calcula un *tabel de contingenţă teoretic* care satisface această ipoteză de independenţă. Se determină apoi *abaterea* (*ecartul*) dintre cele două tabele de contingenţă observat şi teoretic. Dacă această abatere este mică, atunci ea este explicată doar prin întâmplare (hazard) şi ipoteza de independenţă este validată. Dacă din contră, această abatere este prea mare pentru ca doar întâmplarea să o explice, atunci ipoteza de independenţă trebuie să fie respinsă. Ipotezele noastre statistice sunt:

- Ipoteza nulă (H_0): Fumatul (F) și diagnosticul de boală (B) sunt independente.
- Ipoteza alternativă (H_1): Fumatul (F) și diagnosticul de boală (B) sunt dependente.

Calculul tabelului de contingenţă teoretic

Problema este următoarea: dispunând de un eşantion de n = 400 de subiecţi dintre care 160 prezintă boala B, iar 130 sunt fumători, să se determine cum sunt repartizaţi subiecţii în funcţie de cele două caractere (F şi B) dacă se presupune că acestea sunt independente.

Trebuie deci să fie completat următorul tabel de contingenţă teoretic (numit şi *tabel de contingenţă calculat*):

Fumat	Boala		
	B	$\bar{B}$	Total
F			130
$\bar{F}$			270
Total	160	240	400

Folosind ipoteza de independenţă dintre cele două caractere F şi B, putem calcula probabilităţile:

$$\Pr(B \cap F) = \Pr(B) \times \Pr(F)$$

unde:

Pr(B $\cap$ F) este probabilitatea de a avea simultan caracterele B şi F
Pr(B) este probabilitatea de a avea caracterul B
Pr(F) este probabilitatea de a avea caracterul F.

Probabilităţile se calculează astfel:

$$\Pr(\mathrm{B}) = \frac{\text{Numarul de indivizi cu boala B}}{\text{Numarul total de indivizi}} = \frac{160}{400}$$

şi analog,

$$\Pr(\mathrm{F}) = \frac{130}{400},$$

$$\Pr(\mathrm{B} \cap \mathrm{F}) = \frac{\text{Numarul de indivizi cu B si F}}{\text{Numarul total de indivizi}} = \frac{F^t(F,B)}{400},$$

unde $F^t(F,B)$ este frecvenţa teoretică (căutată) din prima căsuţă a tabelului de contingenţă teoretic.

Deci $F^t(F,B)$ se calculează prin formula:

$$F^t(F,B) = \frac{\Pr(\mathrm{B}) \times \Pr(\mathrm{F})}{\mathrm{n}} = \frac{130 \times 160}{400} = 52\,.$$

Observaţii:

1) $F^t(F,B)$ este un raport având la numărător produsul totalurilor parţiale ale liniei şi respectiv coloanei corespunzătoare şi la numitor totalul general.

Analog se pot calcula şi celelalte frecvenţe teoretice, aşa că tabelul de contingenţă teoretic devine:

Fumat	Boala		
	B	$\bar{B}$	Total
F	$\frac{130 \times 160}{400}$	$\frac{130 \times 240}{400}$	130
$\bar{F}$	$\frac{270 \times 160}{400}$	$\frac{270 \times 240}{400}$	270
Total	160	240	400

Acest mod de calcul se aplică şi în cazul general, atunci când cele două caractere studiate au fiecare un număr de modalităţi (valori) de realizare arbitrar (≥ 2).

2) Se observa uşor că având calculată frecvenţa teoretică $F^t(F,B) = 52$ celelalte frecvenţe teoretice nu mai sunt necesar să fie direct calculate prin formulele prezentate în tabele, ci se pot uşor determina prin diferenţe din totalurile de linie sau coloană:

Fumat	Boala		
	B	$\bar{B}$	Total
F	52	130-52	130
$\bar{F}$	160-52	270-(160-52)	270
Total	160	240	400

și se obţine tabelul de contingenţă teoretic:

Fumat	Boala		
	B	$\bar{B}$	Total
F	52	78	130
$\bar{F}$	108	162	270
Total	160	240	400

Se poate astfel constata că pentru un tabel de contingenţă teoretic (2×2), este suficient să calculeze o frecvenţă teoretică pentru a putea determina tabelul în întregime.

Această proprietate se regăseşte şi în cazul general a unui tabel cu L rânduri şi C coloane, unde este uşor de constatat faptul că este suficient să se calculeze primele $(L-1) \times (C-1)$ frecvenţe teoretice, celelalte obţinându-se prin diferenţe. Se va vedea că produsul $(L-1) \times (C-1)$ defineşte numărul de grade de libertate al lui χ^2.

Ecartul dintre cele două tabele de contingenţă

Fie O_i şi T_i frecvenţele observate şi teoretice (i=1,2,..,n) situate în aceleaşi poziţii în cele două tabele. Ecartul între cele două tabele notat cu X^2 se calculează prin formula:

$$X^2 = \sum_{i=1}^{L \cdot C} \frac{(O_i - T_i)^2}{T_i}.$$

Distribuţia χ^2

Se poate arăta că dacă ipoteza de independenţă este satisfăcută, atunci X^2 determinat prin formula de mai sus se supune unei legi de probabilitate χ^2 (Chi-pătrat) cu $(L-1) \times (C-1)$ grade de libertate.

Pentru această lege χ^2 se poate determina valoarea χ^2_α ce corespunde pragului de semnificaţie α (de obicei α=0,05) şi care verifică condiţia:

$$\Pr(\chi^2 \geq \chi^2_\alpha) = \alpha .$$

Cu ajutorul acestei valori se defineşte regiunea critică a testului $[\chi^2_\alpha, \infty)$, unde sub ipoteza nulă, X^2 are o probabilitate α de a se găsi.

Etapele aplicării testului Chi-pătrat

În continuare se vor prezenta cele şase etape ale testului Chi-pătrat utilizat pentru testarea independenţei a două caractere.

Tabelul 8.1. Etapele testului Chi pătrat pentru exemplul 8.1

	Cazul general	**Ilustrarea printr-un exemplu**
Problema	Se încearcă să se determine, cu ajutorul unui eşantion de n subiecţi, dacă două caractere A şi B având L şi respectiv C modalităţi de realizare sunt sau nu independente.	Fumatul (F) şi o boală (B) sunt independente? În acest caz, L=C=2, iar eşantionul observat are n=400 subiecţi repartizaţi în tabelul de contingenţă prezentat mai sus.
Etapa 1. Definirea ipotezei nule H_0	H_0: caracterele A şi B sunt independente.	H_0: fumatul nu are influenţă asupra apariţiei bolii B.
Etapa 2. Definirea unui parametru	$X^2=\sum_{i=1}^{L\cdot C}\frac{(O_i-T_i)^2}{T_i}$ urmează o lege χ^2 cu (L-1) x (C-1) grade de libertate	$X^2=\sum_{i=1}^{L\cdot C}\frac{(O_i-T_i)^2}{T_i}$ urmează o lege χ^2 cu 1 grad de libertate.
Etapa 3. Pragul de semnificaţie	Fie α pragul de semnificaţie al testului.	S-a ales pragul de semnificaţie α = 0.05
Etapa 4. Definirea regiunii critice	Ţinând seama de faptul că X^2 urmează legea χ^2 cu (L-1) x (C-1) grade de libertate se determină valoarea χ^2_α încât $P(\chi^2 \geq \chi^2_\alpha)= \alpha$. Regiunea critică este $[\chi^2_\alpha,\infty)$.	Pentru pragul α =0.05 şi χ^2 cu 1 grad de libertate valoarea χ^2_α = 3,84, astfel că în acest caz regiunea critică este intervalul [3,84 , ∞).
Etapa 5. Calcularea valorii observate a parametrului	Se calculează frecvenţele teoretice $T_i=\frac{totallinie\times totalcoloana}{n}$ Se calculează $X^2=\sum_{i=1}^{L\cdot C}\frac{(O_i-T_i)^2}{T_i}$	Se calculează $X^2=\frac{(80-52)^2}{52}+\frac{(50-78)^2}{78}+\frac{(80-108)^2}{108}+\frac{(190-162)^2}{162}$ =37,2
Etapa 6. Decizia	Dacă $X^2\in[3,84,\infty)$ se respinge H_0 cu un risc de eroare de prima speţă $\leq\alpha$. Dacă $X^2\notin[3,84,\infty)$ atunci H_0 nu se respinge, neputându-se respinge H_0 cu un risc de eroare de speţa a doua β	X^2 >> 3,84 deci ipoteza nulă H_0 se respinge cu un risc inferior lui 5%. În concluzie, *fumatul se asociază cu boala B. Atenție direcția asociației dintre fumat și boală trebuie demonstrată prin parametri medicali.*

Condiţii de aplicare

Testele pentru variabilele calitative, la fel ca orice test statistic, trebuie să îndeplinească anumite cerinte:

- Eșantioanele sunt extrase aleator din populație și păstrează toate caracteristicile acesteia.
- Fiecare subiect face parte dintr-un singur grup sau clasă, acestea fiind bine definite.
- Frecvenţele din tabelul de contingență teoretic teoretice trebuie să fie minim 5 pentru cel puțin 80% din celule (Atenție la tabelele 2x2 toate celulele trebuie să aibă valorile mai mari sau egale cu 5).
- Nicio frecvență nu trebuie să fie nulă (0).

Observaţii:

- Testul χ^2 nu demonstrază decât asocierea fenomenelor nu și direcția relației. În exemplul anterior este necesară și asocierea unui parametru care să descrie din punct de vedere medical direcția și puterea legăturii în speță riscul relativ.
- În cazul tabelelor de contingenţă 2x2 dacă tabelul de contingenţă observat este de forma (tabelul 8.2):

Tabelul 8.2. Tabel de contingență observat

a	b	
c	d	
		n

se poate arăta că :

$$\chi^2 = \frac{(ad-bc)^2 \cdot n}{(a+b)\cdot(c+d)\cdot(a+c)\cdot(b+d)},$$

ceea ce în anumite cazuri simplifică calculele.

8.1.2 Testarea normalităţii unei distribuţii

Pentru testarea normalității se poate utiliza și testul Chi-pătrat (Chi Square goodness of fit), prin care frecvenţele (de clasă sau cumulate) observate sunt comparate cu frecvenţele teoretice (de clasă sau cumulate) determinate utilizând legea normală.

Ipotezele statistice care se formulează pentru acest test sunt:

- Ipoteza nulă (H_0): Eșantionul este normal distribuit.
- Ipoteza alternativă (H_1): Eșantionul nu este normal distribuit.

Analog aplicării testului Chi pătrat pentru testarea independenței, se calculează statistica testului χ^2, care se verifică dacă se găsește în regiunea critică a testului [χ^2_α,∞). Dacă $\chi^2 > \chi^2_\alpha$, atunci ipoteza nulă se respinge, adică populația din care a fost extras eșantionul nu este normal distribuită. În caz contrar, normalitatea se acceptă.

8.1.3 Testul Chi-pătrat de semnificaţie într-un tabel 2 x 2

Scopul testului este să determine dacă există diferențe semnificative între frecvenţele observate pentru două variabile dicotomiale.

Condiţii de aplicare

1.Eșantioanele sunt extrase aleator din populație și păstrează toate caracteristicile acesteia.

2.Fiecare subiect face parte dintr-un singur grup sau clasă, acestea fiind bine definite.

3.Este necesar ca talia celor două eşantioane să fie suficient de mare. Această condiţie este satisfăcută dacă n > 20 şi frecvenţele din tabelul de contingență teoretic teoretice trebuie să fie minim 5, pentru cel puțin 80% din celule și nicio celulă să nu fie 0.

4. Observațiile în cadrul eșantioanelor să fie independente

Atunci când se aplică distribuţii continue unor valori discrete, este necesară aplicarea corecţiei lui Yates, cunoscută și sub denumirea de corecția de continuitate. Prin corecția Yates se obține o aproximare mai bună a distribuției binomiale, dar se obține mai greu semnificație statistică decât la aplicarea în mod direct a testului Chi pătrat.

Algoritmul testului

Atunci când două eşantioane sunt divizate fiecare în două clase, se poate construi următorul tabel de contingenţă 2 x 2 (tabelul 8.3):

Tabelul 8.3. Tabel de contingență 2x2

	C1	C2	Total
Eşantionul 1	a	b	a + b
Eşantionul 2	c	d	c +d
Total	a + c	b + d	n = a + b + c + d

Statistica testului este comparată cu valoarea obţinută din tabelul χ^2 cu 1 grad de libertate. Dacă χ^2 este mai mare decât valoarea critică, vom respinge ipoteza nulă de independenţă între eşantioane şi clase. Altfel spus, cele două eşantioane nu au fost extrase din aceeaşi populaţie.

În continure vom relua exemplul prezentat mai devreme în cazul testului χ^2 pentru corectitudinea potrivirii, dar statistica testului va fi calculată într-un mod diferit, specific situaţiei unui tabel de (2×2) (cu un singur grad de libertate).

Exemplul 8.2

Într-un studiu se urmăreşte în ce măsură fumatul în timpul sarcinii se asociază cu apariţia preeclampsiei. Dintr-un lot de 200 de gravide la care s-a obţinut prin anamneză expunerea la factorul de risc, fumat, s-a obţinut următorul tabel de contingenţă (2×2):

Tabel observat	Preeclampsie prezentă	Preeclampsie absentă	Total
Fumat prezent	20	6	26
Fumat absent	80	94	174
Total	100	100	200

Ipoteza nulă (H_0)**:** Fumatul în timpul sarcinii nu influenţează apariţia preeclampsiei. Cu alte cuvinte, eşantionul cuprinzând pacientele care fumează în timpul sarcinii provine din aceeaşi populaţie ca şi cel al pacientelor care nu fumează în timpul sarcinii, din punct de vedere al frecvenţei preeclampsiei.

$$\chi^2 = \frac{(n-1)(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

Statistica χ^2 se va calcula conform formulei de mai sus:

$$\mathrm{X}^2 = \frac{(200-1)\cdot(20\cdot 94 - 6\cdot 80)^2}{(20+6)\cdot(20+80)\cdot(6+94)\cdot(80+94)} \cong 4{,}7$$

Se va regăsi apoi în tabel valoarea critică χ^2 (0,05) cu 1 grad de libertate şi anume 3,84.

Deoarece 4,7 >3,84 ipoteza nulă poate fi respinsă cu un risc de primă speţă <0,05, validându-se aşadar ipoteza alternativă, adică faptul că frecvența preeclampsiei este diferită între lotul care a fumat și cel care nu a fumat. Încă odată testul statistic demonstrează doar matematic existența unei diferențe semnificative între frecvențe; din punct de vedere medical este necesar să folosim un alt parametru, riscul relativ, pentru descrierea fenomenului din punct de vedere medical.

8.1.4 Testul hi-pătrat de semnificaţie într-un tabel k x 2

Scopul testului este investigarea semnificaţiei diferenţelor dintre k distribuţii de frecvenţe observate, cu clasificare dicotomială.

Condiţii de aplicare

1.Eşantioanele sunt extrase aleator din populație și păstrează toate caracteristicile acesteia.

2.Fiecare subiect face parte dintr-un singur grup sau clasă, acestea fiind bine definite.

3. Este necesar ca talia celor k eşantioane să fie suficient de mare. Frecvențele din tabelul de contingență teoretic teoretice trebuie să fie minim 5, pentru cel puțin 80% din celule și nicio celulă să nu fie 0.

4. Observațiile în cadrul eșantioanelor să fie independente

Algoritmul testului

Atunci când observaţiile din cele k eşantioane sunt divizate fiecare în două clase, se poate construi următorul tabel de contingenţă k x 2 (tabelul 8.4):

Tabelul 8.4. Tabelul de contingență k x 2

	C1	C2	Total
Eşantionul 1	x_1	$n_1 - x_1$	n_1
:	:	:	:
Eşantionul i	x_i	$n_i - x_i$	n_i
:	:	:	:
Eşantionul k	x_k	$n_k - x_k$	n_k
Total	$x = \sum_{i=1}^{k} x_i$	n - x	$n = \sum_{i=1}^{k} n_i$

Statistica testului este:

$$\chi^2 = \frac{n^2}{x(n-x)}\left(\sum_{i=1}^{k}\frac{x_i^2}{n_i} - \frac{x^2}{n}\right).$$

Această valoare se compară cu valoarea critică obţinută din tabelul χ^2 cu (k – 1) grade de libertate.

Dacă χ^2 este mai mare decât valoarea critică, se respinge ipoteza nulă de independenţă între eşantioane şi clase.

Exemplul 8.4

Sunt studiate trei eșantioane formate din trei grupe de vârstă a femeilor gravide. Se dorește să se verifice dacă exisă o posibilă asociere cu apariția preeclampsiei. În total este investigat un număr de 80 de persoane, obţinându-se următorul tabel de contingenţă 3×2:

Tabel observat	Preeclampsie prezentă	Preeclampsie absentă	Total
Vârsta <20 de ani	14	22	36
Vârsta 20 - 40 de ani	18	16	34
Vârsta > 40 de ani	8	2	10
Total	40	40	80

Ipoteza nulă H_0: Prezenţa preeclampsiei nu depinde de grupa de vârstă din care face parte gravida.

Statistica χ^2 se va calcula conform formulei:

$$\chi^2 = \frac{80^2}{40\times 40}\left[\frac{14^2}{36}+\frac{18^2}{34}+\frac{8^2}{10}\right]-\frac{40^2}{80}=5,495$$

Valoarea critică χ^2 (0,05) cu 2 grade de libertate se regăseşte în tabelul testului şi este 5,99. Deoarece valoarea calculată a lui χ^2 este mai mică decât valoarea critică, ipoteza nulă nu poate fi respinsă. Adică, preeclampsia nu depinde de grupa de grupa de vârstă din care face parte gravida.

8.2 Testul exact al lui Fisher

Testul exact al lui Fisher (incomplet tradus Fisher Exact) este un test neparametric care se utilizează atunci când nu este îndeplinita una din condițiile necesare aplicării testului Chi pătrat, și anume atunci când datele pe care le avem provin din eșantioane mici sau cu distribuție slabă (frecvențe sub 5 în tabelul de contingență teoretic în peste 20% din celule, sau o frecvență de 0 într-o celulă). Testul exact al lui Fisher se bazează pe calculul distribuției hipergeometrice.

Algoritmul testului

Atunci când două eşantioane sunt divizate fiecare în două clase, se poate construi un tabel de contingenţă 2 x 2.

În continuare se calculează probabilitatea prag (cut-off) folosind formula:

$$p = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{n!\,a!\,b!\,c!\,d!}$$

Observație: ! inseamnă factorial, adică 3!=1·2·3=6.

Algoritmul continuă cu găsirea tuturor tabelelor de frecventă posibile cu aceleași valori marginale (au aceleași totaluri pe linii și coloane) și calcularea probabilității asociate folosind formula de mai sus. Suma tuturor probabilităților pentru toate combinațiile posibile ale tabelui de contingență cu aceleași valori marginale este egală cu 1.

Determinarea probabilității p corespunzătoare testului exact al lui Fisher se face prin adunarea la valoarea p_{cutoff} calculată, a probabilităților care au valori mai mici decât aceasta, obținute din tabele de contingență care au avut valori mai extreme decât tabelul nostru inițial.

Dacă p<0,05, atunci ipoteza nulă se respinge.

Exemplul 8.5

Presupunem că suntem interesați să determinăm dacă există o legătură între apariția unui accident vascular cerebral (AVC) și administrarea unei anumite substanțe halucinogene. Substanțele halucinogene fiind interzise de legislație, iar AVC-ul la persoanele tinere fiind destul de rar, colectarea unor date așa încât să reușim să îndeplinim criteriile de aplicare a testului Chi pătrat este practic imposibilă. Testul folosit în această situație va fi testul exact al lui Fisher.

Tabel observat	AVC prezent	AVC absent	Total
Consum substanțe halucinogenă	4	1	5
Fără consum de substanțe halucinogenă	0	5	5
Total	4	6	10

Ipoteza nulă (H_0)**:** Nu există legătură între consumul de substanță halucinogenă și apariția AVC.

Calculăm valoarea lui p folosind formula de mai sus:

$$p = \frac{5!\,5!\,4!\,6!}{10!\,4!\,1!\,0!\,5!} = 0{,}024$$

Pentru a putea decide în legătură cu ipoteza nulă, pe lângă probabilitatea pe care am obținut-o (care este probabilitatea obținută din tabelul de frecvență observat), este necesar să calculăm analog, probabilitățile pentru toate posibilele tabele de frecvențe observate (pentru care se păstrează totalurile). Tabele de contingență posibile și probabilitățile corespunzătoare sunt:

	AVC prezent	AVC absent	Total
Consum SH	2	3	5
Fără SH	2	3	5
Total	4	6	10

$$p = \frac{5!\,5!\,4!\,6!}{10!\,2!\,3!\,2!\,3!} = 0{,}476$$

	AVC prezent	AVC absent	Total
Consum SH	3	2	5
Fără SH	1	4	5
Total	4	6	10

$$p = \frac{5!\,5!\,4!\,6!}{10!\,3!\,2!\,1!\,4!} = 0{,}238$$

	AVC prezent	AVC absent	Total
Consum SH	1	4	5
Fără SH	3	2	5
Total	4	6	10

$$p = \frac{5!\,5!\,4!\,6!}{10!\,1!\,4!\,3!\,2!} = 0{,}238$$

	AVC prezent	AVC absent	Total
Consum SH	0	5	5
Fără SH	4	1	5
Total	4	6	10

$$p = \frac{5!\,5!\,4!\,6!}{10!\,0!\,5!\,4!\,1!} = 0{,}024$$

Suma dintre p_{cutoff} și probabilitățile care sunt mai mici sau egale cu p_{cutoff} este:

p=0,024+0,024=0,048

p<0,05, deci ipoteza nulă se respinge, adică există o asociere intre consumul de substanțe halucinogene și apariția AVC.

8.3 Testul McNemar

Testul McNemar este asemănător testului Chi-pătrat, foloseste aceeași distribuție cu acesta (de aici și parametrul statistic identic), utilizându-se în situația în care avem eșantioane perechi.

Condiţii de aplicare

1.Eșantioanele au fost extrase aleator din populație și păstrează toate caracteristicile acesteia.

2.Fiecare subiect face parte dintr-un singur grup (clasă), care este bine definit.

4. Observațiile în cadrul eșantioanelor să fie independente

5. Suma valorilor din celulele b+c>25

Algoritmul testului

Considerăm ca datele sunt prezentate în tabelul 10.8 de contingenţă 2 x 2.

Statistica testului este:

$$\chi^2 = \frac{(b-c)^2}{b+c}$$

Analog cu testul chi-pătrat, această valoare este apoi comparată cu valoarea obţinută din tabelul χ^2 cu 1 grad de libertate.

Dacă χ^2 este mai mare decât valoarea critică, vom respinge ipoteza nulă de independenţă între eşantioane şi clase. Altfel spus, cele două eşantioane nu au fost extrase din aceeaşi populaţie.

Exemplul 8.6

Pacienților diabetici li s-a recomandat o dietă care are ca efect scădera glicemiei. Prin studiul nostru dorim să urmarim dacă această dietă a avut si un rol în scăderea colesterolului. Avem un eșantion de 70 de diabetici:

Tabel observat	*Înainte de dietă*	*După dietă*	Total
Hipercolesterolemie	15	25	40
Normocolesterolemie	10	20	30
Total	25	45	70

Ipoteză nulă (H_0): nu există o legătură între hipercolesterolemie și dieta urmată.

Calculăm valoarea lui χ^2 folosind formula de mai sus:

$$\chi^2 = \frac{(25-10)^2}{25+10} = \frac{225}{35} = 6{,}4$$

Valoarea obținută se compară cu valoarea critică a χ^2 (0,05) cu 1 grad de libertate şi anume 3,84.

Deoarece 6,4 >3,84 ipoteza nulă poate fi respinsă cu un risc de primă speţă <0,05, validându-se aşadar ipoteza alternativă, adică faptul că dieta are efect asupra colesterolului.

8.4 Alegerea testului adecvat pentru compararea distribuțiilor de frecvențe

Pentru alegerea corectă a testului statistic adecvat pentru diferitele situații întâlnite în practică recomandăm utilizarea arborelui decizional din figura 8.1. Menționăm că pentru tabelele de contingență din cauza dimensiunii acestora de numai 4 celule regula de 80% devine regulaabsolutăŞ toate valorile așteptate trebuie să fie mai mari sau egale cu 5

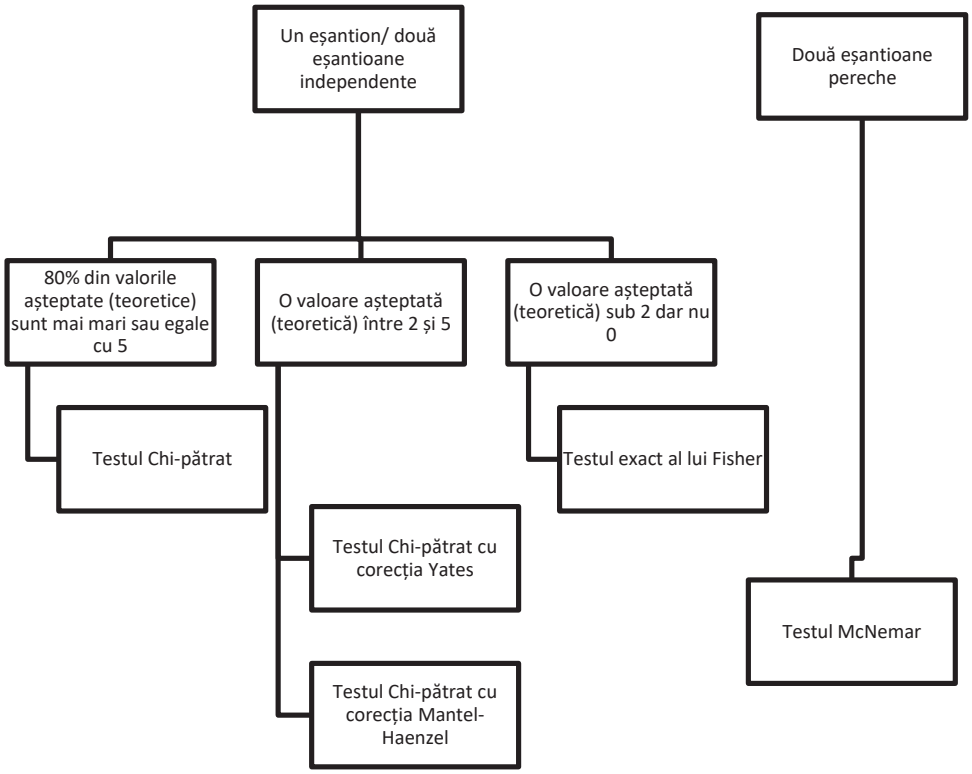


Figura 8.1. Alegerea testului adecvat pentru compararea distribuțiilor de frecvențe

8.5 Exerciții propuse

1. Se dorește să se stabilească dacă există o asociere între fumat și gen (bărbați, femei). Datele au fost colectate aleator și se află în tabelul de mai jos:

Tabel observat	Fumător	Nefumător	Total
Bărbați	15	25	40
Femei	10	20	30
Total	25	45	70

Specificați ipotezele statistice, aplicați și interpretați testul ales.

2. Se dorește să se stabilească dacă există o asociere între fumat și gen (bărbați, femei). Datele au fost colectate aleator și se află în tabelul de mai jos.

Tabel observat	Bărbați	Femei	Total
Fumător	15	10	25
Fost fumător	5	2	7
Nefumător	20	18	38
Total	40	30	70

Precizați care sunt ipotezele testului, numele testului folosit și numărul gradelor de libertate

3. Un experiment a fost efectuat pentru a verifica dacă există o legătură între consumul de alcool și apariția AVC. Dintr-un total de 400 de subiecți, 120 au făcut AVC, iar din aceștia 88 consumau frecvent alcool. Dintre pacienții care nu au făcut AVC, 218 nu consumau frecvent alcool.

a. Creați tabelul de contingență observat
b. Calculați tabelul de contingență teoretic
c. Formulați ipotezele statistice
d. Calculați statistica testului
e. Interpretați rezultatul obținut

9. Corelații și regresii

Obiective didactice:

- Înțelegerea noțiunilor de coeficient de corelație și a celei de funcție de regresie
- Înțelegerea importanței relației direcționale pentru abordarea cauzalității fenomenelor
- Alegerea coeficienților și modelelor de regresie corespunzătoare experimentelor abordate
- Interpretarea indicilor de corelație atât direct cât și împreună cu testele statistice corespunzătoare

9.1 Relații între variabile cantitative

Cuantificarea legăturilor dintre două sau mai multe variabile cantitative se poate face prin stabilirea unor mărimi matematice, indicii de corelație, iar modelarea matematică a acestei legături printr-o funcție care poartă numele de regresie. Validarea indicilor de corelație și a funcțiilor de regresie se face prin teste statistice specifice. Suntem obișnuiți cu noțiunile de proporționalitate sau dependență care pot fi exemplificate prin numeroase exemple de variabile legate, intuitiv, unele de altele: înălțimea și greutatea, vârsta și tensiunea arterială, etc.

Cel mai simplu caz se poate obține prin colectarea a două caracteristici cantitative (variabile) dintr-un eșantion de talie n: $X: X_1, X_2,..., X_n$, $Y: Y_1, Y_2,..., Y_n$.

Regula generală este ca între cele două variabile să se respecte direcționalitatea legăturii de cauzalitate, X determină Y. Altfel, de exemplu, inversarea direcției relației dintre vârsta şi tensiunea arterială ar reprezenta o abordare medicală inedită deoarece orice hipotensor ar reduce vârsta pacienților.

Relația între două variabile cantitative poate fi abordată din două puncte de vedere:

- Calculul numeric al intensității legăturii dintre cele două variabile se realizează cu ajutorul unor indici de corelaţie dintre care cel mai cunoscut este coeficientul de corelaţie.
- Dacă acești indici denotă existența unei relații între variabile se poate determina tipul de legătură care face ca Y să depindă de X, adică determinarea unei funcţii f numită funcţie de regresie, astfel încât Y = f(X). În acest caz, una dintre variabile, care se numeşte ***variabila independentă*** (X), poate fi controlată nealeator, iar cealaltă numită ***variabila dependentă*** (Y) ia valori care nu sunt controlate (variază aleator).

Cea mai simplă și cea mai utilizată funcţie de regresie este cea liniară care aproximează și relația empirică de proporționalitate:

$$f(X) = a + b\,X.$$

9.1.1 Reprezentarea grafică: diagrama de dispersie

Diagrama de dispersie asociată unei tabel de date bidimensional:

$$X: X_1, X_2, \ldots, X_n \qquad Y: Y_1, Y_2, \ldots, Y_n$$

se obţine reprezentând grafic punctele de coordonate (X_i, Y_i) i=1,2,...,n.

Orientarea şi dispersia norului de puncte poate da o primă idee despre relaţia între cele două variabile aflate în studiu, figurile 9.1 și 9.2.

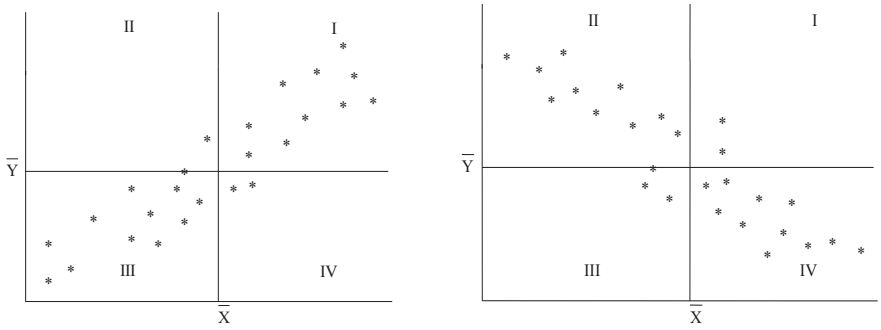


Figura 9.1. Diagrame de dispersie având o tendinţă crescătoare (stânga) sau crescătoare (dreapta)

Existența unei relaţii liniare între cele două variabile va determina ca punctele diagramei să fie dispuse preponderent în anumite cadrane: numeroasele puncte din cadranele I şi III în figura 9.1 (stânga) denotă o relație de proporționalitate directă, iar densitatea de puncte din cadranele II şi IV din partea dreaptă a aceleiași figuri sugerează o relație de proporționalitate inversă. În lipsa unei relaţii între X şi Y, punctele diagramei de dispersie se vor repartiza relativ uniform în cele patru cadrane, figura 9.2.

Nu întotdeauna neîncadrarea într-una dintre situaţiile ilustrate prin figura 9.1 înseamnă că între variabilele X şi Y nu există nici o legătură deoarece este posibil să existe o relaţie descrisă de o funcție matematică mai complicată. În acest caz, aspectul diagramei de dispersie ne indică cea mai adecvată clasă de funcţii care poate fi folosită pentru regresie.

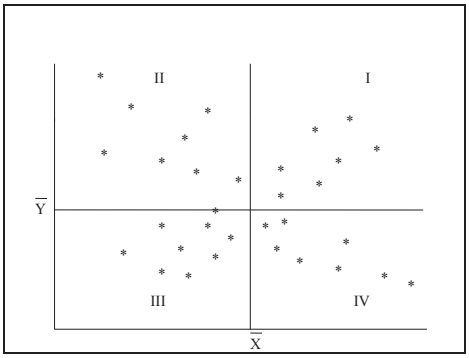


Figura 9.2. Diagramă de dispersie având norul de puncte dispersat

9.2 Indici de corelație

Cei mai importanți indici de corelație sunt suma produselor ecart (SPE), covarianţa COV(X,Y), coeficientul de corelaţie Bravais-Pearson, coeficientul de determinare şi coeficientul de corelaţie Spearman.

9.2.1 Suma produselor ecart

Pentru a se utilizează observaţia că dacă punctul (X_i,Y_i) se află în cadranele I sau III ale diagramei de dispersie atunci produsul este pozitiv, iar atunci când este situat în cadranele II şi IV este negativ. Astfel că, o măsură a intensităţii relaţiei dintre variabilele X şi Y este dată de Suma produselor $(X_i - \overline{X})(Y_i - \overline{Y})$ descrie "intensitatea" relaţiei dintre cele două variabile X şi Y:

$$SPE = \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$$

și este numită suma produselor ecart. Are valore pozitivă pentru un nor de puncte cu alură crescătoare și negativă pentru datele invers proporționale. Un dezavantaj evident al SPE este faptul că acest coeficient depinde de numărul de puncte din seria statistică şi de unităţile de măsură ale variabilelor.

9.2.2 Covarianţa

Pentru a reduce variabilitatea SPE prin împărțirea acestuia la talia eșantionului rezultă covarianța:

$$COV(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}).$$

care are în continuare o exprimare folosind unităţile de măsură ale variabilelor.

9.2.3 Coeficientul de corelaţie Bravais-Pearson

Pentru datele normal distribuite un indicator independent de unităţile de măsură ale celor două variabile se obţine prin împărţirea covarianţei la produsul abaterilor standard şi poartă numele de coeficientul de corelaţie sau coeficientul Bravais-Pearson:

$$r = \frac{COV(X,Y)}{Sx \cdot Sy} = \frac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{\left[n\sum X^2 - \left(\sum X^2\right)\right]\left[n\sum Y^2 - \left(\sum Y^2\right)\right]}}.$$

Dintre proprietăţile coeficientului de corelaţie menţionăm: este întotdeauna cuprins între -1 şi 1 şi cu cât valoarea sa absolută se apropie de 1 cu atât mai mult "intensitatea" relaţiei liniare între cele două variabile va fi mai mare. Când r este pozitiv relaţia între variabilele X şi Y este "pozitivă", adică o creştere a lui X determină în general o creştere a lui X. Când r < 0 relaţia între cele două variabile este "negativă" adică o creştere a lui X are în general ca şi consecinţă o diminuare a lui Y.

Regulile empirice Colton enunţate în 1974 sunt utile pentru comentarea valorilor coeficientului de corelaţie furnizând o interpretare standardizată a acestuia:

- un coeficient de corelaţie de la -0,25 la 0,25 înseamnă o corelaţie slabă sau nulă,
- un coeficient de corelaţie de la 0,25 la 0,50 (sau de la -0,25 la -0,50) înseamnă un grad de asociere acceptabil
- un coeficient de corelaţie de la 0,5 la 0,75 (sau de la -0,5 la -0,75) înseamnă o corelaţie moderată spre bună
- un coeficient de corelaţie mai mare decât 0,75 (sau mai mic decât -0,75) înseamnă o foarte bună asociere sau corelaţie

Testul de semnificaţie pentru coeficientul de corelaţie Pearson

Semnificaţia coeficientului de corelaţie Pearson poate fi evaluată dacă valoarea observată a apărut datorită întâmplării (dacă este semnificativ diferită de zero). Interpretarea probabilităţii furnizate de acest test este că datele experimentale ne permit sau nu ne permit enunţarea existenţei unei relaţii între variabilele luate în calcul.

Din punct de vedere matematic există mai multe posibilităţi de teste de evaluare a semnificaţiei coeficientului de corelaţie (test F, test Student) dar interpretarea acestor teste şi rezultatele produse sunt de cele mai multe ori identice.

Pentru o interpretare corectă coeficientul de corelaţie Pearson trebuie să fie însoţit de testul său de semnificaţie. Interpretarea acestora considerate împreună este sintetizată în tabelul 11.2

Tabelul 11.2 Interpretarea coeficientului de corelaţie conform regulilor lui Colton şi a testului de semnificaţie

Valoarea r	p > 0,05	p < 0,05
-0,25 la 0,25	Nu are semnificatie statistică	Corelaţie slabă sau nulă
0,25 la 0,50 (-0,25 la -0,50)	Nu are semnificatie statistică	Grad de asociere acceptabil
0,5 la 0,75 (-0,5 la -0,75)	Nu are semnificatie statistică	Corelaţie moderată spre bună
> 0,75 (< -0,75)	Nu are semnificatie statistică	Foarte bună asociere sau corelaţie
r < -1; r > 1	Eroare	Eroare

9.2.4 Corelarea rangurilor: coeficientul de corelaţie Spearman

În măsura în care între diferitele grupuri rezultate din utilizarea variabilelor ca şi elemente de grupare nu rezultă semnificaţie statistică (printr-o comparaţie de tip ANOVA) se poate utiliza coeficientul Spearman.

Coeficientul de corelaţie Spearman, notat r_s, este analogul nonparametric al coeficientul de corelaţie Pearson, calculat pentru a fi utilizat în special cu date ordinale.

$$r_s = \frac{6\sum(X-Y)^2}{n(n^2-1)}$$

unde n este numărul de perechi de variabile.

Testul de semnificaţie pentru coeficientul de corelaţie Spearman

Semnificaţia coeficientului de corelaţie Spearman poate fi evaluată dacă valoarea observată a apărut datorită întâmplării (dacă este semnificativ diferită de zero).

9.2.5 Coeficientul de determinare

Coeficientul de determinare este pătratul coeficientului de corelaţie r, adică $d = r^2$. Prin definiţie, coeficientul de determinare reprezintă partea din variaţia totală a lui Y explicată prin relaţia liniară existentă între X şi Y.

În caz extrem, dacă toate punctele se află pe o dreaptă care nu e paralelă cu axa Ox (adică nu are panta nulă), orice variaţie a lui Y este exprimată prin relaţia liniară şi coeficientul de determinare este 1. Din contră, dacă X şi Y sunt independente, adică între

cele două variabile nu există o relaţie liniară, coeficientul de determinare va fi zero.

Acest coeficient, în procente (adică înmulţit cu 100) exprimă procentajul în care variaţia lui Y este dată prin relaţia liniară între cele două variabile.

9.3 Alegerea corectă a metodei de corelație

În figura 9.3 vă prezentăm un algoritm de alegere corectă a metodei de corelație/regresie în funcție de tipul datelor pentru comparațiile între două variabile.

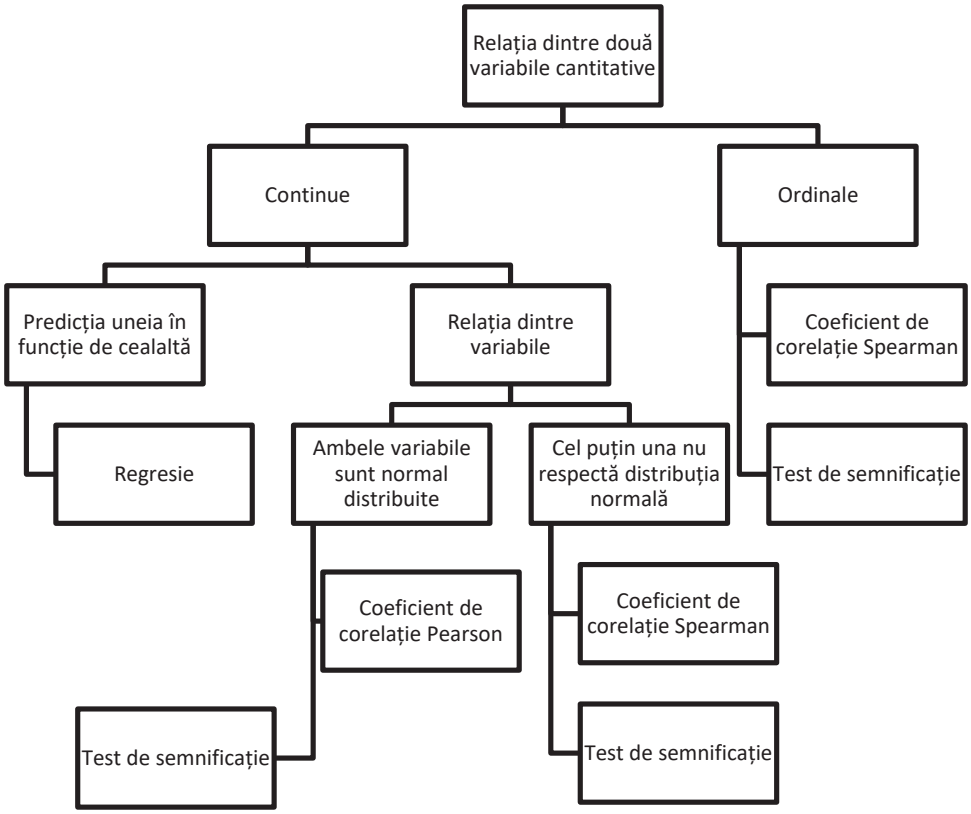


Figura 9.3. Relaţia dintre două variabile cantitative

9.4 Relaţiile neliniare, factorii de cauzalitate

9.4.1 Relațiile neliniare

Coeficienţii de corelaţie nu pot fi interpretaţi fără graficele de dispersie. Astfel, dacă dispersia punctelor în jurul dreptei de regresie nu este oarecare şi ecarturile au o anumită distribuţie sistematică se poate presupune din forma diagramei de dispersie existenţa unei relaţii neliniare între X şi Y. În acest caz un coeficient de corelaţie relativ apropiat în valoare absolută de 1 nu implică în mod necesar existenţa unei relaţii liniare între variabile, ci doar faptul că norul de puncte este aproximat destul de bine printr-o dreaptă.

De asemenea, un coeficient de corelaţie destul de mic în valoare absolută nu implică în mod necesar absenţa relaţiei între variabile, ci doar faptul că dreapta de regresie în porţiunea în care se află norul de puncte se abate destul de mult de la curba neliniară pe care "se află" punctele, cauzând în acest mod o variaţie reziduală mare.

Un coeficient de corelaţie mare nu înseamnă întotdeauna că cea mai bună relaţie între variabile este cea liniară, ca şi faptul că un coeficient de corelaţie scăzut nu implică în mod necesar absenţa oricărei relaţii între variabile, ci doar eventual a unei relaţii liniare.

Această relaţie ar putea fi neliniară, tipul ei putând fi intuit din diagrama de dispersie.

Testarea în aceste cazuri a mai multe modele matematice de corelaţie este necesară, iar modelul ales este cel cu coeficientul de corelaţie maxim, interpretarea într-un astfel de caz este că modelul aproximează cel mai bine fenomenul observat pe toată lungimea evoluției sale.

9.4.2 Factorii de cauzalitate

De cele mai multe ori atunci când se stabileşte corelaţia între două variabile se pate evidenția logic că variaţia uneia este cauza variaţiei celeilalte. Această relaţie de la cauză la efect nu trebuie acceptată fără o analiză suplimentară sau experimente complementare, pentru a trage concluzii corecte.

Înainte de a invoca o relaţie de cauzalitate, trebuie verificat dacă pentru cele două variabile studiate nu există o a treia care determină un efect cofondator (le determină pe cele două), fapt ce le face să mimeze o relaţie cauzală.

De exemplu, dacă există o corelaţie pozitivă între venitul unei persoane şi nivelul colesterolului, nu se poate trage concluzia că creşterea venitului determină creşterea colesterolului. Astfel, de exemplu, dacă se consideră şi vârsta, ca a treia variabilă, se poate constata că există o corelaţie pozitivă a ei atât cu venitul cât şi cu nivelul colesterolului.

Valoarea coeficientului de corelație este strict interpretabilă din punct de vedere matematic și este interpretată după analiza de cauzalitate. Analize suplimentare sau corelări ştiinţifice conexe sunt necesare pentru a trage concluzii cauzale în baza mărimii

coeficientului de corelaţie.

Relația de cauzalitate premerge existența corelației matematice și argumentele de plauzibilitate științifică sunt primele care trebuiesc luate în calcul. Autorii recomandă calculul coeficienților de corelație numai între variabilele pentru care această relație a fost demonstrată în prealabil sau este justificată de natura studiului.

9.5 Regresia liniară simplă

Dacă coeficientul de corelaţie este relativ mare în valoare absolută, poate fi util ca norul de puncte să fie substituit printr-o dreaptă de regresie a cărei ecuaţie ne permite prezicerea (exprimarea) valorilor uneia dintre variabile în funcţie de valorile celeilalte.

Aproximarea relației liniare dintre două variabile se poate face prin trei modele matematice care vor să minimizeze abaterile: dreapta Y(X) când se ţine seama de abaterile pe ordonată, dreapta x(Y) pentru abaterile pe abscisă și dreapta celor mai mici dreptunghiuri pentru abaterile pe ordonată şi abscisă considerate simultan. Toate modelele încercă să minimizeze eroarea de aproximare a norului de puncte printr-o dreaptă.

9.5.1 Aproximarea prin dreapta Y(X)

Se defineşte o dreaptă de regresie a variabilei Y în funcţie de variabila X printr-o ecuaţie de forma:

$$Y=a + bX.$$

Coeficienţii a şi b sunt calculaţi astfel încât suma pătratelor abaterilor valorilor estimate $\hat{Y}_i = a + bX_i$ de la valorile observate Y_i să fie minimă (metoda celor mai mici pătrate). Adică se caută a şi b (coeficienţii dreptei de regresie), astfel încât să se minimizeze suma pătratelor abaterilor de la dreapta de regresie, adică :

$$\min_{a,b\in R} \sum_{i=1}^{n} (a + bX_i - Y_i)^2$$

Se poate demonstra în acest caz că valorile lui a şi b pentru care este atins minimul sumei precedente sunt date prin formulele:

$$b = \frac{COV(X,Y)}{S_X} \quad a = \overline{Y} - b \cdot \overline{X},$$

b fiind coeficientul unghiular al dreptei de regresie (panta), iar ***a*** fiind ordonata la origine care se calculează prin formula precedentă ţinând seama de faptul că dreapta de regresie trece prin punctul de coordonate ($\overline{X}, \overline{Y}$).

9.5.2 Aproximarea prin dreapta X(Y)

Dreapta de regresie a variabilei X în funcţie de variabila Y are ecuaţia:

X =c + dY

şi se determină la fel cu dreapta de regresie Y(X), utilizând principiul celor mai mici pătrate, adică astfel încât:

$$\min_{c,d\in R}\sum_{i=1}^{n}(c+d\ Y_i - X_i)^2$$

Se obţine:

$$d=\frac{COV(X,Y)}{S_Y}\quad c=\overline{Y}-d\cdot\overline{X}$$

9.5.3 Dreapta celor mai mici dreptunghiuri

Dreapta de regresie a celor mai mici dreptunghiuri este o dreaptă de ecuaţie:

y=e+fx.

Notând cu

$$\hat{Y}_i = e+f\,X_i, \hat{X}_i=\frac{Y_i-e}{f}$$, i=1,2,...,n,

se determină e şi f astfel încât suma:

$$\sum_{i=1}^{n}\left|(X_i-\hat{X}_i)(Y_i-\hat{Y}_i)\right|$$

să fie minimă (după e şi f în R).

Valorile lui e şi f pentru care minimul este atins sunt următoarele:

$$f=sign(SPE)\cdot\frac{S_Y}{S_X}\quad e=\overline{Y}-f\overline{X}\ .$$

9.5.4 Varianţa reziduală

Varianţa reziduală asociată unei dreptei de regresie Y(X) este egală cu media aritmetică a pătratelor abaterilor reziduale punctuale (abaterilor punctelor diagramei de dispersie de la dreapta de regresie), adică

$$S_R^2=\frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2,$$

unde $\hat{Y}_i = a + bX_i$, i=1,2,…,n.

Varianţa reziduală S_R^2 reprezintă o măsură a dispersiei norului de puncte în jurul dreptei de regresie.

În general, cele trei drepte de regresie nu coincid, iar dreapta celor mai mici dreptunghiuri este cuprinsă între unghiul ascuţit al celorlalte două drepte de regresie. În funcţie de situaţia concretă studiată, îi revine utilizatorului sarcina de a alege pe cea mai potrivită pentru a descrie relaţia dintre cele două variabile. De exemplu, în cazul în care X poate fi considerată ca şi cauză a variabilei Y, atunci dreapta de regresie Y(X) este de regulă cea mai potrivită pentru a fi aleasă.

9.5.5 Dreptele de regresie şi coeficientul de corelaţie

Poziţia relativă a celor trei drepte de regresie, divergenţa lor, poate fi pusă în relaţie cu dispersia punctelor şi deci cu coeficientul de corelaţie.

Astfel, dacă |r|=1, atunci cele trei drepte de regresie coincid. Din contră, cu cât |r| este mai mic, cu atât cele trei drepte se distanţează una de alta, unghiul dintre Y(X) şi X(Y) crescând odată cu scăderea lui |r|.

La limită, când r=0, Y(X) şi X(Y) sunt perpendiculare. Y(X) e paralelă cu axa absciselor, iar X(Y) cu cea a ordonatelor.

Dreapta de regresie a celor mai mici dreptunghiuri, situată întotdeauna în unghiul dintre cele două, are următoarele caracteristici:

- Dacă varianţa lui Y este superioară celei a lui X, dreapta celor mai mici dreptunghiuri (YMR) va fi mai aproape de Y(X) iar în caz contrar de X(Y).
- Dacă variaţiile lui X şi Y sunt egale , YMR va fi bisectoarea unghiului format de dreptele Y(X) şi X(Y).

9.6 Utilizarea funcţiilor de regresie

Funcţiile de regresie sunt adesea utilizate pentru prezicerea (exprimarea) unei valori a lui Y pentru o valoare a lui X şi invers.

Să presupunem că funcţia de regresie e de tipul Y(X). Când se determină valoarea funcţiei (adică a lui Y), pentru un X cuprins ín intervalul $[X_{min}, X_{max}]$, atunci se efectuează o operaţie de interpolare, iar când X se află în afara intervalului se spune că este vorba de o extrapolare.

Acest gen de utilizare este frecvent întâlnită în probleme de etalonare. Se măsoară cele două variabile pentru o serie de valori cunoscute, apoi pe baza dreptei de regresie se prezic valorile necunoscute.

Cu atât mai mare este exactitatea prezicerii Y = a +bX cu cât X ales este mai aproape

de valoarea medie $\overline{X}$. La distanţă mare de medie predicţia utilizând modelul furnizat de regresia liniară devine riscantă.

9.7 Alte modele de corelație și regresie

9.7.1 Regresia liniară multiplă

Pentru modelarea fenomenelor naturale în cazul mai complex al prezenţei a mai mult de două variabile una dintre variabile este considerată variabilă dependentă, iar celelalte se consideră independente, prima exprimându-se în funcţie de celelalte.

Astfel fiind date variabilele:

$$X_i: X_{i1},...,X_{in} \ , i=1,2,...,m$$

$$Y: \ Y_1,..., Y_n$$

se caută o relaţie de forma:

$$Y = a + b_1X_1+...+b_mX_m,$$

unde coeficienţii a şi b_i (i=1,...,m) se determină astfel încât să minimizeze expresia:

$$\sum_{i=1}^{n}(Y_i-(a+b_1X_{1i}+...+b_mX_{mi}))^2.$$

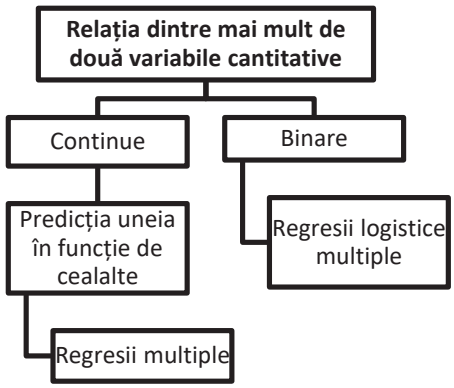


Figura 9.4. Relația dintre mai mult de două variabile cantitative

Regresia liniară multimensională reprezintă un model matematic al legăturii dintre variabila dependentă şi variabilele independente care în cazul în care varianţa reziduală asociată este mică poate substitui norul de puncte în spaţiul m+1 dimensional asociat tabelului de date. Coeficientul b_i al regresiei arată "intensitatea" influenţei variabilei X_i la variaţia variabilei dependente Y ca şi sensul acestei influenţe (prin semnul lui b_i).

Relațiile liniare multidimensionale trebui separate în analiza statistică de cele în care rezultatul este o variabilă dicotomială care sunt rezolvate prin regresia logistică discutată în capitolul 9.6.3 (vezi figura 9.4).

9.7.2 Regresii neliniare

Teoria regresiilor este un instrument eficient şi frecvent utilizat în modelarea relației dintre variabile chiar și când între acestea nu putem descrie o relației liniară. Dependența dintre variabile este asumată printr-o funcție oarecare: una dintre variabile este considerată variabilă dependentă iar celelalte se consideră independente, prima exprimându-se în funcţie de celelalte.

De cele mai multe ori funcțiile alese depind de programul statistic folosit, de forma observată pe graficul de dispersie și de aplicațiile descrise în literatura de specialitate. Putem clasifica aceste tipuri de regresie astfel:

- Regresii polinomiale
- Regresii gausiene
- Regresii sigmoidale
- Regresii hiperbolice
- Regresii exponentiale
- Regresii logaritmice
- Regresii tridimensionale
- Etc.

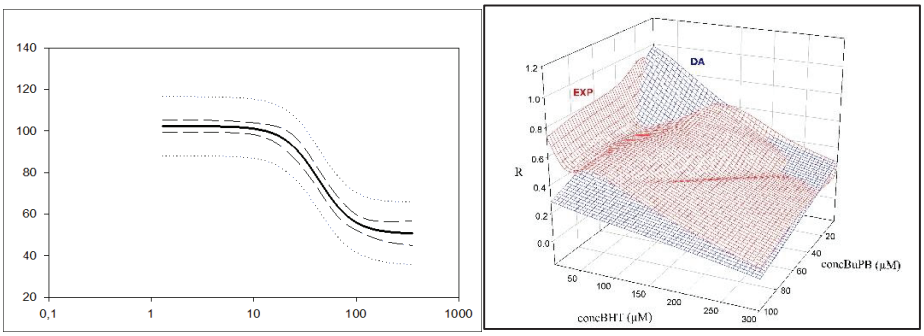


Figura 9.5. Exemple de regresii: sigmoidală (stânga) și tridimensională (dreapta)

Fiecare tip de ecuație are aplicațiile ei și alegerea corectă determină o mai bună aproximare a fenomenului. În general se recomandă utilizarea regresiei care minimizează eroarea aproximării (cea cu cel mai mare coeficient de corelație). De exemplu dinamica relațiilor farmacologice sau toxicologice doză-efect este aproximată de o funcție sigmoidă,

regresia sigmoidală permițând calcularea matematică atât a concentrațiilor substanței la care nu se obține efect, a concentrațiilor substanței pentru care efectul este maxim, iar punctul de inflexiune al curbei corespunde unui important parametru farmacologic IC50/EC50 (concentrația inhibitorie/efective la care se observă jumătatea răspunsului).

9.7.3 Regresia logistică

În multe studii medicale se dorește obținerea de predicții pentru o variabilă dependentă discretă de tip binar (afectat da/nu, vindecat da/nu). Se urmărește care sunt variabilele care influențează variabila dependentă, în ce măsură contribuie acestea la predicție. Metoda matematică de regresie care aproximează o variabilă dicotomială prin una sau mai multe variabile cantitative sau calitative predictive poartă numele de regresie logistică. Pentru calculele necesare valorile posibile ale variabilei prezise vor fi notate cu 0 și cu 1.

Dacă la calculul dreptei de regresie formula utilizată era:

$$Y=a + bX.$$

care poate varia în afara domeniului asumat pentru Y, [0,1] în acest caz se face o transformare prin utilizarea logaritmului natural:

$$\ln[p/(1-p)] = a + bX$$

sau

$$[p/(1-p)] = \exp(a + bX).$$

Modelul regresiei logistice este o simplă transformare a modelului liniar. Distribuția "logistică" este o funcție sigmoidă și permite calculul simplu al probabilităților. Distribuția "logistică" are drept rezultat o estimare a probabilității în intervalul [0,1].

$$p = 1/[1 + \exp(-a - bX)]$$

Interpretarea rezultatelor acestui tip de regresie este mai puțin intuitivă, de exemplu panta dreptei „b” nu mai este interpretabilă ca rata schimbării lui Y în funcție de X, ci mai degrabă o rată a schimbării în funcție de logaritmul rației (raportului) șansei (ODDS RATIO) de modificare a lui X. Rația șansei (ODDS RATIO) se definește în acest caz ca raportul dintre șansa de a apărea evenimentul împărțită la șansa de a apărea evenimentul contrar sau ca raportul dintre șansa ca Y=1 când X=1 împărțită la șansa ca Y=1 când X=0 în cazul în care X este o variabilă calitativă. Validarea legăturii dintre predictori și variabila dependentă se face cu teste specifice de semnificație. Calculul coeficienților de regresie se face prin metoda iterativă a estemării verosimilității maxime (maximum likelihood estimation - MLE).

9.8 Exerciţii propuse

1. Pentru 9 pacienţi cu anemie aplastică se evaluează procentul de reticulocite şi numărul limfocitelor

NR	RETICULOCITE %	LIMFOCITE (nr/ mm^2)
1	3,6	1700
2	2,0	3078
3	0,3	1820
4	0,3	2706
5	0,2	2086
6	3,0	2299
7	0,0	666
8	1,0	2088
9	2,2	2013

A. Trasaţi dreapta de regresie dintre procentul reticulocitelor şi numărul limfocitelor

B. Calculaţi coeficientul de corelaţie r şi coeficientul de determinare r^2 şi interpretaţi nivelul de corelaţie între reticulocite şi limfocite.

C. Care este semnificaţia lui r?

D. Calculaţi varianţa reziduală.

2. Folosind tabelul de mai jos calculaţi:

Nr	Durata spitalizării	Vârsta	Sex
1	5	30	F
2	10	73	F
3	6	40	F
4	11	47	F
5	5	25	F
6	14	82	M
7	30	60	M
8	11	56	F
9	17	43	F
10	3	50	M
11	9	59	F
12	3	4	M
13	8	22	F
14	8	33	F
15	5	20	F
16	5	32	M
17	7	36	M
18	4	69	M
19	3	47	M
20	7	22	M

A. Determinaţi cea mai bună relaţie liniară dintre durata spitalizării şi vârstă.

B. Verificaţi semnificaţia acestei relaţii.

C. Care este coeficientul de corelaţie r dintre durata spitalizării şi vârstă? Interpretaţi.

D. Comentaţi legătura dintre coeficientul de corelaţie r şi funcţia de regresie determinată la punctul a.

3. Dacă coeficientul de determinare r^2 dintre două seturi de măsurători este de 0,7 care din următoarele afirmaţii este adevărată?

A. valoarea unei măsurători creşte cu 0,7 când cealaltă creşte cu 1
B. 49% dintre observaţii se găsesc pe dreapta de regresie
C. 70% din observaţii se găsesc pe dreapta de regresie
D. 70% din varianţa unei măsurători este datorată celeilalte
E. 49% din varianţa unei măsurători este datorată celeilalte

4. Într-un studiu s-au calculat valorile coeficientului de corelaţie între greutate şi inaltimea pacienţilor rezultând o valoare r= -0,88; probabilitatea testului de semnificatie este 0,055. Cum interpretaţi rezultatul obţinut?

5. Dacă între două variabile cantitative se stabileşte în urma calculului coeficientului de corelaţie că există o dependenţă invers proporţionala acesta poate fi:

A. -0,78
B. -1
C. 1
D. 0,78
E. 0

Bibliografie generală

1. Borenstein M, Hedges L, Rothstein H. Introduction to Meta-Analysis. 2007 [Accessed September 9, 2016]. Available from: https://www.meta-analysis.com/downloads/Meta%20Analysis%20Fixed%20vs%20Random%20effects.pdf
2. Bulmer MG. Principles of Statistics. Dover, 1979.
3. Data Visualization Checklist. In: Stephanie Evergreen & Ann K. Emery. Presenting Data Effectively, 2014.
4. Dawson B, Trapp RG Basic & Clinical Biostatistics. 4th ed. New York: Lange Medical Books-McGraw-Hill, Medical Pub. Division, 2004.
5. Dodge Y. The Concise Encyclopedia of Statistics. Springer Science, 2008.
6. Dragomirescu L. Biostatistică pentru începători. Editura Constelaţii, Bucureşti, 1998.
7. Drugan T, Colosi H, Bolboacă S, Achimaş A, Ţigan Ş. Inferenţa statistică a datelor medicale. Cluj-Napoca, Alma Mater, 2003.
8. Drugan T, Achimaş A, Ţigan Ş. Aplicaţii medicale ale statisticii. Cluj-Napoca, Editura Medicală Iuliu Haţieganu, 2010.
9. Field A. Discovering Statistics Using SPSS. 3rd ed. SAGE; 2009.
10. Hertzog MA. Considerations in determining sample size for pilot studies. Res Nurs Health. 2008;31:180-91.
11. Jevons WS. A serious fall in the value of gold ascertained, and its social effects set forth. E. Stanford: London, 1863.
12. Jost SD. Statistics and Data Analysis. Course documents. Independent Two-sample z-test. [Internet] 2016 [cited 2016 September]:1-3. Available from: URL: http://facweb.cs.depaul.edu/sjost/csc423/.
13. Ott RL, Longnecker M. An Introduction to Statistical Methods and Data Analysis, :Brooks / Cole Publishers; 2010.
14. Pearson K. On the dissection of asymmetrical frequency curves. Philos Trans Roy Soc Lond. Ser. A 1894;185:71-110.
15. Petrie A, Sabin C. Medical statistics at a glance. 3rd Ed. Oxford: Wiley-Blackwell; 2009.
16. Plackett RL. Studies in the history of probability and statistics. VII. The principle of the arithmeticmean. Biometrika 1958;45;130-5.
17. Ray R. Testing the Difference of Two Means. [Internet] 2009 [cited 2016 September]:1-4. Available from: URL: http://web02.gonzaga.edu/faculty/rayr/ma121/Section9.2.pdf.
18. Riffenburgh RH. Statistics in Medicine. 2nd ed. Elsevier Academic Press.
19. Rosner B. Fundamentals of Biostatistics. 7th edition, Brooks/Cole, Cengage Learning 2010.
20. Ţigan Ş, Achimaş A, Drugan T, Gălătuş R, Gui D. Informatică şi statistică aplicate în medicină. Editura SRIMA, 2000.