Basic Statistics

Stefanescu, Răzvan and Dumitriu, Ramona

Dunarea de Jos University Galati, Dunarea de Jos University Galati

9 September 2007

Răzvan Ştefănescu
Ramona Dumitriu

BAZELE STATISTICII

Galați, 2007

CUPRINS

Capitolul 1 - Introducere în ştiința statisticii

De regulă, o prezentare generală a unei ştiințe se desfăşoară pe trei coordonate: obiectul de studiu, conceptele de bază şi metodele utilizate. În plus, în cazul unei ştiințe aflată în plină evoluție, aşa cum este în prezent statistica, este indicată şi abordarea transformărilor pe care aceasta le-a suferit în timp.

1.1. Obiectul de studiu al ştiinței statisticii

Se consideră că obiectul de studiu al ştiinței statisticii este reprezentat de aşa-numitele *fenomene colective* – o noțiune destul de complexă. Pentru înțelegerea acesteia, vom începe prin a defini fenomenul drept o evoluție a materiei de la o stare la alta. Astfel de evoluții se află sub influența unor condiții de mediu, date de acțiunea unor diverşi factori. În raport cu modul de manifestare, pot fi delimitate două categorii de fenomene:

1. fenomene tipice;
2. fenomene colective.

1. **Fenomenele tipice** au în comun faptul că în condiții de mediu identice vor duce întotdeauna la aceleaşi rezultate. Mecanismul unui astfel de fenomen este, de regulă, destul de simplu, cu un număr redus de factori de influență. Se spune că fenomenele tipice sunt guvernate de aşa numite *legi deterministe*, în care relația cauză-efect este precisă. Cunoaşterea acestor legi şi a condițiilor de mediu permite anticiparea cu certitudine a rezultatelor fenomenelor tipice.

2. **Fenomenele colective** sunt caracterizate prin faptul că în condiții de mediu identice pot conduce la rezultate diferite. În general, mecanismul unui fenomen colectiv este relativ complex, cuprinzând un număr mare de factori. Despre aceste fenomene se spune că în loc de a fi guvernate de legi deterministe, depind în mare măsură de hazard. Din acest motiv, rezultatele unui fenomen colectiv nu pot fi anticipate decât în condiții de incertitudine.

Tabelul 1.1. Trăsături definitorii ale fenomenelor tipice şi ale fenomenelor colective

Tip de fenomene / Aspecte	Fenomene tipice	Fenomene colective
Comportament în condiții de mediu identice	- o singură formă de manifestare	- mai multe forme de manifestare
Caracteristici ale mecanismelor de desfăşurare	- mecanisme simple, cu un număr redus de factori şi cu legi deterministe	- mecanisme complexe, cu factori de influență numeroşi, în care intervine hazardul
Certitudinea sau incertitudinea asupra viitoarelor rezultate	- certitudine	- incertitudine

O sinteză a caracteristicilor care permit delimitarea dintre fenomenele tipice şi cele colective, aşa cum au fost definite anterior, este prezentată în tabelul 1.1 Totuşi, la o analiză atentă, pot fi identificate puncte slabe pentru toate cele trei criterii de departajare. Astfel, în practică este imposibil ca un fenomen să se producă de mai multe ori în condiții de mediu cu adevărat identice, ceea ce face îndoielnică separarea fenomenelor pe baza acestui criteriu. În ce priveşte aspectul mecanismului de desfăşurare a fenomenelor, se poate obiecta că, în fapt, orice fenomen este influențat, mai mult sau mai puțin, de o infinitate de factori, fiecare cu un mod de acțiune diferit, astfel încât mintea omenească nu este capabilă să înțeleagă legile care îl guvernează. În fine, criteriul certitudinii sau incertitudinii asupra viitoarelor rezultate este atacabil deoarece termenul de certitudine are mau mult sens abstract în timp ce în realitate nimic nu este cert. În concluzie, o delimitare riguroasă între fenomenele tipice şi cele colective nu este posibilă din cauza unor limite cognitive care ne împiedică să înțelegem pe deplin realitatea.

Într-o anumită măsură, putem depăşi aceste limite construind aşa - numite *modele*, care sunt reprezentări simplificate, aproximative ale realității. Atunci când construim un model al unui fenomen, luăm în considerare numai factorii a căror influență o considerăm relevantă pentru desfăşurarea fenomenului şi apoi stabilim legi pentru a exprima acțiunea acestora. În această operațiune intervin şi percepţiile noastre, bineînţeles subiective, asupra fenomenului studiat. Atunci când considerăm că acesta este apropiat de ideea de fenomen tipic, elaborăm un model în care acţiunea factorilor de influenţă este exprimată prin legi deterministe. Un astfel de model are avantajul că permite folosirea unor tehnici simple în analiza şi previziunea fenomenului studiat, însă poate da rezultate eronate atunci când impactul unor factori ce nu au fost luaţi în considerare se dovedeşte semnificativ. Pe de altă parte, atunci când considerăm că o evoluţie întruneşte în mare măsură caracteristicile unui fenomen colectiv, o putem studia elaborând un model în care acţiunea factorilor de influenţă este exprimată prin aşa-numite *legi stocastice*. Practic, printr-o astfel de lege acceptăm că fenomenul studiat poate fi influenţat, pe lângă factorii pe care i-am considerat relevanţi în cadrul modelului, şi de alţi factori, pe care, din diferite motive, nu i-am introdus în mod explicit în model. Evident, modelele cu legi stocastice induc o complexitate deosebită analizei şi previziunii fenomenelor studiate însă oferă, totodată, o imagine mai apropiată de realitate în comparaţie cu modelele cu legi deterministe.

Dintr-o perspectivă pragmatică, cea mai importantă trăsătură a fenomenelor colective este reprezentată de incertitudinea asupra mecanismelor de producere şi asupra viitoarelor rezultate. Din acest motiv, studiul acestor fenomene se concentrează, în mare măsură, asupra modalităţilor de a face faţă acestei incertitudini. Chiar dacă nu putem înţelege pe deplin mecanismul unui fenomen colectiv, studiul acestuia ne poate releva o serie de elemente esenţiale ale acestuia, pe care le-am putea folosi în enunţarea unor relaţii cauză-efect şi în previziunea rezultatelor. În acest scop, cercetarea unui fenomen colectiv poate aborda mai multe aspecte:

- caracterizarea efectelor fenomenului;
- identificarea factorilor relevanţi de influenţă asupra fenomenului şi stabilirea modului de acţiune a acestora;
- estimarea rezultatelor posibile ale fenomenului şi a şanselor de producere a acestora.

De regulă, un proces de cercetare statistică se desfăşoară în trei etape:

1. **culegerea datelor statistice**, în care se înregistrează aspecte ale fenomenului studiat;
2. **prelucrarea statistică a datelor**, în care, prin procedee specifice ştiinţei statisticii, sunt determinaţi indicatori ce caracterizează fenomenul cercetat;

3. **analiza statistică**, în care, prin interpretarea indicatorilor statistici sunt relevate trăsăturile esențiale ale fenomenului studiat, elaborându-se modele asupra desfăşurării acestuia şi previzionându-se evoluțiile viitoare (fig. 1.1.)

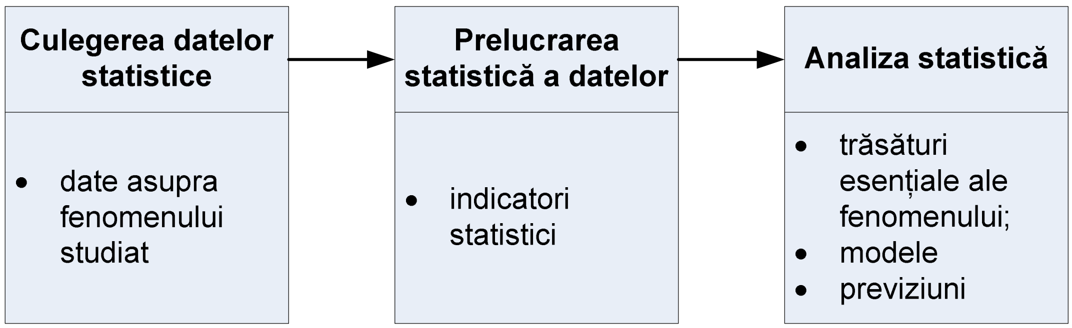


Figura 1.1. Etapele unei cercetări statistice

1.2. Concepte de bază ale ştiinței statisticii

Rigoarea care ar trebui să caracterizeze studiul din cadrul unei ştiințe este condiționată de aplicarea într-o manieră unitară a procedeelor de cercetare. Îndeplinirea acestei condiții impune ca toate conceptele utilizate în cercetare să fie definite în mod precis. În ce priveşte ştiința statisticii, cercetarea din cadrul acesteia are la bază mai multe concepte:

a) populația statistică;
b) unitatea statistică;
c) caracteristicile statistice;
d) evenimentele;
e) variabilele aleatoare;
f) funcţii probabilistice.

a) **Populația statistică** (numită şi *colectivitate statistică*) este reprezentată de o mulțime de elemente studiate pentru a se cerceta starea la un moment dat sau evoluția în timp a unuia sau mai multor fenomene. Populațiile statistice pot îmbrăca diferite forme, în funcție de scopurile şi modalitățile de cercetare a fenomenelor colective. Atunci când se studiază starea unui fenomen la un moment dat, elementele populației statistice reflectă manifestarea din acel moment a fenomenului (de exemplu, dacă se analizează salariul mediu, într-o anumită lună, pentru o ramură a economiei naționale, populația statistică este formată din ansamblul salariaților care lucrează, în luna respectivă, în acea ramură). În schimb, dacă se cercetează evoluția în timp a unui fenomen, elementele populației statistice trebuie să reflecte dinamica manifestării fenomenului în perioada de timp studiată (de exemplu, pentru a se analiza evoluția salariului mediu dintr-o ramură a economiei naționale, pe parcursul unui an, populația statistică poate fi formată din valorile salariului mediu din acea ramură, înregistrate în cele douăsprezece luni ale anului studiat).

În anumite faze ale cercetării, o populație statistică poate fi divizată în mai multe părți, pentru fiecare dintre acestea fiind aplicate metode diferite de analiză.

b) O **unitate statistică** este o componentă a mulțimii care formează o populație statistică. În funcție de metodele şi scopurile cercetării, o unitate statistică poate corespunde unui element indivizibil al populației statistice, fiind numită, în acest caz, *unitate simplă*, sau poate consta dintr-un grup de astfel de elemente, situație în care este numită *unitate compusă*. De exemplu, dacă populația statistică este reprezentată de ansamblul studenților de la o anumită specializare, pot fi stabilite unități simple, fiecare dintre acestea corespunzând unui student, sau pot fi definite unități compuse, constând în grupe, ani de studiu etc.

c) **Caracteristicile statistice** reprezintă însuşirile prin care sunt descrise, în cadrul unei cercetări, unitățile statistice. În raport cu modul de descriere, pot fi delimitate două tipuri de caracteristici statistice:

- caracteristici calitative, care descriu unitățile statistice prin cuvinte;
- caracteristici cantitative, care descriu unitățile statistice prin numere.

d) În cadrul ştiinței statisticii, un **eveniment** este un rezultat posibil sau o combinație de elemente posibile, ale unui fenomen studiat. Acest concept are implicații directe în cadrul previziunilor asupra evoluțiilor viitoare ale fenomenelor. În cazul unui fenomen colectiv, care are mai multe rezultate posibile, previziunile se fac sub forma unor mulțimi de evenimente, numite *câmpuri*. În funcție de metodele utilizate în cadrul previziunii, elementele unui câmp de evenimente pot fi prezentate sub diferite forme: valori numerice, descrieri în cuvinte etc. Un eveniment este numit *elementar* atunci când nu poate fi descompus în mai multe evenimente, şi *compus*, atunci când reprezintă un ansamblu de evenimente elementare. De exemplu, dacă se aruncă un zar, realizarea unuia dintre cele şase numere posibile poate fi considerată drept un eveniment elementar. Prin reuniunea unora dintre acestea pot fi constituite evenimente compuse: obţinerea unui număr par, a unui număr mai mic decât patru etc. Relaţiile dintre evenimente, care sunt foarte importante din perspectiva aprecierii şanselor de producere a acestora, pot fi studiate prin operaţiuni specifice teoriei mulţimilor: reuniuni, intersecţii etc. Două evenimente se numesc *mutual exclusive* atunci când este imposibilă realizarea lor simultană (altfel spus, când intersecţia lor este mulţimea vidă). Astfel, în cazul aruncării unui zar evenimentul de obţinere a numărului doi este mutual exclusiv cu evenimentul de obţinere a unui număr impar însă nu se află în aceeaşi relaţie cu evenimentul de obţinere a unui număr mai mic decât trei.

Un câmp de evenimente este numit *complet* atunci când elementele sale conţin toate rezultatele posibile ale fenomenului studiat. Utilizarea unui câmp complet de evenimente, care este o condiţie necesară pentru o previziune riguroasă, este însă adeseori foarte dificil de realizat în practică.

e) O **variabilă aleatoare** este o aplicaţie prin care fiecărui element al unui câmp de evenimente îi este asociată o valoare numerică, ceea ce facilitează cuantificarea efectelor fenomenului studiat. În funcţie de modul în care sunt atribuite valorile numerice, se pot delimita două tipuri de variabile aleatoare:

e_1) variabile aleatoare discrete;

e_2) variabile aleatoare continue.

e_1) Valorile numerice ale unei **variabile aleatoare discrete**, care pot fi finite sau infinite, sunt atribuite evenimentelor în mod discontinuu, în salturi. De exemplu, în cazul aruncării unui zar, sunt atribuite evenimentelor cele şase numere posibile, nu şi valorile intermediare dintre acestea.

e_2) La o **variabilă aleatoare de tip continuu** sunt atribuite evenimentelor absolut toate valorile numerice de pe un interval de variaţie. În acest caz, evident, valorile numerice sunt în mod obligatoriu infinite. De exemplu, dacă se studiază cantitatea de precipitaţii care va surveni în cursul unui an, poate fi luat în calcul un interval de variaţie care să cuprindă un număr infinit de valori numerice.

Alegerea între variabilele de tip discret sau continuu pentru a cuantifica efectele unui fenomen se face în funcţie de metodele de cercetare statistică utilizate. Uneori, cu toate că rezultatele posibile ale unui fenomen ar acoperi absolut toate valorile numerice ale unui interval de variație, se preferă, mai ales atunci când măsurătorile nu au o precizie prea mare, atribuirea unor valori în salturi pentru elementele câmpului de evenimente. Alteori, deşi evenimentele ar putea fi descrise prin numere întregi, se preferă să se opereze cu intervale de valori numerice.

f) **Funcțiile probabilistice** sunt utilizate în scopul cuantificării şanselor de apariție a rezultatelor posibile ale unui fenomen. O probabilitate poate fi definită drept o descriere cantitativă, printr-un număr mai mare sau egal decât zero şi mai mic sau egal decât 1, a şanselor de producere a unui eveniment. O funcție probabilistică este o aplicație prin care este asociată câte o probabilitate pentru fiecare element al unui câmp de evenimente. În general, funcțiile probabilistice sunt stabilite asupra unor variabile aleatoare, ceea ce facilitează atribuirea de probabilități.

1.3. Trăsături ale tehnicilor statistice

Incertitudinea inerentă în cazul fenomenelor colective, predispune adeseori la interpretări subiective. Pentru a contracara acest neajuns, tehnicile statistice au la bază, în general, algoritme destul de stricte, cu reguli precise care nu lasă prea multă libertate de acțiune pentru cei care le utilizează. Ca o consecință, cercetările statistice solicită, de regulă, o atenție deosebită pentru amănunte. Această calitate, care îi face uneori pe statisticieni să pară cam prea pedanți şi cam prea birocrați, poate conferi, totuşi, o rigoare deosebită unui demers ştiințific, chiar şi în condiții de incertitudine.

Există mai multe criterii de clasificare a tehnicilor statistice:

a) criteriul etapei de cercetare statistică;

b) criteriul modului de studiere în timp a fenomenelor;

c) criteriul cuprinderii populației statistice în cadrul cercetării.

a) În raport cu **criteriul etapei de cercetare**, tehnicile statistice pot fi grupate în trei categorii:

a_1) tehnici de culegere a datelor;

a_2) tehnici de prelucrare statistică a datelor;

a_3) tehnici de analiză statistică.

a_1) **Tehnicile de culegere a datelor** sunt folosite pentru a înregistra aspecte ce caracterizează fenomenele studiate. Această categorie cuprinde variate procedee: interviuri, recensăminte, experimente etc.

a_2) **Tehnicile de prelucrare statistică a datelor** sunt folosite pentru a se obține informații utilizate în cadrul analizei. Cele mai multe dintre aceste tehnici sunt cantitative, astfel încât informațiile rezultate îmbracă, de regulă, forma unor mărimi numerice.

În raport cu aspectele vizate se pot delimita, în cadrul categoriei tehnicilor de prelucrare statistică a datelor, mai multe subcategorii:

- *tehnici statistice de sistematizare a datelor*, care facilitează prelucrările statistice ulterioare şi care pot oferi informații asupra amplorii şi intensității fenomenelor studiate;
- *tehnici de reprezentare grafică*, pe baza cărora pot fi sesizate unele aspecte esențiale ale unui fenomen cercetat: sensul evoluției în timp, forma funcției probabilistice asociate, legăturile cu alte fenomene etc.;
- *tehnici de determinare a valorilor tipice*, prin care sunt determinate mărimi reprezentative pentru o populație statistică;
- *tehnici de calcul al unor indicatori ai dispersiei, asimetriei sau boltirii*, pe baza cărora se poate aprecia măsura în care valorile tipice sunt reprezentative pentru ansamblul populației statistice;
- *tehnici de calcul al unor indicatori ai legăturilor dintre fenomene*, care pot oferi informații asupra factorilor care influențează un fenomen cercetat;
- *tehnici de determinare a unor numere indice*, care facilitează comparațiile între unitățile statistice etc.

a_3) **Tehnicile de analiză statistică** sunt utilizate pentru interpretarea informațiilor obținute prin prelucrarea datelor statistice, în vederea înțelegerii fenomenelor studiate. În această categorie sunt incluse procedee de stabilire a legilor de dependență a unui fenomen cercetat față de factorii semnificativi de influență, de prognoză a evoluțiilor viitoare a fenomenelor colective etc.

b) Pe baza **criteriului modului de studiere în timp a fenomenelor** pot fi delimitate două categorii de tehnici statistice:

b_1) tehnici statistice de analiză statică;

b_2) tehnici statistice de analiză dinamică.

b_1) **Tehnicile statistice de analiză statică** sunt utilizate pentru a studia starea unui fenomen la un moment dat. Adeseori, analiza statică este asemănată cu o fotografiere, care surprinde toate aspectele obiectului fotografiat fără a putea releva însă transformările pe care acesta le-a suferit în timp.

b_2) **Tehnicile statistice de analiză dinamică** sunt folosite pentru a studia evoluția în timp a unui fenomen. Un procedeu de analiză dinamică poate fi asemănat unei filmări care surprinde transformările în timp.

c) După **criteriul cuprinderii populației statistice în cadrul cercetării**, tehnicile statistice pot fi împărțite în două categorii:

c_1) tehnici de cercetare statistică totală

c_2) tehnici de cercetare statistică parțială.

c_1) **Tehnicile de cercetare statistică totală** presupun utilizarea unor date provenite de la toate unitățile unei populații statistice. Astfel de procedee, care sunt avantajoase din perspectiva rigorii demersului științific, se pot dovedi, în cazul unui volum mare al populației statistice, deosebit de costisitoare şi de cronofage.

c_2) **Tehnicile de cercetare statistică parțială** presupun utilizarea unor date care provin numai de la anumite unități ale populației statistice şi extinderea ulterioară a informațiilor astfel obținute la nivelul ansamblului populației. De exemplu, sondajele de opinie prin care se estimează convingerile şi opțiunile unei populații la un moment dat, utilizează datele obținute pentru o parte infimă a ansamblului colectivității. Cu toate că, în principiu, conferă unui demers științific o rigoare mai mică decât cercetarea statistică totală, cercetarea statistică parțială îi este adeseori preferată acesteia datorită operativității şi costurilor mici pe care le implică.

Fiecare dintre categoriile de tehnici statistice prezentate anterior cuprinde o gamă variată de procedee, ceea ce ridică, în cadrul unei cercetări statistice, problema alegerii metodelor optime. Într-o astfel de decizie sunt luate în considerare, de regulă, trei aspecte:

- costurile implicate;
- perioada de timp necesară pentru aplicare;
- acuratețea rezultatelor.

Alegerea este, adeseori, destul de dificilă în condițiile în care, în general cele trei criterii de decizie se află în relații concurențiale (de exemplu, un procedeu care asigură o acuratețe mare a rezultatelor implică, de regulă, un cost ridicat şi o perioadă lungă de timp pentru implementare).

În principiu, faptul că tehnicile statistice sunt bazate în mare măsură pe tehnici cantitative asigură o anumită obiectivitate studiului fenomenelor colective. Totuşi, aceasta nu înseamnă că cercetarea statistică nu este expusă greşelilor, arbitrariului sau subiectivităţii. Uneori, tehnicile statistice sunt utilizate greşit sau sunt chiar manipulate pentru a se ajunge la rezultatele dorite de utilizatori, situaţii care au condus la comentarii ironice de genul *„statistica este ştiinţa prin care se poate demonstra orice”* sau, în cuvintele lui Mark Twain: *„Există trei tipuri de minciuni: simple, gogonate şi statistici”*.

1.4. Evoluţia în timp a ştiinţei statisticii

Termenul de statistică (în limba germană *statistik*, provenit din italianul *statisto*) a fost folosit pentru prima oară de către omul de ştiinţă Gottfried Achenwall, la jumătatea secolului XVIII, cu sensul de ansamblu de informaţii organizate într-o formă care să faciliteze analiza. Totuşi, forme incipiente de cercetare statistică au fost folosite cu mult înainte de consacrarea termenului de statistică. Astfel, istoricii au demonstrat că acum mai bine de 4000 de ani în Egiptul Antic se practicau inventarieri ale averii statului. Tot din perioada antică, Vechiul Testament ne oferă informaţii despre un recensământ organizat în Israel de regele David, iar alte date furnizate de istorici descriu înregistrări statistice practicate în China Antică şi Roma Antică. Aceste înregistrări statistice aveau un scop pragmatic: buna conducere a statului necesita informaţii asupra resurselor materiale, financiare şi umane disponibile. Cu timpul, tehnicile statistice utilizate în administraţia publică s-au diversificat, devenind totodată mai sofisticate.

Încă din Evul Mediu s-au manifestat preocupări pentru folosirea tehnicilor statistice nu doar în administraţia publică ci şi în cadrul cercetărilor ştiinţifice. Începutul a fost făcut cu unele aspecte demografice (natalitatea, mortalitatea etc.) pe care oameni de ştiinţă din acea perioadă (John Grund, Halley ş.a.) le-au studiat prin procedee statistice.

Au urmat alte tentative de studiere a unor fenomene colective, care au relevat necesitatea unor tehnici care să faciliteze o cercetare riguroasă, chiar şi în condiţii de incertitudine. Astfel de procedee au fost concepute în cadrul ştiinţei matematicii, ceea ce a făcut ca multă vreme statistica să fie considerată drept o componentă a acesteia. De altfel, delimitarea dintre cele două ştiinţe a rămas până astăzi ambiguă. Există matematicieni care privesc statistica drept un capitol al matematicii după cum există statisticieni care consideră că matematica este un capitol al statisticii.

Treptat, procedeele statistice au început să fie folosite frecvent în cadrul cercetărilor din diferite ştiinţe: fizica, astronomia, chimia, biologia etc., ajungându-se ca până la urmă să fie considerate indispensabile pentru un demers ştiinţific riguros. Iniţial erau folosite doar procedee ţinând de ceea ce numim astăzi *statistica descriptivă*, care au ca obiect descrierea fenomenelor: reprezentări grafice, determinarea valorilor tipice, aprecierea dispersie şi a asimetriei etc. Cu timpul însă, s-au dezvoltat şi tehnici ale aşa-numitei *statistici inductive*, care are ca obiect generalizarea rezultatelor unor cercetări parţiale.

În ultimele decenii, pe lângă administraţia publică şi cercetările ştiinţifice, procedeele statistice au început să fie folosite pe scară largă în conducerea unor organizaţii din domeniul economic şi social. Tot în ultimele decenii, s-a iniţiat combinarea tehnicilor statistice cu procedee bazate pe inteligenţa artificială, câştigându-se astfel o operativitate deosebită.

Aplicarea metodelor statistice în diferite domenii de activitate a condus la o diversificare a tehnicilor, acestea trebuind să fie adaptate condiţiilor în care sunt utilizate. Această situaţie a condus la o diferenţiere în cadrul ştiinţei statisticii a două componente: statistica matematică şi statistica aplicată. **Statistica matematică** are ca obiect formularea, pe baza principiilor ştiinţei matematicii, a unor tehnici de cercetare statistică. În ce priveşte **statistica aplicată**, aceasta are ca obiect adaptarea tehnicilor statisticii matematice la condiţiile concrete ale domeniilor în care sunt utilizate. În cadrul statisticii aplicate se delimitează prin particularităţile procedeelor, mai multe ramuri: statistica economică, statistica managerială, statistica fizicii, statistica biologiei, statistica chimiei, statistica sociologică, statistica ingineriei, statistica medicinii etc.

Capitolul 2 - Culegerea datelor statistice

2.1. Coordonatele culegerii datelor statistice

Culegerea datelor, care reprezintă începutul unui demers de cercetare statistică, are un rol determinant asupra calității acestuia. Indiferent de rigoarea tehnicilor utilizate în etapele ulterioare, dacă datele colectate sunt eronate, rezultatele cercetării vor fi, de asemenea eronate, situație cunoscută sub denumirea de „fenomenul GIGO” *(garbage in – garbage out).*

Este necesar ca operațiunile de culegere a datelor să fie circumscrise unor caracteristici ale cercetării statistice din care fac parte: scopul acesteia, caracterul regulat sau extraordinar, domeniul de aplicare, acuratețea solicitată etc. Aceste aspecte sunt luate în considerare atunci când se stabilesc principalii parametri ai culegerii datelor: populația statistică, sursele datelor, caracteristicile statistice la care se vor raporta datele, instrumentele de colectare a datelor ş.a.m.d.

Într-o cercetare statistică pot fi utilizate două tipuri de date:

- *date primare*, culese special pentru acel demers;
- *date secundare*, care au fost obținute anterior, pentru alte scopuri.

În general, procurarea datelor secundare este mult mai puțin costisitoare şi consumă mult mai puțin timp în comparație cu obținerea datelor primare. Adeseori, datele secundare sunt preluate din comunicate oficiale ale unor instituții publice. De exemplu, în cercetările statistice asupra activității unei firme pot fi folosite date asupra unor indicatori macroeconomici: rata inflației, rata şomajului, salariul mediu, cursurile valutare ş.a.m.d., care au un caracter public. În ciuda avantajelor incontestabile pe care le oferă, utilizarea datelor secundare este, totuşi, limitată, acestea având, de regulă, un rol complementar. Din perspectiva utilizării populației statistice pot fi delimitate două forme de culegere a datelor statistice:

a) culegerea datelor prin recensăminte;

b) culegerea datelor prin sondaje.

a) **Culegerea datelor prin recensăminte** presupune investigarea tuturor unităților populației statistice prin care se studiază un fenomen. Atunci când numărul de unități statistice este foarte mare (aşa cum este, de exemplu, cazul recensămintelor asupra populației umane) un recensământ necesită folosirea unui volum mare de personal, nu întotdeauna pe deplin calificat, ceea ce implică probleme organizatorice importante, costuri ridicate precum şi posibilitatea unor erori de înregistrare semnificative. Din aceste motive, astfel de recensăminte se efectuează destul de rar (de exemplu, recensămintele asupra populației se utilizează, de regulă, o dată la zece ani). În schimb, atunci când numărul de unități statistice este relativ redus, recensământul poate fi un mijloc destul de simplu, de ieftin şi de precis de culegere a datelor statistice (de exemplu, pentru o firmă cu un număr mic de clienți nu este prea dificil să obțină date despre toți aceştia).

b) **Culegerea datelor prin sondaj** presupune ca în loc de a se colecta date de la toate unitățile populației statistice să fie investigată doar o parte a acesteia, numită *eşantion*, urmând ca în cadrul cercetării statistice informațiile obținute pe baza datelor de la eşantion să fie extinse asupra întregii populații statistice. În comparație cu recensămintele, sondajele necesită, de regulă, un volum mult mai redus de personal, ceea ce permite ca toți lucrătorii utilizați să fie calificați, şi face mai uşoară coordonarea, implicând totodată costuri mai mici şi restrângând posibilitatea erorilor de înregistrare. Totuşi, culegerea datelor prin sondaje este expusă aşa-numitelor *erori de reprezentativitate*, care derivă din posibilitatea ca eşantionul ales să nu fie suficient de reprezentativ pentru ansamblul populației statistice.

2.2. Instrumente de culegere a datelor statistice

În acest subcapitol vor fi prezentate succint patru tipuri de instrumente utilizate destul de frecvent în culegerea datelor statistice:

- chestionarea statistică;
- observația statistică;
- experimentul statistic;
- panelul statistic.

2.2.1. Chestionarea statistică

O chestionare statistică e reprezentată de un ansamblu de întrebări, cuprinse într-un aşa numit *chestionar*, adresate unor persoane cu privire la percepţiile şi reacţiile acestora faţă de un fenomen studiat. În cadrul chestionării statistice se detaşează, prin importanţă, trei aspecte:

- forma de anchetare;
- acurateţea datelor culese;
- proiectarea chestionarelor.

2.2.1.1 Forme de anchetare

Din perspectiva formelor de anchetare se pot distinge două tipuri de chestionări statistice:

a) interviuri;
b) chestionări scrise.

a) Interviurile îmbracă forma unor discuţii purtate cu persoanele anchetate de către lucrători specializaţi, numiţi *operatori de interviuri*. Principalul avantaj al interviurilor constă în faptul că permite interactivitatea dintre operatorul de interviuri şi persoana intervievată. În cadrul discuţiilor, operatorul interviurilor îi poate lămuri persoanei anchetate sensul unor întrebări dificile, o poate convinge pe aceasta să atingă subiecte mai delicate sau îi poate adresa întrebări suplimentare, neprevăzute în chestionarul stabilit iniţial, pentru a lămuri anumite aspecte.

Totuşi, în culegerea datelor prin interviuri intervin şi câteva dificultăţi semnificative:

- operatorul de interviuri trebuie să fie, în mod obligatoriu, o persoană calificată;
- adeseori interviurile consumă perioade de timp destul de lungi;
- reticenţa de a acorda interviuri pe care o manifestă unele dintre persoanele alese pentru a fi anchetate etc.

Interviurile pot fi realizate atât pe cale orală cât şi prin telefon. Cele pe cale orală sunt mai costisitoare şi consumă mai mult timp decât cele telefonice însă facilitează într-o măsură mai mare interactivitatea dintre operatorul de interviuri şi persoanele anchetate.

b) **Chestionările scrise** se materializează în distribuirea, către persoanele anchetate, a unui chestionar, cu rugămintea de a se răspunde la întrebările acestuia. Faţă de interviuri, chestionările scrise au unele avantaje incontestabile: sunt mai operative, mai uşor de organizat şi mai puţin costisitoare. În acelaşi timp însă, la această formă de chestionare lămurirea persoanelor anchetate asupra înţelesului unor întrebări este mai dificilă.

2.2.1.2. Acurateţea datelor culese prin chestionări statistice

Chestionarea statistică se diferenţiază faţă de celelalte instrumente ale culegerii de date statistice prin oportunităţile pe care le oferă pentru înţelegerea comportamentului uman în legătură cu fenomenele studiate. Totuşi, tocmai faptul că se află în relaţie directă cu comportamentul uman face ca acest instrument să se afle expus într-o măsură considerabilă erorilor.

Există mai multe surse de erori asupra datelor culese prin chestionări statistice:

- neînţelegerea sensului unora dintre întrebări;
- nesinceritatea sau refuzul de a supune adevărul la întrebări delicate pentru persoanele anchetate (apartenenţa la o minoritate religioasă sau sexuală, practica de a oferi mită, metodele manageriale aplicate etc.);
- neseriozitatea unora dintre respondenţi etc.

Pentru combaterea acestor surse de erori pot fi aplicate diferite remedii:formularea clară şi în termeni simpli a întrebărilor adresate, abordarea cu tact a subiectelor delicate, selectarea prealabilă a persoanelor anchetate etc.

2.2.1.3. Proiectarea chestionarelor statistice

Pentru ca o chestionare statistică să îşi atingă obiectivele vizate este indicată o proiectare minuţioasă a chestionarelor utilizate, în raport cu aspectele asupra cărora se doreşte colectarea datelor. Chiar şi în cazul unui interviu, unde întrebările iau adeseori naştere în mod spontan în cadrul discuţiilor, se recomandă pregătirea din timp a unor întrebări cheie. pentru proiectarea chestionarelor statistice pot fi formulate mai multe reguli, a căror respectare condiţionează rigoarea colectării datelor.

Regula nr. 1. *Întrebările trebuie formulate de pe o poziţie neutră faţă de aspectele abordate*

Această neutralitate trebuie să îngăduie persoanelor anchetate să îşi exprime opiniile fără a fi influenţate de modul de formulare a întrebărilor. De exemplu, o întrebare de genul *„Dumneavoastră doriți, aşa cum doreşte cea mai mare parte a populației României, integrarea în Uniunea Europeană?”* are dezavantajul că poate influența unele persoane să răspundă afirmativ numai ca să nu apară ca având o opinie separată față de majoritate, după cum alte persoane vor răspunde negativ tocmai pentru a se distanța de majoritate. În practică, regula formulării întrebărilor de pe o poziție neutră este încălcată uneori în mod voit, tocmai pentru a se obține anumite rezultate, aşa cum a fost, în trecut, cazul unor sondaje de opinie din România.

Regula nr. 2. *Întrebările trebuie astfel formulate încât sensul acestora să fie uşor de înțeles.*

Termenii întrebărilor trebuie stabiliți în funcție de unele caracteristici ale grupului de persoane anchetat (relația în care se află cu aspectele studiate, pregătirea profesională, vârsta etc.). În principiu, trebuie evitate formulările prea complicate sau cuvintele folosite rar.

Regula nr. 3. *Întrebările chestionarului trebuie organizate într-o succesiune logică, de la generalități la aspecte particulare*

S-a constatat că o abordare prea bruscă a unor aspecte particulare conduce adeseori la răspunsuri pripite întrucât persoanele anchetate nu au avut timp să intre în temă. Este indicat, din acest motiv, să se înceapă cu întrebări cu caracter general ajungându-se treptat la întrebări cu caracter particular. De exemplu, dacă se culeg date asupra cererii potențiale pentru sortimentele comercializate de o firmă producătoare de dulciuri se poate porni cu o întrebare de genul *„Consumați frecvent dulciuri?”*, ajungându-se abia în final la întrebări de amănunt, cum ar fi: *„Prin ce anume considerați că se diferențiază produsele firmei noastre față de alte sortimente de dulciuri?”*

Într-o anchetă, unele întrebări pot fi adresate doar anumitor categorii de respondenți, stabilite pe baza răspunsurilor de la unele întrebări anterioare. De exemplu, pe baza întrebării referitoare la consumul de dulciuri, respondenții pot fi împărțiți în două categorii: cei care consumă frecvent dulciuri şi cei care nu fac aceasta decât rar sau deloc. Pentru persoanele din prima categorie ancheta ar putea fi direcționată pentru obținerea de date asupra preferințelor în materie de dulciuri, cu întrebări de genul *„În cazul în care consumați frecvent dulciuri, care sunt sortimentele dumneavoastră preferate?”*. Pentru respondenții din a doua categorie este mai important de aflat motivele reținerii față de dulciuri, prin întrebări de genul: *„În cazul în care nu consumați frecvent dulciuri, ce anume vă face să nu le preferați?”*. Se poate observa că fiecare din cele două întrebări începe cu o condiție (numită *filtru*) folosită mai ales în cazul chestionărilor scrise pentru a indica persoanele care vor trebui să răspundă.

Regula nr. 4. *Tipul întrebărilor trebuie adaptat la circumstanțele chestionării*

În cadrul chestionarelor se disting, în funcție de forma în care trebuie date răspunsurile, două tipuri de întrebări:

- întrebări cu răspuns deschis;
- întrebări cu variante prestabilite de răspuns.

Întrebările cu răspuns deschis îl lasă pe respondent să răspundă în forma pe care o doreşte. De exemplu, în cadrul unei anchete printre studenți asupra oportunităților oferite de o facultate, o astfel de întrebare ar putea fi formulată astfel: *„De ce ați ales facultatea noastră?”*, lăsându-se celor interogați libertatea în a-şi prezenta motivele. În acest fel poate fi evitată influențarea persoanelor anchetate în formularea răspunsurilor. Totuşi, întrebările cu răspuns deschis nu sunt prea potrivite pentru respondenții care au dificultăți în a-şi exprima opiniile clar şi rapid. Un alt dezavantaj este reprezentat de faptul că răspunsurile oferite sunt destul de greu de sistematizat.

Întrebările cu variante prestabilite de răspuns oferă respondenților posibilitatea de a opta între mai multe răspunsuri posibile. De exemplu, dacă se efectuează o anchetă asupra modului în care locuitorii unei localități ar primi înființarea unui parc de distracții, s-ar putea pune mai multe întrebări cu variante prestabilite de răspuns: *„Aveți cunoştință despre un plan de înființare a unui parc de distracții în localitatea dumneavoastră? DA/NU”*, sau *„În cazul în care sunteți împotriva proiectului parcului de distracții, ce anume veți face pentru a vă susține punctul de vedere?*

- □ *voi adresa scrisori consilierilor locali;*
- □ *voi adresa scrisori parlamentarilor din județul nostru;*
- □ *voi semna o petiție împotriva construirii parcului de distracții;*
- □ *voi participa la demonstrații împotriva construirii parcului de distracții;*
- □ *voi participa la tentativele de întrerupere a construcției parcului;*

□ *nu voi face nimic concret.”*

Uneori, la astfel de întrebări se trece la scalarea răspunsurilor, pentru a se putea măsura atitudinea respondenților. De exemplu, pentru a se estima cât de mult aprobă sau dezaprobă localnicii înființarea unui parc de distracții, ar putea fi formulată următoarea întrebare: *„Credeți că parcul de distracții ar trebui construit?”*, cu următoarele răspunsuri posibile:

Agreez ideea foarte mult	Agreez ideea	Nu ştiu	Dezaprob ideea	Dezaprob foarte mult ideea

Aceste întrebări au avantajul unei sistematizări simple şi operative a răspunsurilor. Totuşi, uneori răspunsurile oferite nu reflectă exact opiniile respondenților ci sunt, mai degrabă, cele mai apropiate față de acestea dintre variantele de răspuns oferite. Aceasta se întâmplă mai ales atunci când variantele de răspuns nu sunt suficient de dezvoltate pentru a cuprinde toate nuanțele opiniilor respondenților. De exemplu, datele culese pe baza unei întrebări de genul: *„Sunteți de acord cu instalarea de baze ale armatei SUA pe teritoriul României? DA/NU”* nu oferă neapărat o imagine completă a opiniilor celor interogați. O parte dintre cei care ar da răspunsuri negative ar putea accepta, în fapt, bazele militare în anumite condiții: dreptul de veto al statului român asupra utilizării acestor baze în conflictele militare, valabilitatea jurisdicției româneşti pentru militarii americani etc.

2.2.2. Observația statistică

O observație statistică este o înregistrare a unor aspecte ale manifestării unui fenomen cercetat. Acest instrument de culegere de date statistice este utilizat destul de frecvent în variate domenii: în studiul comportamentului oamenilor sau al altor viețuitoare, în cercetări asupra activității economice, unde se înregistrează diferiți parametri: volumul producției, al vânzărilor etc., în cercetarea unor procese fizice, chimice etc. Atunci când se cercetează comportamentul oamenilor sau al altor viețuitoare se recomandă, de regulă, ca observarea statistică să se desfăşoare fără ştirea celor studiați, tocmai pentru a nu le afecta comportamentul (în acest scop se folosesc filmările cu camere ascunse, oglinzi cu vedere unilaterală ş.a.m.d.). Pentru înregistrarea unor date ce privesc parametrii tehnici ai unor fenomene pot fi folosite aparate de măsură în combinaţie cu tehnologii informaţionale. În domeniul activităţii economice, observaţiile statistice sunt circumscrise adeseori unui sistem informaţional, fiind organizate în raport cu caracteristicile acestuia.

Acurateţea datelor obţinute prin observări statistice depinde în mare măsură de modalităţile de înregistrare. La acest instrument de colectare a datelor statistice sunt relevante două categorii de erori: erorile umane şi erorile date de deficienţe tehnice.

Principalele avantaje ale observaţiilor statistice sunt reprezentate de costurile în general reduse şi de relativa simplitate a aplicării. Totuşi, utilizarea acestui instrument are şi unele limite, în special în cazul cercetării comportamentului uman unde poate oferi date asupra manifestării dar nu şi asupra motivaţiilor acestuia.

2.2.3. Experimentul statistic

Un experiment statistic constă în provocarea, în mod artificial dar în condiţii cât mai apropiate de cele naturale, a unui proces, pentru a i se putea studia manifestarea. Experimentele statistice au aplicaţii în diferite domenii: în cercetări asupra unor fenomene fizice, chimice, biologice, sociologice, psihologice ş.a.m.d., în fundamentarea unor decizii manageriale etc. Caracteristicile unui experiment statistic trebuie adaptate domeniului în care acesta trebuie aplicat. De exemplu, dacă se doreşte obţinerea de date asupra modului în care ar putea fi primit un sortiment de produs nou, acesta este distribuit, înainte de lansarea pe piaţă, unui grup de persoane ale căror reacţii vor fi studiate. Acurateţea datelor obţinute prin experimente statistice depinde în mare măsură de gradul în care condiţiile de desfăşurare a acestora sunt apropiate de condiţiile naturale. Experimentele statistice sunt indicate îndeosebi în cazul unor procese inedite, care nu pot fi studiate pe baza experienţelor din trecut. Totuşi, utilizarea acestora este adeseori destul de complexă şi de costisitoare.

2.3.4. Panelul statistic

Un panel statistic constă în interogarea periodică a unui grup de persoane cu privire la un acelaşi fenomen. Acest instrument este indicat în cercetarea unor procese pentru care este de aşteptat ca percepţiile populaţiei să se modifice substanţial în timp: politici guvernamentale, campanii promoționale, lansări de sortimente noi de produse etc.

Un panel statistic poate fi considerat drept un ansamblu de chestionări efectuate cu regularitate cu aceleaşi persoane anchetate. Totuşi, organizarea culegerii de date prin acest instrument este mult mai dificilă decât prin chestionările statistice obişnuite deoarece nu sunt prea uşor de găsit persoane dispuse să răspundă întrebărilor cu regularitate, uneori de-a lungul unor perioade lungi de timp. În plus, acuratețea datelor culese poate avea de suferit deoarece tocmai faptul că au fost alese în grupul supus chestionării poate determina schimbări în comportamentul unor persoane anchetate. De asemenea, culegerea datelor este expusă şi riscului descompletării, din diferite motive (decese, schimbarea localității de domiciliu etc.) a grupului de persoane anchetate.

Capitolul 3 - Prelucrarea primară a datelor statistice

3.1. Coordonate ale prelucrării primare a datelor statistice

După ce datele statistice au fost culese, este necesară transpunerea lor într-o formă care să faciliteze caracterizarea fenomenelor colective. Ansamblul procedeelor utilizate în acest scop poartă denumirea de prelucrare primară a datelor statistice. Rezultatele acestor operațiuni pot îmbrăca mai multe forme:

- serii statistice;
- tabele statistice;
- reprezentări grafice.

3.1.1. Seriile statistice

O serie statistică este o modalitate de organizare a unei populații statistice sub forma unui şir în care fiecărei unități îi sunt asociate valori ale uneia sau mai multor caracteristici. În studiul fenomenelor colective pot fi utilizate forme variate de serii statistice. În continuare, vom prezenta o clasificare a acestora în raport cu două criterii: numărul şi tipul caracteristicilor folosite în descrierea unităților populației.

În funcție de **numărul caracteristicilor** pot fi delimitate două tipuri de serii statistice:

- **serii unidimensionale**, în care unitățile populației statistice sunt descrise printr-o singură caracteristică;
- **serii multidimensionale** (bidimensionale atunci când se folosesc două caracteristici, tridimensionale atunci când sunt folosite trei caracteristici ş.a.m.d.), în care unitățile populației statistice sunt descrise prin mai multe caracteristici.

În raport cu **tipul caracteristicilor statistice**, se pot departaja trei categorii de serii statistice:

- **serii atributive**, în care sunt utilizate alte tipuri de caracteristici decât cele de spațiu sau de timp;
- **serii de spațiu**, în care caracteristicile utilizate descriu locul de manifestare a fenomenului studiat;
- **serii de timp**, în care caracteristicile folosite descriu evoluția în timp a fenomenului studiat.

3.1.2. Tabele statistice

Un tabel statistic este un tabel în ale cărui celule sunt înscrise valorile asociate unei serii statistice, grupate pe linii şi pe coloane în raport cu unităţile statistice şi caracteristicile folosite în descrierea fenomenului studiat.

Tabelele statistice se pot folosi atât în calculul unor mărimi cât şi în reprezentarea aspectelor definitorii ale fenomenelor colective. În ambele cazuri, exigenţele cercetării statistice impun câteva reguli în construirea tabelelor.

Tabelul 3.1. Calculul cifrei de afaceri

Nr. crt.	**Sortiment de produs**	**Preţ (p_i) [RON/buc]**	**Cantităţi vândute (q_i) [buc]**	**Cifră de afaceri aferentă sortimentului ($CA_i = p_i \times q_i$) [RON]**
(0)	**(1)**	**(2)**	**(3)**	**(4) = (2) × (3)**
1	Televizor model 1	100	100	10 000
2	Televizor model 2	50	40	2 000
3	CD player	10	80	800
4	Total	x	x	12 800
5	Simbol pentru total	x	x	$\sum_{i=1}^{n} CA_i$

Regula nr. 1: *Unităţile statistice şi caracteristicile statistice trebuie înscrise în tabele cu elemente de identificare care să le diferenţieze în mod clar.* Atunci când tabelul serveşte în determinarea unor mărimi, este indicată folosirea, pentru caracteristicile utilizate, a unor simboluri, înscrise între paranteze rotunde şi, dacă este cazul, a formulelor de calcul. De exemplu, în tabelul 3.1., care a fost utilizat pentru determinarea cifrei de afaceri a unei firme care comercializează produse electronice, unităţile statistice sunt reprezentate de sortimentele de produse iar caracteristicile statistice de preţuri şi de cantităţi vândute. În situaţia în care firma comercializează două sortimente de televizoare, acestea au trebuit diferenţiate prin precizarea modelului. La cele două caracteristici au fost precizate simbolurile, iar în ce priveşte cifra de afaceri a fiecărui sortiment, dată de produsul dintre preţ şi cantitatea vândută, a fost înscrisă şi formula de calcul. În coloana prin care s-au calculat cifrele de afaceri pentru cele trei sortimente au fost înscrise totalul acestora (care reprezintă de fapt valoarea mărimii care trebuia calculată) şi simbolul asociat acestuia.

Regula nr. 2: *Se recomandă ca unităţilor statistice să le fie asociat un număr de ordine* (numit şi *număr curent*) *care să faciliteze regăsirea datelor.* Atunci când tabelul este utilizat în determinarea unor mărimi, este indicat ca şi fiecărei caracteristici statistice să îi fie asociat un număr de ordine, înscris în paranteze rotunde, iar pentru valorile calculate în cadrul tabelului să se indice modul de calcul prin intermediul numărului de ordine. De exemplu, în tabelul 3.1., formula (4) = (2) × (3) indică faptul că pentru fiecare linie care corespunde unei unităţi statistice, valoarea din coloana nr. 4 este dată de produsul valorilor din coloanele cu numerele de ordine 2 şi 3.

Tabelul 3.2. Indicatori macroeconomici în România în perioada 1990 – 1995

Indicatori	U.M.	1990	1991	1992	1993	1994[a)]	1995 [b)]
1. Produsul intern brut	mld. lei	857,9	2203,9	6029,2	20035,7	49794,8	72248,9
2. Export FOB total [c)]	mil. USD	5775,4	4265,7	4363,4	4892,2	6151,3	7519,5
3. Import FOB total [c)]	mil. USD	9202,5	5372,0	5784,1	6020,1	6562,4	8750,0
4. Rata inflaţiei ca nivel la sfârşitul perioadei [1)]	%	37,7	222,8	199,2	295,5	61,7	27,8
5. Rata	%	-	3,0	8,2	10,4	10,9	8,9

şomajului 2)							

Sursa: Banca Națională a României, Raport Anual 1995

1) Decembrie an curent față de decembrie an anterior
2) La sfârşitul perioadei;

a) Datele privind conturile naționale pentru anul 1994 sunt semidefinite;
b) Datele privind conturile naționale pentru anul 1995 sunt provizorii;
c) Datele operative pentru anul 1995.

Regula nr. 3: *Pentru fiecare caracteristică înscrisă într-un tabel statistic trebuie precizată unitatea de măsură.* În funcție de spațiul disponibil şi de particularitățile seriei statistice, unitățile de măsură pot fi înscrise între paranteze pătrate, lângă denumirea caracteristicii statistice (tabelul 3.1.), într-o linie sau coloană specială (tabelul 3.2.) sau, atunci când toate caracteristicile statistice din tabel au aceeaşi unitate de măsură, deasupra colțului din dreapta al tabelului (tabelul 3.3).

Tabelul 3.3. Rate ale dobânzilor practicate în sistemul bancar din România în perioada septembrie – decembrie 1993

- procente pe an -

Nr. crt.	**Luna**	**Rată medie la credite pentru clienți nebancari**	**Rată medie la depozite pentru clienți nebancari**	**Rată medie la operațiuni interbancare**
1	Sept.	53,8	33,8	...
2	Oct.	74,4	38,0	...
3	Nov.	83,6	41,4	61,7
4	Dec.	86,4	42,5	61,3

Sursa: Banca Națională a României, Raport Anual 1995

Regula nr. 4: Într-un tabel statistic trebuie indicate, de regulă prin note explicative, valorile provizorii, care ar putea fi modificate în urma unor calcule ulterioare, în care vor fi folosite date mai precise. De exemplu, în tabelul 3.2., indicatorii macroeconomici determinați pentru anii 1994 şi 1995 au un caracter provizoriu, întrucât în anul 1996, când au fost calculați, nu fuseseră definitivate conturile naționale (în care sunt înregistrate operațiunile comerciale şi financiare la nivel macroeconomic). S-a luat deci în considerare posibilitatea ca după definitivarea conturilor naționale să fie determinate alte valori ale indicatorilor. În plus, valorile exporturilor din anul 1995 nu au o acuratețe foarte ridicată deoarece în anul 1996, când au fost calculate, nu erau disponibile date complete asupra comerțului exterior.

Tabelul 3.4. Capitaluri proprii ale societăților bancare din România în perioada septembrie-decembrie 1995

- milioane lei; sfârşitul perioadei -

Nr. crt.	Luna	Capital statutar	Fond de rezervă	Profit net 1)	Alte fonduri
1	Sept.	636.915	372.718	x	818.059
2	Oct.	648.611	383.219	x	871.610
3	Nov.	661.210	393.918	x	1.001.217
4	Dec.	661.895	408.420	1.134.551	536.971

1) Până în luna decembrie 1995 a fost inclus la poziția „Alte pasive – profit” şi „Alte pasive – Vărsăminte şi prelevări din profit”

Sursa: Banca Națională a României, Raport Anual 1995

Regula nr. 5: *Se recomandă pentru mărimile la care se practică mai multe metode de determinare să fie precizat, pentru a se evita confuziile, modalitatea de obținere.* Aceasta se poate face fie prin indicații în celulele tabelului fie prin note explicative în afara celulelor. De exemplu, în tabelul 3.2, pentru rata anuală a inflației, indicator ce poate fi determinat în mai multe moduri (prin raportare a nivelului mediu al prețurilor din anul curent față de cel din anul anterior; prin raportare a nivelului prețului de la sfârşitul anului curent față de cel de la sfârşitul anului anterior ş.a.m.d.) s-a considerat necesar să se precizeze că s-a calculat pe baza prețurilor în vigoare în luna decembrie. De asemenea, pentru rata şomajului, un alt indicator care poate fi calculat în diferite moduri (pe baza numărului de şomeri de la sfârşitul perioadei sau a numărului mediu din perioada analizată) s-a precizat că s-au luat în calcul datele culese la sfârşitul fiecărui an. În ce priveşte indicatorul profitului net al societăților, prezentat în tabelul 3.4., s-a considerat necesar să se precizeze că lipsa valorilor în primele trei luni are drept cauză „înscrierea, în acea perioadă, a profitului la poziția „Alte pasive” şi nu la cea de „Capitaluri proprii”.

Regula nr. 6: *Într-un tabel statistic nu trebuie să existe celule necompletate.* Această regulă a fost instituită îndeosebi pentru a se evita completările ulterioare, de către alte persoane decât autorii tabelului, şi pentru a nu se oferi imaginea de nefinalizare. Pentru ca celulele tabelului să poată fi completate în orice condiții, au fost stabilite câteva simboluri speciale:

- **semnul „…”**, atunci când nu se cunosc datele care ar trebui să figureze în celule (de exemplu, în tabelul 3.3., acest semn figurează pentru valorile ratei medii la operațiunile interbancare din lunile septembrie şi octombrie deoarece abia din luna noiembrie 2005 s-a trecut la un calcul riguros al parametrilor operațiunilor interbancare);
- **semnul „x”,** atunci când într-o celulă nu trebuie să figureze nici un fel de dată (de exemplu, în tabelul 3.4. nu este cazul ca în primele trei luni să se treacă date pentru profitul net ca parte a capitalurilor proprii deoarece în această perioadă profitul era înregistrat la poziția „Alte pasive”);
- **semnul „-”**, atunci când data are o valoare nulă (de exemplu, în tabelul 3.2. acest semn figurează pentru rata şomajului din anul 1990 întrucât, cel puțin oficial, în acel an nu existau şomeri în România);
- semnul **„0”**, atunci când prin rotunjire rezultă valoarea zero (de exemplu, pentru valorile subunitare mai mici decât 0,5 în cazul numerelor afişate, fără zecimale, pentru cele mai mici decât 0,05 pentru numerele afişate cu o singură zecimală ş.a.m.d.).

Regula nr. 7. *Pentru datele prezentate într-un tabel statistic trebuie să se precizeze sursele din care au fost preluate*. Menționarea surselor oferă indicii asupra acurateței datelor folosite şi asupra responsabilităților în ce priveşte veridicitatea acestora.

3.1.3. Reprezentarea grafică a datelor statistice

Reprezentările grafice facilitează sesizarea rapidă a unor aspecte esențiale ale fenomenelor studiate. Sunt folosite, de asemenea, şi în cadrul unor tehnici de determinare a unor indicatori. Astfel de procedee, cu toate că uneori nu conferă o acuratețe prea mare au, față de calculele analitice, avantajul operativității. Datorită posibilității de sinteză rapidă pe care o oferă, reprezentările grafice sunt folosite destul de frecvent şi în fundamentarea unor decizii manageriale. În ultimii ani, dezvoltarea tehnologiilor informaționale a făcut posibilă realizarea operativă a reprezentărilor grafice, oricât de complexe ar fi acestea.

Tehnicile de reprezentare grafică a datelor statistice formează o gamă foarte largă, diversificată în raport cu obiectivele cercetării şi cu tipurile de date folosite. Marea varietate a acestor procedee, ca şi faptul că acestea nu sunt folosite prea des în calcule de precizie, au făcut ca regulile asupra reprezentărilor grafice să nu fie la fel de stricte şi de universal valabile precum cele utilizate pentru tabelele statistice. Pot fi, totuşi, menționate câteva recomandări generale, menite să inducă o anumită rigoare în reprezentarea grafică a datelor statistice:

- pentru datele statistice reprezentate trebuie precizate unitățile de măsură;
- este indicat ca graficele să fie proporționale cu valorile datelor reprezentate iar relațiile de proporționalitate să fie precizate printr-o aşa-numită *scară a graficului*;
- atunci când se folosesc simboluri, acestea trebuie explicate în cadrul unei aşa – numite *legende a graficului*;
- pentru datele statistice reprezentate trebuie să se indice sursele din care s-au obținut.

3.2. Prelucrarea primară a datelor cu caracteristici atributive

3.2.1. Prelucrarea primară prin serii atributive simple

O serie atributivă simplă prezintă o populație statistică desemnând pentru fiecare unitate câte o valoare din fiecare caracteristică atributivă. Astfel de serii sunt uşor de alcătuit, însă utilizarea lor în calculele statistice ulterioare poate fi destul de dificilă în cazul unui număr mare de unități statistice.

Reprezentarea seriilor atributive simple se poate face atât prin tabele cât şi pe cale grafică. Dintre tehnicile grafice se detaşează, prin acuratețe, *reprezentările în coordonate carteziene* (numite astfel în onoarea lui Descartes, cel care le-a descoperit) în care valorile asociate unităților statistice sunt transpuse într-un sistem de axe. Cel mai adesea se foloseşte un sistem de două axe: una orizontală, numită *axa ordonatelor* şi una verticală, numită axa *absciselor.* La intersecția acestora se află punctul de origine, notat cu 0, cu coordonate de valori nule. Reprezentările în coordonate carteziene ale seriilor atributive simple sunt indicate pentru aprecierea influenței pe care un factor o are asupra unui fenomen. În acest scop sunt reprezentate două caracteristici statistice:

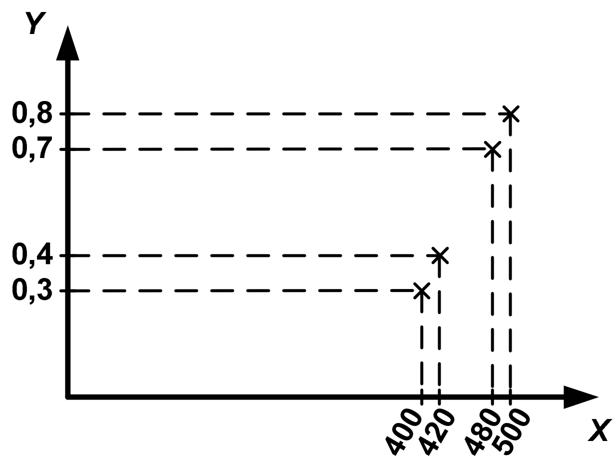
- o caracteristică numită *variabilă independentă*, reprezentată pe axa absciselor si care reflectă stările factorului de influență;
- o caracteristică numită *variabilă dependentă,* reprezentată pe axa ordonatelor şi care reflectă manifestarea fenomenului supus influenței.

Pentru fiecare axă a unei reprezentări în coordonate carteziene trebuie stabilită câte o scară de valori prin raportarea la spațiul disponibil a valorii absolute maxime a datelor prezentate. Atunci când valoarea absolută minimă este depărtată de origine se poate câştiga spațiu translatând intervalul de valori mai aproape de intersecția axelor. O astfel de operațiune trebuie indicată pe grafic prin aplicarea unui simbol al secționării la axa la care s-a făcut translatarea.

Exemplul 3.1. În tabelul 3.5. sunt prezentate valorile suprafețelor comerciale şi ale vânzărilor lunare de mărfuri pentru cele patru puncte de desfacere ale unei firme. Se cere să se studieze, pe cale grafică, influența pe care suprafața comercială o are asupra vânzărilor dintr-un punct de desfacere.

Tabelul 3.5. Suprafețele comerciale şi vânzările lunare de mărfuri pentru patru puncte de desfacere

Nr. crt.	**Suprafață comercială [m^2]**	**Vânzări lunare de mărfuri [mil. RON]**
(0)	(1)	(2)
1	400	0,3
2	420	0,4
3	480	0,7
4	500	0,8

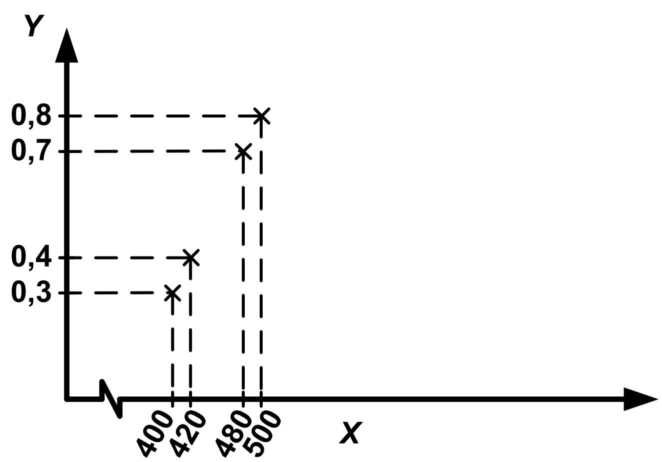


Scara: Ox – 1 cm = 50 m^2 Oy – 1 cm = 0,1 mil. RON

Figura 3.1. Reprezentarea grafică a unei serii atributive simple fără translatarea intervalului de valori.

Rezolvare: Reprezentarea grafică s-a făcut printr-un sistem de două axe:

- axa absciselor (Ox) a fost repartizată variabilei independente, adică suprafeței comerciale;
- axa ordonatelor (Oy) a fost repartizată variabilei dependente, adică vânzărilor lunare de mărfuri (fig. 3.1.).



Scara:
Ox – 1 cm = 10 m^2; Oy – 1 cm = 0,1 mil. RON

Fig. 3.2. Reprezentarea grafică a unei serii atributive simple cu translatarea intervalului de valori

Se poate observa, totuşi, că pe axa absciselor valoarea minimă a suprafeţei comerciale este foarte îndepărtată de origine, ceea ce deschide posibilitatea obţinerii de spaţiu suplimentar. După cum se poate observa în figura 3.2., prin translatarea intervalului de valori se poate trece la o scară de cinci ori mai mică, ceea ce uşurează interpretările.

Reprezentările grafice evidenţiază faptul că vânzările de mărfuri sunt cu atât mai mari cu cât suprafaţa comercială este mai mare (ceea ce are semnificaţia unei legături directe între cele două variabile) iar punctele reprezentate au tendinţa de liniaritate.

Prin reprezentări grafice poate fi studiată şi dependenţa unui fenomen faţă de mai mulţi factori, folosindu-se mai multe axe, însă în acest caz analiza devine destul de complexă.

3.2.2. Prelucrarea primară prin distribuţii de frecvenţe

3.2.2.1. Conceptul de distribuţie de frecvenţe

O distribuţie de frecvenţe este o serie atributivă care prezintă o populaţie statistică prin două elemente:

- formele pe care le îmbracă o caracteristică atributivă a populaţiei (sau, în cazul unei serii multidimensionale, caracteristicile atributive ale populaţiei);
- frecvenţele absolute, care reprezintă, pentru fiecare din formele caracteristicii, numărul de unităţi statistice.

În funcţie de natura caracteristicii atributive pot fi delimitate două tipuri de distribuţii de frecvenţe:

- distribuţii homograde;
- distribuţii heterograde.

3.2.2.1.1. Distribuţii homograde

La o distribuţie homogradă caracteristica atributivă este calitativă, adică nu poate fi exprimată numeric. De exemplu, în descrierea unei persoane pot fi utilizate mai multe caracteristici calitative: cetăţenia, etnia, sexul, religia, profesia, studiile, culoarea ochilor, a părului etc. Fiecare dintre aceste caracteristici poate îmbrăca mai multe stări distincte net una faţă de alta, iar numărul de unităţi statistice înregistrate pentru o astfel de stare reprezintă frecvenţa absolută asociată acesteia. De exemplu, seria prezentată în tabelul 3.6. este o distribuție homogradă în care stările caracteristicii calitative sunt reprezentate de cauzele care au determinat plecarea unor angajați de la o firmă, iar frecvențele absolute sunt date de numărul de persoane asociat fiecărei cauze.

Tabelul 3.6. Numărul şi cauzele ieşirilor de personal din cadrul unei firme

Nr. crt.	**Cauze ale ieşirilor de personal**	**Număr de persoane**
(0)	(1)	(2)
1	Restructurarea activității	7
2	Pensionări	17
3	Plecări la cerere	6
4	Acte de indisciplină	4

Unitățile statistice în raport cu care se face repartizarea frecvențelor pot îmbrăca diferite forme în funcție de scopul în care se realizează distribuția. De exemplu, dacă se studiază repartizarea suprafețelor de depozitare a unor materiale de construcție, unitățile statistice pot lua forma unor unități de suprafață. În acest caz, frecvența absolută asociată unui sortiment de material va fi dată de numărul de unități de suprafață ocupate de acesta (tab. 3.7).

Tabelul 3.7. Repartizarea suprafețelor de depozitare ale unor materiale de construcție

Nr. crt.	**Sortiment de materiale de construcții**	**Suprafață ocupată [m^2]**
(0)	(1)	(2)
1	Cărămizi	120
2	Bolțari	40
3	Ciment	20

Este indicat ca elementele unei distribuții homograde să fie organizate în ordinea descrescătoare a frecvențelor astfel încât să fie evidențiate (deoarece vor fi citite primele) stările cele mai reprezentative (cu cele mai mari frecvențe) ale caracteristicii calitative, aşa cum se poate observa în tabelele 3.6. şi 3.7.

Se pot face însă excepții pentru situațiile în care există o ordine naturală sau conceptuală între stările caracteristicii. De exemplu, atunci când se distribuie personalul unei firme în raport cu studiile absolvite, se începe cu nivelul cel mai redus de pregătire ajungându-se la cel mai înalt (tabelul 3.8.).

Tabelul 3.8. Distribuția personalului unei firme în raport cu studiile absolvite

Nr. crt.	Studii absolvite	Număr de angajați
(0)	(1)	(2)
1	Absolvenți de şcoli profesionale	20
2	Absolvenți de licee	50
3	Absolvenți de învățământ superior	10

Alcătuirea unei distribuții homograde este în general simplă în condițiile în care distincția dintre stările unei caracteristici calitative nu este prea dificilă. Practic, o distribuție homogradă poate fi realizată chiar în cadrul operațiunii de culegere a datelor, prin clasificarea acestora în raport cu stările unei caracteristici (sau mai multor caracteristici) calitative. Atunci când sunt utilizate mai multe caracteristici calitative, se alcătuieşte o distribuție homogradă combinată: se începe cu gruparea după o primă caracteristică după care grupele obținute se împart în subgrupe în raport cu o altă caracteristică ş.a.m.d. În operațiile de grupare a datelor după caracteristici calitative vom folosi următoarele notații: n_i^x este frecvența absolută grupei i formată după caracteristica x; K_x este numărul de grupe formate în raport cu caracteristica x.

Exemplul 3.2. În tabelul 3.9. sunt prezentate datele culese asupra unui grup de 12 angajați au unei firme cu privire la trei caracteristici calitative: departamentul care lucrează, studiile absolvite şi sexul. Se cere să se alcătuiască distribuții homograde unidimensionale în raport cu cele trei caracteristici precum şi distribuții combinate de două şi trei caracteristici.

Tabelul 3.9. Date culese asupra unui grup de 12 angajați cu privire la trei caracteristici calitative

Nr. crt.	Departamentul în care lucrează 1)	Studiile absolvite 2)	Sex 3)
(0)	(1)	(2)	(3)
1	DP	SS	B
2	DP	SL	B

3	DP	SP	B
4	DM	SS	F
5	DM	SS	F
6	DM	SL	B
7	DF	SS	F
8	DF	SS	B
9	DR	SS	B
10	DR	SL	F
11	DL	SS	F
12	DL	SP	B

1) Pentru datele asupra departamentelor în care lucrează cei 12 angajaţi au fost folosite următoarele simboluri: DP – departamentul de producţie; DM – departamentul de marketing; DF – departamentul de finanţe-contabilitate; DR – departamentul relaţiilor cu personalul; DL – departamentul de logistică.

2) Pentru datele asupra studiilor absolute au fost folosite următoarele simboluri: SP – absolvent de şcoală profesională, SL – absolvent de liceu, SS – absolvent de studii superioare.

3) pentru datele asupra sexului angajaţilor au fost folosite următoarele simboluri: B – sex bărbătesc, F – sex femeiesc. **Rezolvare**: Distribuţiile homograde unidimensionale, destul de simplu de realizat, sunt prezentate în tabelele 3.10., 3.11. şi 3.12.

Tabelul 3.10. Distribuţia angajaţilor în raport cu departamentul în care lucrează

Nr. crt.	**Studiile absolvite**	**Frecvenţa absolută** (n_i^x)
(0)	(1)	(2)
1	DP	3
2	DM	3
3	DF	2
4	DR	2
5	DL	2
6	Total	12
7	Simbol pentru total	$\sum_{i=1}^{K_x} n_i^x$

Tabelul 3.11. Distribuția angajaților în raport cu studiile absolvite

Nr. crt.	**Sex**	**Frecvența absolută** (n_i^x)
(0)	(1)	(2)
1	SS	7
2	SL	3
3	SP	2
4	Total	12
5	Simbol pentru total	$\sum_{i=1}^{K_y} n_i^y$

Tabelul 3.12. Distribuția angajaților după sex

Nr. crt.	**Sex**	**Frecvența absolută** (n_i^x)
(0)	(1)	(2)
1	B	7
2	F	5
3	Total	12
4	Simbol pentru total	$\sum_{i=1}^{K_z} n_i^z$

Dintre variantele posibile de distribuții combinate după două caracteristici, în acest exemplu va fi prezentată, în tabelul 3.13., gruparea după departament şi după studiile absolvite.

Tabelul 3.13. Distribuțiile angajaților în raport cu departamentul în care lucrează şi cu studiile absolvite:

Studii absolvite / **Departament**	**SS**	**SL**	**SP**	**Total**
DP	1	1	1	3
DM	2	1	-	3
DF	2	-	-	2
DR	1	1	-	2
DL	1	-	1	2

Total	7	3	2	12

Gruparea combinată după cele trei caracteristici este prezentată în tabelul 3.14.

Tabelul 3.14. Distribuția angajaților după cele trei caracteristici

Grupe după departament	**Subgrupe după studii absolvite şi sex**						**Total**
	SS		**SL**		**SP**		
	B	**F**	**B**	**F**	**B**	**F**	
DP	1	-	1	-	1	-	3
DM	-	2	1	-	-	-	3
DF	1	1	-	-	-	-	2
DR	1	-	-	1	-	-	2
DL	-	1	-	-	1	-	2
Total	3	4	2	1	2	-	12

3.2.2.1.2. Distribuții heterograde

La o distribuție heterogradă caracteristica atributivă este cantitativă, adică poate fi cuantificată. De exemplu, o persoană poate fi descrisă prin mai multe caracteristici cantitative: înălțime, greutate etc. Spre deosebire de distribuțiile homograde, unde distincția dintre stările caracteristicilor se face de la sine, la distribuțiile heterograde este necesară stabilirea unor intervale de variație ale valorilor caracteristicii cantitative. Acestea vor reprezenta formele caracteristicii în raport cu care se vor grupa unitățile statistice. Frecvența absolută a unei grupe va fi dată de numărul de unități statistice ale căror date le încadrează în intervalul de valori asociat grupei.

În raport cu trăsăturile caracteristicii cantitative se pot delimita două tipuri de distribuții heterograde:

a) distribuții heterograde de tip discret;
b) distribuții heterograde de tip continuu.

a) La o **distribuție heterogradă de tip discret** caracteristica poate lua doar valori numărabile, în salturi. În tabelul 3.15 este prezentată o astfel de distribuție, în care caracteristica de grupare este numărul de vânzători din cadrul punctelor de desfacere ale unei firme. Pentru valorile discrete ale caracteristicilor este necesar să se prevină echivocul asupra repartizării unităților statistice pe intervale de variație. În tabelul 3.15, de exemplu, deoarece s-a precizat că limita inferioară este inclusă în interval, un punct de vânzare cu 10 lucrători va fi repartizat la a doua grupă, iar unul cu 20 de lucrători, la a treia. Astfel de precizări pot fi făcute şi prin prezentarea intervalelor de variaţie drept închise sau deschise (tabelul 3.16.).

Tabelul 3.15. Repartizarea punctelor de desfacere ale unei firme în raport cu numărul de vânzători

Nr. crt.	**Număr de vânzători**	**Număr de puncte de desfacere**	**Centru de interval**
(0)	(1)	(2)	(3)
1	0 – 5	4	2,5
2	5 – 10	3	7,5
3	10 – 15	2	12,5

Notă: Limita inferioară este inclusă în interval

b) La o **distribuţie heterogradă de tip continuu** se consideră că o caracteristică poate lua orice valoare dintr-un interval. În tabelul 3.16 este prezentată o distribuţie de tip continuu în care caracteristica de grupare este reprezentată de suprafaţa punctelor de vânzare ale unei firme. Şi la astfel de serii se recomandă delimitarea fără echivoc a intervalelor de variaţie, deşi, în principiu, nu ar trebui să apară confuzii asupra repartizării unităţilor statistice pe intervale de variaţie (de exemplu, în ce priveşte tabelul 3.16., este imposibil ca suprafaţa unui punct comercial să fie exact 50 m^2; nu poate fi decât mai mare sau mai mică, astfel încât se poate încadra uşor la un interval de variaţie).

Tabelul 3.16. Repartizarea punctelor de desfacere ale unei firme în raport cu suprafaţa

Nr. crt.	Suprafaţa comercială	Număr de puncte de desfacere	Centru de interval
(0)	(1)	(2)	(3)
1	[30 – 50]	3	40
2	[50 – 70]	4	60
3	[70 – 90]	2	80

Pentru fiecare grupă a unei distribuții heterograde poate fi stabilită o valoare reprezentativă, numită *centrul de interval*, care reprezintă mijlocul intervalului de variație asociat grupei. O astfel de valoare are semnificația unui rezultat al factorilor ce acționează în mod permanent în cadrul grupei, în absența unor factori accidentali, temporari. În general, în determinarea indicatorilor statistici pe baza distribuțiilor heterograde se consideră că toate unitățile statistice dintr-o grupă au o valoare a caracteristicii de grupare egală cu centrul de interval, ceea ce uşurează calculele şi permite relevarea aspectelor esențiale.

Centrul de interval al unei grupe poate fi calculat prin formula:

$$X_i^{'} = \frac{X_i - X_{i-1}}{2} \qquad (3.1.)$$

în care: - $X_i^{'}$ este centrul de interval al grupei *i*;

- X_{i-1} este limita inferioară a intervalului de variație asociat grupei *i*;

- X_i este limita superioară a intervalului de variație asociat grupei *i*.

În alegerea numărului şi lungimii intervalelor de variație ale unei distribuții heterograde sunt luate în considerare atât aspecte care țin de relevarea factorilor permanenți şi accidentali de influență cât şi considerente ale facilitării calculelor. Se apreciază că lungimile mari ale intervalelor de variație permit evidențierea factorilor permanenți de influență însă pot duce la ignorarea efectelor factorilor accidentali, ceea ce impietează asupra acurateței analizei statistice. În acelaşi timp. lungimile mici ale intervalelor de variație, care reflectă într-o măsură mai mare influența factorilor accidentali, fac dificilă distincția dintre aceştia şi factorii permanenți. Se recomandă să se evite apariția unor grupe cu frecvența absolută nulă, iar limitele intervalelor să nu conțină prea multe zecimale. Destul de frecvent, pentru facilitarea calculelor, se folosesc distribuții heterograde cu intervale de variație egale, numite şi *serii de variație cu grupe egale*.

Pentru distribuțiile heterograde cu intervale de variație egale, stabilirea grupelor se poate face pornind de la o lungime dată a intervalelor de variație fie de la un număr dat al acestor intervale. În alcătuirea grupelor pe baza lungimii date a intervalelor de variație se poate utiliza un algoritm descris de următoarele reguli:

1) se stabilesc limitele variației caracteristicii de grupare:

X_{min}, care este cea mai mică valoare a caracteristicii; X_{max}, care este cea mai mare valoare a caracteristicii;

2) se determină primul interval de variație având ca limită inferioară valoarea X_{min}, iar ca limită superioară valoarea $X_{min} + dx$, unde dx este lungimea aleasă a intervalului de variație;
3) pentru intervalele de variație următoare, limita inferioară va fi reprezentată de limita superioară a intervalului anterior, iar limita superioară se obține adăugând la limita inferioară lungimea intervalului;
4) operațiunea de stabilire a grupelor se încheie atunci când s-a ajuns la o limită superioară de interval care să fie mai mare sau egală decât X_{max}.

Exemplul 3.3.: În tabelul 3.17. sunt prezentate datele culese asupra productivității muncii pentru un grup de 10 angajați ai unei firme. Se cere să se grupeze aceste date într-o distribuție heterogradă în care lungimea fiecărui interval de variație să fie egală cu 20 RON/lună.

Tabelul 3.17. Date culese asupra productivității muncii pentru un grup de 12 angajați

RON/lună

Nr. crt.	Productivitatea medie lunară a muncii
(0)	(1)
1	102,46
2	122,38
3	124,21
4	111,42
5	142,04
6	90,75
7	129,46
8	149,28
9	99,00
10	132,42
11	128,61
12	111,62

Rezolvare: Din trecerea în revistă a datelor rezultă limitele variației caracteristicii de grupare:

X_{min} = 90,75 RON/lună; X_{max} = 149,28 RON/lună.

În acest caz, pentru a se opera cu numere rotunde, vom începe alcătuirea grupelor nu de la 90,75 ci de la 90 RON/lună. Vor rezulta, astfel, intervalele de variație prezentate în figura 3.3. şi grupele prezentate în tabelul 3.18.

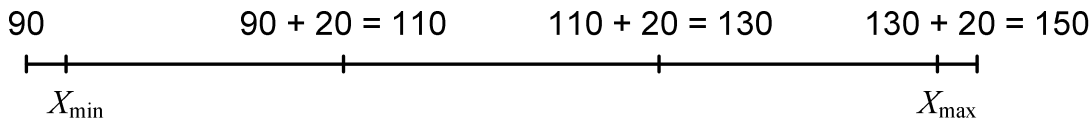


Figura 3.3. Stabilirea intervalelor de variație

Tabelul 3.18. Gruparea angajaților pe baza unor intervale cu o lungime dată

Nr. crt.	**Interval de variație (X_{i-1} ; X_i) [RON/lună]**	**Frecvență absolută** $\left(n_i^x\right)$
(0)	(1)	(2)
1	(90 – 110]	3
2	(110 – 130]	6
3	(130 – 150]	3
4	Total	12
5	Simbol pentru total	$\sum_{i=1}^{Kx} n_i^x$

În alegerea lungimii unui interval de variație poate fi folosită formula recomandată de statisticianul Herbert Sturges:

$$d_x = \frac{X_{max} - X_{min}}{1 + 3{,}322 \times \lg N} \qquad (3.2.)$$

unde N este numărul de unități statistice care trebuie grupate.

Pentru stabilirea grupelor unei distribuții heterograde cu intervale egale pornind de la numărul acestora se poate folosi următorul algoritm:

1) se stabilesc limitele variației caracteristicii de grupare X_{min} şi X_{max};
2) se determină lungimea unui interval de grupare (d_x) pe baza formulei:

$$d_x = \frac{X_{max} - X_{min}}{K_x} \qquad (3.3.)$$

unde K_x este numărul ales de grupe;

3) se împarte intervalul [$X_{\min}$; $X_{\max}$] în K_x intervale de variație, fiecare cu lungimea d_x.

Este de menționat faptul că valoarea d_x nu poate fi rotunjită prin micşorare (situație în care există riscul ca unele unități statistice să nu mai fie cuprinse în nicio grupă) ci doar prin majorare.

Atunci când sunt utilizate mai multe caracteristici cantitative, se poate proceda la fel ca în cazul distribuțiilor homograde, alcătuindu-se distribuții combinate, în care grupele obținute în raport cu o caracteristică sunt împărțite în subgrupe în raport cu alte caracteristici.

Exemplul 3.4.: În tabelul 3.19 prezentate datele culese asupra cifrei de afaceri şi profitului pentru un grup de 12 firme dintr-o ramură industrială. Se cere:

a) să se grupeze datele în trei grupe cu intervale de variație egală în raport cu cifra de afaceri;
b) să se grupeze datele în patru grupe cu intervale de variație egală în raport cu profitul;
c) să se realizeze gruparea combinată a datelor în raport cu cele două caracteristici, pe baza grupărilor anterioare.

Tabelul 3.19. Date asupra cifrei de afaceri şi profitului pentru un grup de 12 firme

mil. RON

Nr. crt.	**Cifră de afaceri**	**Profit**
(0)	(1)	(2)
1	10,8	2,5
2	29,6	6,0
3	30,0	6,0
4	12,8	4,5
5	10,0	2,0
6	40,0	10,0
7	38,0	9,8
8	24,3	7,2
9	29,8	6,8
10	22,4	5,4
11	32,4	7,5
12	22,8	5,5

Rezolvare: În acest exemplu, pentru facilitarea calculelor, vom nota cu X caracteristica cifrei de afaceri şi cu Y caracteristica profitului.

a) Pentru gruparea după caracteristica X se pleacă de la limitele variaţiei:

X_{min} = 10,0 mil. RON; X_{max} = 40,0 mil. RON şi de la numărul de grupe $K_x = 3$

Se obţine apoi lungimea unui interval de variaţie prin formula:

$$d_x = \frac{X_{\max} - X_{\min}}{K_x} = \frac{40-10}{3} = 10 \text{ mil. RON}$$

Se trece în continuare la determinarea celor trei grupe, prezentate în tabelul 3.20.

Pentru repartizarea unităţilor statistice asupra grupelor s-a considerat că fiecare interval de variaţie este închis pentru limita inferioară şi deschis pentru cea superioară. S-a făcut o excepţie pentru ultimul interval (această excepţie nu face însă ca lungimile intervalelor de variaţie să nu mai fie egale) astfel încât grupa să includă şi firma cu cea mai mare cifră de afaceri.

Tabelul 3.20. Gruparea celor 12 firme în raport cu cifra de afaceri

Nr. crt.	**Interval de variaţie (X_{i-1} ; X_i) [mil. RON]**	**Frecvenţă absolută $\left(n_i^x\right)$**
(0)	(1)	(2)
1	[10 ; 20)	3
2	[20 ; 30)	5
3	[30 ; 40]	4
4	Total	12
5	Simbol pentru total	$\sum_{i=1}^{Kx} n_i^x$

b) La gruparea după caracteristica Y se porneşte de la numărul de grupe $K_y = 4$ şi de limitele variaţiei: $y_{min} = 2$; $y_{max} = 10$. Se determină apoi lungimea unui interval de variaţie prin formula:

$d_y = \frac{y_{max} - y_{min}}{K_y} = \frac{10-2}{4} = 2$ mil. RON. Sunt stabilite, în continuare, cele patru grupe, prezentate în tabelul 3.21.

Tabelul 3.21. Gruparea celor 12 firme în raport cu profitul

Nr. crt.	**Interval de variație (y_{i-1}; y_i) [mil. RON]**	**Frecvență absolută $\left(n_i^y\right)$**
(0)	(1)	(2)
1	[2-4)	2
2	[4 – 6)	3
3	[6 – 8)	5
4	[8 – 10]	2
5	Total	12
6	Simbol pentru total	$\sum_{i=1}^{K_y} n_i^y$

Din aceleaşi considerente ca pentru gruparea după caracteristica X, la gruparea după caracteristica Y s-a ales ca intervalul de variație al ultimei grupe să fie închis la ambele capete.

c) Gruparea combinată a celor 12 firme în raport cu cifra de afaceri şi cu profitul este prezentată în tabelul 3.22.

Distribuțiile heterograde cu intervale de variație egale sunt, aşa cum a rezultat din exemplele anterioare, uşor de alcătuit iar utilizarea lor, după cum se va vedea în capitolele ulterioare, facilitează calculele statistice. Totuşi, folosirea lor pentru a descrie o populație statistică nu este întotdeauna posibilă. Uneori, stabilirea unor intervale de variație egale conduce la situații în care în unele grupe frecvențele absolute sunt foarte mici sau chiar nule. O astfel de situație ar apărea dacă s-ar încerca gruparea localităților din România, în intervale de variație egală, după numărul de locuitori, în condițiile în care mai mult de jumătate dintre localități au mai puțin de 100 000 de locuitori, nici o localitate nu se încadrează în intervalul cuprins între 500 000 şi 2 000 000 de locuitori, iar o singură localitate, capitala, are mai mult de 2 000 000 de locuitori. Într-un astfel de caz, rezolvarea poate veni de la folosirea unor intervale inegale de variație, în care frecvențele să fie ceva mai uniform distribuite. Pentru stabilirea intervalelor inegale de variație sunt folosite diverse tehnici: cumularea mai multor intervale egale, intervale cu lungimi aflate în progresie geometrică etc.

Tabelul 3.22. Gruparea combinată a celor 12 firme în raport cu cifra de afaceri şi cu profitul

Intervale de variație după profit ($y_{i-1} - y_i$) [mil. RON] / **Intervale de variație după cifra de afaceri ($x_{i-1} - x_i$) [mil. RON]**	**[2–4)**	**[4–6)**	**[6–8)**	**[8–10]**	**Total**
[10 – 20)	2	1	-	-	3
[20 – 30)	-	2	3	-	5
[30 – 40]	-	-	2	2	4
Total	2	3	5	2	12

3.2.2.2. Relația dintre distribuțiile de frecvențe şi distribuțiile de probabilitate

Pentru a aborda relația dintre distribuțiile de frecvențe şi distribuțiile de probabilități va trebui, mai întâi, să introducem noțiunea de *frecvență relativă a unei grupe.* Aceasta reprezintă ponderea pe care frecvența absolută a unei grupe oarecare, dintr-o distribuție de frecvențe, o are în totalul frecvențelor absolute şi poate fi calculată prin formula: $n_{r_{i_i}}^x = \frac{n_i^x}{\sum_{j=1}^{K_x} n_j^x}$ (3.4.)

în care: $n^x_{r_i}$ este frecvenţa relativă a grupei i; n^x_i este frecvenţa absolută a grupei i;

k_x este numărul de grupe format după caracteristica x.

Uneori, frecvenţele relative sunt exprimate într-o formă procentuală, mai ales atunci când se doreşte evidenţierea ponderii pe care o grupă o are în ansamblul seriei.

În cadrul statisticii, frecvenţele relative sunt utilizate atât pentru evidenţierea structurii unei populaţii statistice cât şi în anticiparea evoluţiei unui fenomen colectiv. Anticipările au la bază aşa-numitul *postulat al stabilităţii frecvenţelor relative*, care enunţă că dacă se vor face în condiţii asemănătoare mai multe culegeri de date statistice, fiecare cu un număr suficient de mare de unităţi statistice, atunci frecvenţele relative pentru un anumit eveniment nu vor diferi prea mult dintre ele. Altfel spus, revenind la ultimele două exemple, dacă în activitatea de producţie nu vor surveni schimbări importante, este de aşteptat ca şi în viitor rebuturile remaniabile să reprezinte 15% iar cele definitive 10% din rezultatele producţiei, tot aşa cum, dacă nu vor interveni schimbări importante în activitatea firmei distribuitoare de energie termică, 80% dintre restanţierii acesteia vor fi reprezentaţi de familii cu venituri lunare pe membru de familie cuprinse între 100 şi 200 de euro.

Anticipările pe baza frecvenţelor relative pot îmbrăca forma probabilităţilor. Legătura dintre cele două noţiuni a fost realizată prin aşa-numita *Lege a numerelor mari,* formulată în anul 1713 de către Jacob Bernoulli. În esenţă, această lege stipulează că dacă un eveniment A s-a produs de n ori într-o serie de N experimente identice şi independente (adică rezultatele unui experiment nu le pot influenţa pe celelalte), atunci se poate considera, cu condiţia ca N să fie suficient de mare, că probabilitatea de realizare a evenimentului A este dată de relaţia: $P_{(A)} = \frac{n}{N}$ (3.5.)

Numărul N al experimentelor poate fi asimilat totalului unităţilor dintr-o populaţie statistică (astfel spus, totalului frecvenţelor absolute) întrucât unităţile statistice pot fi considerate drept forme de înregistrare ale manifestării unui fenomen studiat. De asemenea, numărul n care arată de câte ori s-a produs un eveniment în cadrul experimentelor poate fi asimilat frecvenţei absolute a unei grupe, deoarece o grupă poate fi considerată drept o reuniune de unităţi statistice pentru care fenomenul studiat s-a manifesta în acelaşi mod (s-a transpus într-un acelaşi eveniment). De exemplu, dacă evenimentul este reprezentat de producerea unui rebut remaniabil, grupa asociată acestuia va cuprinde toate piesele care au fost rebutate dar care pot fi remaniate. De asemenea, dacă evenimentul este reprezentat de faptul că o familie cu un venit lunar mediu cuprins între 100 şi 200 de euro nu şi-a achitat factura pentru energie termică, grupa asociată acestuia va include toate familiile restanțiere cu venituri cuprinse în acea categorie. În aceste condiții, relația 3.5. devine:

$$P_{(A)} = \frac{n}{N} = \frac{n_i^x}{\sum_{i=1}^{K_x} n_i} = n_{r_i}^x \qquad (3.6.)$$

ceea ce ar sugera echivalența dintre probabilități şi frecvențele relative. Totuşi, această echivalență nu poate rezista la o analiză riguroasă deoarece condițiile precizate în legea numerelor mari, referitoare la realizarea unui număr suficient de mare de experimente identice şi independente, nu pot fi îndeplinite în practică. În consecință, frecvențele relative trebuie considerate, mai degrabă, nişte aproximări ale probabilităților. Estimarea probabilităților pe baza frecvențelor relative, deşi nu are întotdeauna o mare acuratețe, este folosită, totuşi, destul de frecvent în practică datorită facilității calculelor.

3.2.2.3. Reprezentarea grafică a distribuțiilor de frecvențe

În reprezentarea grafică a distribuțiilor de frecvențe sunt folosite variate metode. Dintre acestea vom prezenta trei tipuri care se detaşează prin frecvența utilizării:

- diagrame pentru frecvențe absolute;
- diagrame de structură;
- reprezentări în coordonate carteziene.

3.2.2.3.1. Reprezentarea grafică a distribuțiilor de frecvențe

Prin diagramele pentru frecvențe absolute pot fi reprezentate grafic atât distribuții homograde cât şi distribuții heterograde. În esență, un astfel de procedeu constă în reprezentarea fiecărei grupe printr-o figură geometrică a cărei suprafață este direct proporțională cu frecvența absolută a grupei. În raport cu figurile geometrice folosite se pot delimita mai multe tipuri de diagrame pentru frecvențe absolute:

a) diagrame în cercuri;
b) diagrame în pătrate;
c) diagrame în dreptunghiuri etc.

a) Reprezentările grafice prin **diagrame în cercuri** nu sunt prea facile ca urmare a unor dificultăți în stabilirea unor suprafețe proporționale cu frecvențele absolute. După cum se ştie, suprafața unui cerc este dată de relația: $S_C = \pi \cdot r_C{}^2$

(3.7.)

unde: S_C este suprafața cercului; r_C este raza cercului.

În aceste condiții, rădăcinile pătrate ale razelor cercurilor trebuie să fie proporționale cu frecvențele absolute pentru ca acestea, la rândul lor, să fie proporționale cu suprafețele cercurilor. Această proporționalitate se exprimă prin relația: $K_C \cdot n_i^x = S_{C_i} = \pi \cdot r_{C_i}^2$ (3.8.)

unde K_C este o constantă, care reflectă relația dintre suprafața cercului asociat unei grupe *i* şi frecvența absolută a acesteia. Rezultă că raza cercului asociat unei grupe poate fi calculată prin formula:

$$r_{C_i} = \sqrt{\frac{K_C \times n_i^x}{\pi}} \qquad (3.9.)$$

Constanta K_C, ce reprezintă, în fapt, scara la care se desenează diagrama în cercuri, este aleasă luându-se în considerare spațiul disponibil şi avantajele operării cu raze exprimate în numere întregi.

b) Reprezentările grafice prin **diagrame în pătrate** sunt caracterizate prin dificultăți similare celor care apar în cazul diagramelor în cercuri. După cum se ştie, suprafața unui pătrat (S_P) este dată de pătratul laturii sale *(a_P)*:

$$S_P = a_P^2 \qquad (3.10.)$$

Reiese că rădăcinile pătrate ale laturilor pătratelor trebuie să fie proporționale cu frecvențele absolute, astfel încât acestea să fie, la rândul lor, proporționale cu suprafețele pătratelor. Această condiție se exprimă prin relația:

$$K_P \cdot n_i^x = S_{P_i} = a_{P_i}^2 \qquad (3.11.)$$

în care K_P este o constantă care reflectă raportul dintre suprafața pătratului asociat unei grupe *i* şi frecvența absolută a acesteia. Rezultă că latura pătratului asociat unei grupe poate fi obținută pe baza relației:

$$a_{P_i} = \sqrt{K_P \cdot n_i^x} \qquad (3.12.)$$

Scara la care se desenează o diagramă în pătrate, exprimată prin constanta K_P este aleasă, la fel ca în cazul diagramelor în cercuri, luându-se în considerare spaţiul disponibil şi avantajele operării cu laturi ce reprezintă numere întregi.

c) **Diagramele în dreptunghiuri** sunt mai simplu de desenat faţă de cele în cercuri sau în pătrate, motiv pentru care sunt folosite mai frecvent decât acestea. Aşa cum se ştie, suprafaţa unui dreptunghi (S_D) este dată de produsul dintre latura sa verticală $\left(l_{D_v}\right)$ şi latura sa orizontală $\left(l_{D_o}\right)$: $S_D = l_{D_v} \times l_{D_o}$ (3.13.)

Dacă se alege pentru toate dreptunghiurile diagramei o aceeaşi latură orizontală, atunci doar latura verticală va trebui să fie proporţională cu frecvenţa absolută. Această proporţionalitate este exprimată prin relaţia:

$$K_D \cdot n_i^x = S_{D_i} = l_{D_{v_i}} \times l_{D_{o_i}} \qquad (3.14.)$$

în care K_D este o constantă cu rolul de a reflecta raportul dintre suprafaţa dreptunghiului asociat unei grupe i şi frecvenţa absolută a acesteia. Rezultă că latura verticală a dreptunghiului unei grupe poate fi calculată prin formula:

$$l_{D_{v_i}} = \frac{K_D}{l_{D_o}} \times n_i^x \qquad (3.15.)$$

Ca şi pentru celelalte diagrame latura orizontală şi constanta K_D sunt alese luând-se în considerare spaţiul disponibil şi avantajele operării cu laturi ce reprezintă numere întregi.

3.2.2.3.2. Diagrame de structură

Prin diagramele de structură sunt reprezentate ponderile, date de frecvenţele relative, pe care grupele unei distribuţii homograde sau heterograde le deţin în ansamblul populaţiei reprezentate. În acest scop sunt folosite diferite figuri geometrice împărţite în sectoare ale căror suprafeţe sunt proporţionale cu frecvenţele relative ale grupelor. În raport cu figurile geometrice utilizate se pot delimita mai multe tipuri de diagrame de structură:

a) cercul de structură;
b) pătratul de structură;
c) dreptunghiul de structură.

a) Reprezentarea unei distribuții de frecvențe printr-un **cerc de structură** presupune împărțirea acestuia în mai multe sectoare, fiecare dintre acestea având o suprafață proporțională cu frecvența relativă a unei grupe. După cum se ştie, suprafața unui sector de cerc (S_{SC}) este dată de relația:

$$S_{SC} = \frac{U_{SC}}{360} \times \pi \times r_C^2, \qquad (3.16.)$$

unde U_{SC} este unghiul sectorului de cerc.

În condițiile în care r_C are aceeaşi valoare pentru toate sectoarele, rezultă că suprafețele acestora vor fi diferențiate prin intermediul unghiurilor. Pentru o grupă i, unghiul asociat sectorului acesteia va fi dat de relația:

$$U_{SC_i} = n_{r_i}^x \times 360 \qquad (3.17.)$$

b) Atunci când o distribuție de frecvențe este reprezentată printr-un **pătrat de structură**, acesta este împărțit în mai multe părți, ale căror ponderi în suprafața totală sunt egale cu frecvențele relative ale grupelor.

Reprezentarea este destul de simplă, prin împărțirea unei laturi orizontale sau verticale a pătratului în raport cu frecvențele relative ale grupelor, după formula: $l_i = n_{r_i}^x \times l_P$ (3.18.)

unde l_i este latura porțiunii de pătrat care revine grupei i.

c) Reprezentările distribuțiilor de frecvențe prin **dreptunghiuri de structură** sunt oarecum asemănătoare celor care utilizează pătrate de structură, având avantajul că permit o evidențiere mai bună a ponderii grupelor prin manevrarea raportului dintre lungimile celor două titluri.

3.2.2.3.2. Reprezentări în coordonate carteziene

Reprezentările în coordonate carteziene sunt folosite exclusiv pentru distribuțiile heterograde. Printr-o astfel de tehnică sunt reprezentate pe o axă orizontală intervalele de variație ale grupelor, iar pe o axă verticală frecvențele absolute. În practică sunt folosite trei variante de reprezentări carteziene ale distribuțiilor heterograde:

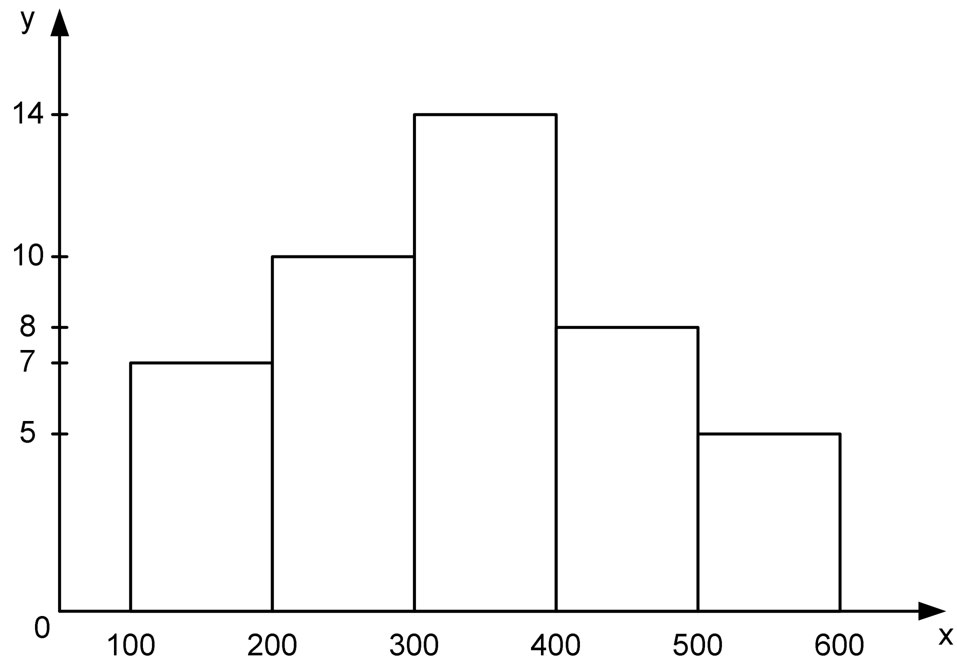
a) histograma;
b) poligonul frecvențelor;
c) curba frecvențelor.

a) O **histogramă** constă în construirea, pentru fiecare grupă dintr-o distribuție heterogradă, a câte unui dreptunghi cu latură orizontală corespunzând intervalului de variație şi cu latura verticală proporțională cu frecvența absolută.

Exemplul 3.5. În tabelul 3.22. este prezentată o distribuție heterogradă ce descrie repartizarea punctelor de desfacere ale unei firme în raport cu veniturile din vânzări. Se cere să se reprezinte aceste date printr-o histogramă.

Tabelul 3.22. Repartizarea punctelor de desfacere ale unei firme în raport cu veniturile din vânzări

Nr. crt.	Interval de variație $(x_{i-1}; x_i)$[mii RON]	Frecvență absolută (n_i^x)
(0)	(1)	(2)
1	[100 ; 200)	7
2	[200 ; 300)	10
3	[300 ; 400)	14
4	[400 ; 500)	8
5	[500 ; 600)	5



Scara: Ox: 1 cm = 100 mii RON; Oy: 1 cm = 2 puncte de desfacere

Fig. 3.4. Reprezentarea printr-o histogramă a distribuției punctelor de vânzare în funcție de venituri

Rezolvare: în figura 3.4. este prezentat modul de trasare a histogramei seriei.

b) Desenarea unui **poligon al frecvențelor** presupune parcurgerea următorului algoritm:

1) la grupele distribuției heterograde se adaugă două, cu frecvența absolută nulă şi cu intervale de variație în continuarea celor de la extremități;
2) pe axa absciselor se trasează intervalele de variație ale grupelor;
3) pentru fiecare grupă se determină câte un punct având drept coordonată orizontală mijlocul intervalului de variație şi drept coordonată verticală frecvența absolută;
4) punctele astfel obținute sunt unite printr-o linie poligonală.

c) O **curbă de frecvențe** se obține trasând o linie curbă prin punctele (sau, cel puțin cât mai aproape de acestea) determinate prin algoritmul folosit pentru desenarea poligonului frecvențelor. Trasarea liniei curbe se poate face fie din ochi, fie prin procedee matematice sofisticate.

Cele trei tehnici prezentate anterior sunt recomandate îndeosebi pentru distribuțiile heterograde cu caracteristici de tip continuu şi cu intervale egale de variație. Atunci când sunt utilizate pentru reprezentarea unor serii cu caracteristici discrete sau cu intervale inegale de variație, aceste procedee pot suferi unele modificări. De exemplu, se recomandă ca în cazul seriilor cu caracteristici discrete să se lase spații libere între reprezentările intervalelor de variație pentru a nu se da impresia continuității. De asemenea, pentru seriile cu intervale de variație inegale se recomandă stabilirea coordonatelor verticale astfel încât suprafețele dreptunghiurilor delimitate de acestea să fie proporționale cu frecvențele relative.

Reprezentările distribuțiilor heterograde în coordonate carteziene sunt utilizate frecvent în aprecierea formei pe care o îmbracă funcția probabilistică asociată unui fenomen. De regulă, în cadrul acestor reprezentări coordonatele de pe ordonate sunt proporționale cu frecvențele absolute şi, implicit, cu cele relative. Cum frecvențele relative constituie nişte aproximări ale probabilităților, rezultă că reprezentările în coordonate carteziene pot fi considerate nişte aproximări ale funcțiilor probabilistice. Forma pe care o are o reprezentare carteziană constituie un indiciu în alegerea tipului de funcție probabilistică asociat fenomenului studiat. În statistica matematică au fost definite mai multe forme ale distribuțiilor heterograde determinate pe baza reprezentărilor în coordonate carteziene. În continuare vom prezenta trei dintre acestea:

1) distribuția în formă de clopot;
2) distribuția în formă de J;
3) distribuția în formă de U.

1) O **distribuție în formă de clopot** corespunde unei repartiții normale de tip Gauss-Laplace. În figura 3.5. este reprezentată grafic o astfel de distribuție. Centrului intervalului valoric al seriei îi corespunde o grupă cu frecvența maximă iar frecvențele celorlalte grupe se diminuează în raport cu aceasta, spre stânga şi spre dreapta, cu un acelaşi ritm, într-o simetrie perfectă. Se poate aprecia că

această formă de distribuție descrie manifestarea în condiții naturale a celor mai multe dintre fenomenele colective. Pentru distribuțiile în formă de clopot se consideră că valorile tipice, situate în intervalul din centru cu frecvența absolută maximă, au un grad mare de reprezentativitate pentru ansamblul seriei, ceea ce uşurează caracterizarea fenomenelor.

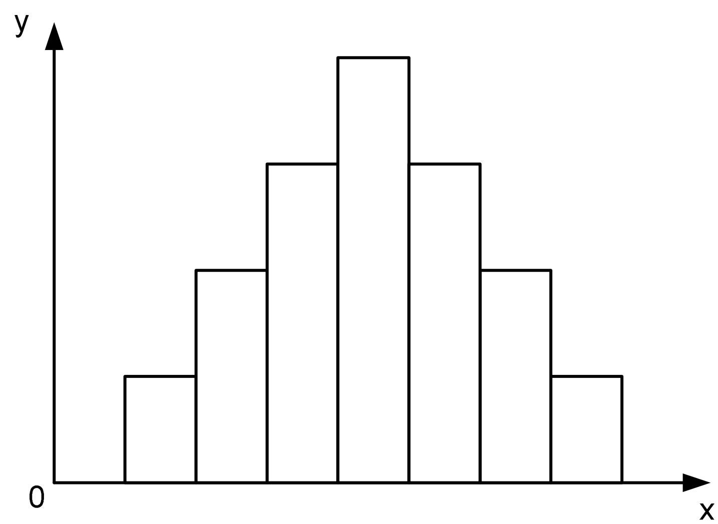


Fig. 3.5. Histograma unei distribuții în formă de clopot

2) O **distribuție în formă de J** (numită şi *curba lui Pareto*, în onoarea economistului Vilfredo Pareto) este caracterizată prin dispunerea frecvenței maxime într-unul din intervalele de variație extreme, frecvența celorlalte grupe scăzând treptat şi atingând un minim la cealaltă extremitate (fig. 3.6.). Astfel de situații apar îndeosebi atunci când se studiază distribuțiile averilor sau veniturilor în cadrul unor comunități polarizate sub aspectul bogăției, în care o mare parte a familiilor se situează în intervale valorice inferioare ale veniturilor sau averilor. La distribuțiile în formă de J se consideră că valorile tipice nu au o reprezentativitate prea mare, ceea ce induce unele dificultăți în caracterizarea fenomenelor studiate.

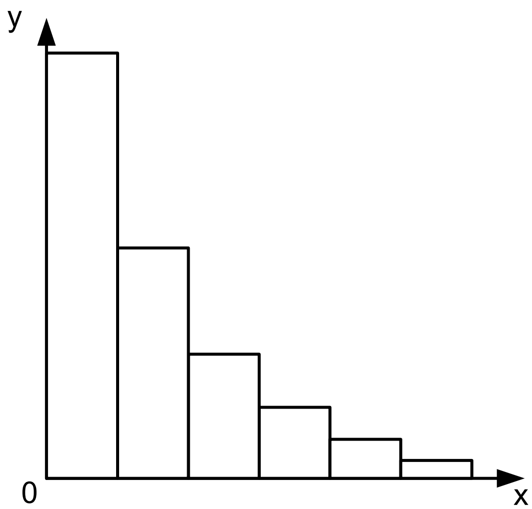


Fig. 3.7. Histograma unei distribuții în formă de J

3) O **distribuție în formă de U** poate fi descrisă drept opusul unei distribuții în formă de clopot. În figura 3.7. este prezentată histograma unei astfel de distribuții. În centrul intervalului valoric al seriei se află o grupă cu frecvența minimă, frecvențele celorlalte grupe crescând treptat în raport cu aceasta, spre stânga şi spre dreapta, într-o simetrie perfectă. Astfel de distribuții se întâlnesc destul de rar, în studiul unor fenomene meteorologice, biologice etc. Se consideră că valorile tipice ale unei distribuții în formă de U nu au un grad mare de reprezentativitate pentru ansamblul seriei, ceea ce face dificil studiul fenomenului.

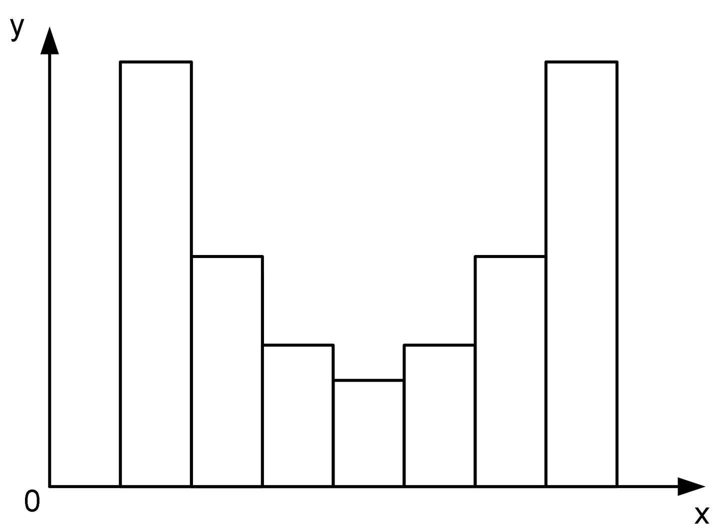


Fig. 3.7. Histograma unei distribuții în formă de U

Cele trei forme prezentate mai sus sunt, într-o anumită măsură, nişte abstractizări, care apar, în realitate, destul de rar într-o „formă pură”. Adeseori, în practică, o distribuție heterogradă este încadrată în una dintre aceste forme abstracte, cu toate că nu îi întruneşte toate însuşirile. De exemplu, sunt considerate drept distribuții în formă de clopot sau în formă de U serii care nu sunt perfect simetrice, sau drept distribuții în formă de J serii la care frecvențele nu cresc sau descresc continuu.

3.3. Prelucrarea primară a datelor statistice prin serii în spațiu

O serie în spațiu este un şir de date asupra unui fenomen, diferențiate pe baza locurilor în care acesta s-a manifestat. În practică, sunt întâlnite destul de frecvent fenomene care se manifestă diferit în locuri diferite. De exemplu, un produs nou lansat poate fi primit foarte bine în unele regiuni şi mai puțin bine în alte regiuni. Printr-o serie în spațiu, astfel de diferențe pot fi relevate şi puse în legătură cu unele circumstanțe care le-au favorizat. Alcătuirea seriilor în spațiu este condiționată de posibilitatea de obținere a datelor simultan din mai multe locuri. În cazul în care datele asupra manifestării unui fenomen în spațiu sunt culese în perioade de timp diferite, comparațiile dintre acestea pentru relevarea diferențelor îşi pierd din rigoare. Datele pe baza cărora se constituie seriile în spațiu pot îmbrăca variate forme: cantitative, calitative, cronologice etc. Aceste serii pot fi atât unidimensionale cât şi multidimensionale. Într-o anumită măsură, o serie în spațiu poate fi considerată drept o distribuție homogradă, în care locul joacă rolul unei caracteristici atributive calitative.

Pentru reprezentările grafice ale seriilor în spațiu pot fi utilizate atât tehnicile de reprezentare specifice distribuțiilor homograde (diagrame de reprezentare a frecvențelor absolute, diagrame de structură etc.) cât şi o tehnică specială, numită *cartogramă*. Aceasta constă în reprezentarea unor aspecte ale seriilor în spațiu prin intermediul unor hărți geografice (la nevoie stilizate), în care sunt evidențiate locurile pentru care s-au cules date. Adeseori, frecvențele absolute sunt descrise prin *pictograme* – simboluri ale datelor prezentate.

3.4. Prelucrarea primară a datelor statistice prin serii în timp

3.4.1. Conceptul de serie în timp

O serie în timp (numită şi *serie cronologică*) poate fi definită drept un şir prin care sunt prezentate date cu privire la stările unui fenomen în diferite momente sau perioade de timp ale manifestării sale. Şirul este ordonat, de regulă, în ordinea cronologică a momentelor şi perioadelor de timp. Seriile în timp sunt practic indispensabile pentru analiza dinamică prin care sunt studiaţi parametrii unor evoluții. Datele prezentate printr-o serie în timp pot îmbrăca diferite forme:date calitative, date cantitative de tip discret sau continuu, date asupra locurilor în care se manifestă fenomenul studiat etc.

În alcătuirea unei serii în timp trebuie luate în considerare momentele şi perioadele de timp pentru care se culeg datele precum şi modalitățile de prezentare a acestora. În funcție de scopurile analizei dinamice, orizontul de timp pentru care se culeg datele statistice poate lua diverse valori: de la câteva secunde, aşa cum se întâmplă când se studiază anumite procese fizice sau chimice, până la mai multe decenii, atunci când se cercetează unele evoluții în domeniul social sau economic. De regulă, cu cât orizontul de timp pentru care se culeg datele este mai mare, cu atât prelucrarea acestora este mai complexă. De exemplu, pentru datele exprimate în unități monetare şi care privesc perioade lungi de timp trebuie luată în considerare diminuarea puterii de cumpărare a banilor, ca urmare a inflației, astfel încât valorile acestora sunt prelucrate pentru a permite comparațiile în raport cu o aceeaşi valoare a banilor.

Momentele şi perioadele de timp în care se culeg date sunt stabilite, de regulă, prin împărțirea orizontului de timp în intervale cu lungimi egale (sau, cel puțin, aproximativ egale). Lungimea acestor intervale este aleasă în raport cu unele trăsături ale fenomenului studiat (regularitate, durată etc.) şi cu posibilitățile de culegere a datelor. De exemplu, în analiza economică, momentele pentru care sunt culese datele sunt stabilite adeseori la sfârşit de an, întrucât bilanțurile contabile ale firmelor se alcătuiesc, de regulă, pe baza situației din această perioadă.

Uneori, prin seriile în timp sunt studiate nu doar tendințele generale ale unor evoluții ci şi variațiile periodice ale fenomenelor cercetate. Astfel de variații periodice se produc ca urmare a unor factori ce acționează semnificativ numai în anumite perioade de timp: în unele anotimpuri ale unui an, în unele zile ale săptămânii, în anumite ore dintr-o zi etc. De exemplu, în cursul unui an, vânzările de înghețată înregistrează, de regulă, valori maxime în lunile de vară şi valori minime în lunile de iarnă. De asemenea, în cursul unei săptămâni, vânzările de bilete la cinematografe sunt, de regulă, mai mari în zilele de sâmbătă şi duminică față de celelalte zile. Variațiile periodice ale unui fenomen sunt cercetate, de regulă, pe baza unor date culese la intervale de timp determinate prin divizarea celor alese pentru obținerea datelor asupra tendinței generale.

În funcție de modul de prezentare a datelor se pot delimita două tipuri de serii în timp:

- *serii în timp simple*, la care datele reflectă situația unui fenomen în momentul sau perioada de timp pentru care au fost culese;
- *seriile în timp cumulate*, la care datele reflectă situația unui fenomen până la momentul sau perioada pentru care au fost culese.

În domeniul economic, seriile în timp cumulative sunt folosite îndeosebi pentru a se evidenția realizarea unor planificări: pentru venituri sau cheltuieli bugetare, pentru nivelul producției etc. O serie în timp cumulată poate fi obținută dintr-o serie în timp simplă, adunând, la valoarea numerică a datei pentru un anumit moment, valorile numerice ale datelor pentru momentele anterioare. De exemplu, în tabelul 3.30. este prezentată situația îndeplinirii planului pentru primul semestru al anului 2005 la producția unui sortiment, pe baza a trei indicatori: producția efectivă lunară, producția efectivă cumulată pentru fiecare lună şi ponderea producției efective cumulate pentru producția planificată pentru întregul semestru.

Tabelul 3.30. Situația îndeplinirii planului pentru producția unui sortiment, în primele şase luni ale anului 2005

Nr. crt.	Luna	Producția efectivă [bucăți]		Pondere a producției efective cumulată în producția planificată pentru întregul semestru[1]
		Lunară	Cumulată de la 1 ianuarie 2005	

(0)	(1)	(2)	(3)	(4) = (3)/10.000
1	Ianuarie	1 900	1 900	0,19
2	Februarie	1 100	3 000	0,30
3	Martie	2 200	4 200	0,42
4	Aprilie	2 500	6 700	0,67
5	Mai	2 700	8 400	0,84
6	Iunie	1 400	9 800	0,98

1) Pentru întregul semestru al anului 2005 a fost planificată o producție de 10 000 bucăți

3.4.2. Reprezentarea grafică a seriilor în timp

Dintre tehnicile aplicat în reprezentarea grafică a seriilor în timp se detaşează, prin frecvența utilizării, două categorii:

- reprezentările în coordonate carteziene;
- reprezentările prin diagrame polare.

3.4.2.1. Reprezentările grafice ale seriilor în timp prin coordonate carteziene

Reprezentările grafice ale seriilor în timp prin coordonate carteziene (numite şi *historiograme* sau *cronograme*) sunt asemănătoare celor folosite pentru reprezentarea seriilor atributive sau seriilor de loc, cu deosebirea că pe axa absciselor coordonatele corespund unor momente sau perioade de timp. Distanțele dintre reprezentările momentelor sau perioadelor de timp pe axa absciselor trebuie să fie proporționale (sau, cel puțin, aproximativ proporționale) cu intervalele de timp dintre acestea . Pe axa ordonatelor sunt reprezentate, la o scară convenabilă, valorile datelor culese. Fiecărei perechi de date şi momente (sau perioade) de timp îi corespunde un punct obținut prin intersectarea dreptelor trasate perpendicular pe cele două axe, în dreptul valorilor corespunzătoare. Există mai multe variante de reprezentare grafică a seriilor în timp prin coordonate carteziene:

- *historiograma prin linii drepte*, care unesc punctele corespunzătoare coordonatelor datelor culese şi momentelor (sau perioadelor) de timp;
- *historiograma prin linii curbe*, care unesc aceleaşi puncte (acest tip de reprezentare este mai greu de realizat însă

sugerează într-o măsură mai mare decât precedentul continuitatea evoluției de la o perioadă de timp la alta);
- *historiograma prin bare,* care constă în reprezentarea datelor prin dreptunghiuri cu latura verticală egală cu coordonatele de pe abscisă şi cu latura orizontală, cu o mărime convenabilă, poziționată în dreptul coordonatelor de pe abscisă.

3.4.2.2. Reprezentările grafice ale seriilor în timp prin diagrame polare

Reprezentările grafice prin diagrame polare îmbracă forma unor sectoare de cerc, concentrice dar cu raze diferite. Sunt utilizate două forme de reprezentare a valorilor prin diagrame polare:
- prin proporționalitatea dintre suprafețele sectoarelor de cerc şi valorile pe care acesta le reprezintă;
- prin proporționalitatea dintre razele sectoarelor de cerc şi valorile pe care acestea le reprezintă.

În practică, este folosită mai frecvent a doua formă de reprezentare, care este mai facilă. De regulă, unghiurile sectoarelor de cerc sunt proporționale cu intervalele de timp la care se referă. În consecință, atunci când aceste intervale sunt aproximativ egale, unghiul unui sector de cerc este obținut prin împărțirea celor 360^{o} ale unui cerc la numărul de intervale de timp.

Spre deosebire de historiograme, care sugerează, în mare măsură, un traseu parcurs de la început până la sfârşit, diagramele polare, la care nu se face o delimitare clară între un moment de început şi unul de sfârşit, sugerează, mai degrabă, un proces reluat permanent. Sunt, din acest motiv, indicate pentru studiul variațiilor periodice ale unor fenomene.

Capitolul 4 – Valori tipice

4.1. Considerații generale asupra valorilor tipice

Într-un capitol anterior am definit valorile tipice drept mărimi reprezentative pentru caracteristicile unei populații statistice. În cadrul cercetărilor statistice, aceste mărimi servesc la identificarea trăsăturilor esențiale ale fenomenelor colective.

Valorile tipice îmbracă o formă numerică, ceea ce constituie un avantaj considerabil din perspectiva cuantificării acestor trăsături. Totuşi, tocmai această însuşire duce la unele constrângeri în folosirea lor. Dacă în ce priveşte datele cantitative, valorile tipice sunt destul de uşor de identificat, în cazul datelor calitative e nevoie de procedee destul de complexe pentru a le transpune într-o formă numerică.

În cercetările statistice sunt folosite mai multe categorii de mărimi prin care să fie reprezentat ansamblul unităților unei populații statistice. În acest capitol vor fi abordate doar trei tipuri de mărimi dintre cele utilizate frecvent în practică:

- *mărimile medii*, care sunt obținute raportând toate valorile unei serii la numărul unităților statistice;
- *valoarea mediană*, calculată în raport cu poziția centrală dintr-o serie ordonată;
- *modul* (numit şi *dominanta*) calculat în raport cu frecvența maximă dintr-o distribuție heterogradă.

În mod obligatoriu, valorile tipice sunt determinate pe baza seriilor statistice. Din acest motiv, modalitățile de calcul ale acestor mărimi trebuie adaptate la tipurile seriilor statistice. În cazul seriilor simple, valorile tipice sunt determinate în raport cu numărul şi valorile asociate unităților statistice. În schimb, pentru distribuțiile heterograde valorile tipice sunt calculate pe baza intervalelor de variație şi frecvențelor asociate grupelor.

Un alt aspect important al valorilor tipice, abordat şi în capitolele anterioare, este constituit de reprezentativitatea pe care o astfel de mărime o are pentru ansamblul populației statistice pe care o caracterizează. În cadrul statisticii matematice au fost dezvoltate mai multe criterii de apreciere a reprezentativității valorilor tipice în raport cu particularitățile seriilor statistice. Pe baza acestora se poate aprecia, pentru o serie statistică anume, care sunt mărimile care îi caracterizează aspectele esențiale.

4.2. Mărimi medii

Mărimile medii sunt considerate drept indicatorii care reflectă în cea mai mare măsură impactul factorilor esențiali de influență asupra fenomenelor colective. În acest subcapitol vor fi prezentate succint patru categorii de mărimi medii:

- media aritmetică;
- media geometrică;
- media armonică;
- mediile de ordin superior.

4.2.1. Media aritmetică

4.2.1.1. Calculul mediilor aritmetice

În raport cu tipurile seriilor statistice se pot delimita două modalități de calcul al mediei aritmetice:

a) modalități specifice seriilor simple;
b) modalități specifice distribuțiilor heterograde.

a) Pentru o **serie simplă**, media aritmetică este obținută raportând totalul valorilor la numărul de unități statistice. În acest caz, formula de calcul are forma: $\overline{X}_0 = \frac{\sum_{i=1}^{N} X_i}{N}$ (4.1.)

în care: $\overline{X}_0$ este media aritmetică, după o caracteristică a seriei simple; X_i este valoarea caracteristicii X asociată unității statistice i; N este numărul de unități statistice din cadrul seriei simple.

b) Pentru o **distribuție heterogradă**, calculul mediei aritmetice are la bază intervalele de variație şi frecvențele asociate grupelor. Formula de calcul este următoarea:

$$\overline{X} = \frac{\sum_{i=1}^{K_x} X_i^{'} \cdot n_i^X}{\sum_{i=1}^{K_x} n_i^X} \qquad (4.2.)$$

unde:

- $\overline{X}$ este media aritmetică a distribuției heterograde în raport cu caracteristica *X*;
- K_X este numărul de grupe al seriei în raport cu caracteristica *X*;
- $X_i^{'}$ este centrul de interval al grupei *i* formată după caracteristica *X*;
- n_i^X este frecvența absolută a grupei *i*.

Exemplul 4.1. În tabelul 4.1. este prezentată o distribuție heterogradă, care descrie repartizarea angajaților unei firme în raport cu veniturile salariale ale acestora. Se cere să se determine venitul salarial mediu pentru angajații firmei.

Tabelul 4.1. Repartizarea angajaților unei firme în raport cu veniturile salariale

Nr. crt.	Grupe după venituri salariale [RON/lună]	Frecvență absolută (n_i^X)
(0)	(1)	(2)
1	[300 – 500)	20
2	[500 – 700)	50
3	[700 – 900)	80
4	[900 – 1.100)	40
5	[1 100 – 1 300)	10

Rezolvare:

În tabelul 4.2. este prezentat calculul termenilor formulei 4.2. În raport cu aceştia rezultă:

$$\overline{X} = \frac{\sum_{i=1}^{K_x} X_i^{'} \cdot n_i^X}{\sum_{i=1}^{K_x} n_i^X} = \frac{154.000}{200} = 770 \text{ RON/lună}$$

Tabelul 4.2. Calcule intermediare pentru determinarea mediei aritmetice

Nr. crt.	Grupe după venituri salariale [RON/lună]	Frecvență absolută (n_i^X)	Centru de interval (X') [RON/lună]	$X' \times n_i^X$ [RON/lună]
(0)	(1)	(2)	(3)	(4)=(3)x(2)
1	[300 – 500)	20	400	8.000
2	[500 – 700)	50	600	30.000
3	[700 – 900)	80	800	64.000
4	[900 – 1.100)	40	1.000	40.000
5	[1.100 – 1.300)	10	1.200	12.000
6	Total	200	×	154.000
7	Simbol pentru total	$\sum_{i=1}^{K_x} n_i^x$	×	$\sum_{i=1}^{K_x} X_i^{'} \cdot n_i^X$

Deşi diferită de modalitatea de calcul pentru seriile simple, formula de determinare a mediei aritmetice pentru distribuțiile heterograde are la bază, la fel ca prima, raportarea sumei valorilor la numărul total de unități statistice. Pentru a demonstra aceasta, vom reaminti ceea ce am menționat într-un capitol anterior, anume că în unele calcule statistice se consideră că toate unitățile statistice dintr-o grupă au o valoare egală cu cea a centrului intervalului de variație. În aceste condiții, suma valorilor din acea grupă este dată de produsul dintre numărul de unități statistice (adică frecvența absolută a grupei) şi centrul intervalului de variație. Rezultă că suma tuturor valorilor seriilor, care poate fi obținută adunând sumele valorilor din toate grupele seriei, este reprezentată de numărătorul din relația (4.2.).

Pe de altă parte, numărul total de unități statistice ale unei distribuții heterograde poate fi obținut adunând frecvențele absolute ale tuturor grupelor (altfel spus, se însumează toate unitățile, din fiecare grupă), ceea ce reprezintă valoarea numitorului din relația (4.2.).

În concluzie, relația (4.2.) prin care se calculează mediile aritmetice ale distribuțiilor heterograde, poate fi considerată drept un raport dintre suma valorilor şi numărul de unități statistice.

Media aritmetică a unei serii simple şi cea a unei distribuții heterograde pot diferi substanțial atunci când aproximarea valorilor unei grupe prin centrul de interval al acesteia este mult îndepărtată de realitate. De regulă, cu cât numărul de grupe este mai mare cu atât diferența dintre cele două medii aritmetice este mai mică.

4.2.1.2. Reprezentativitatea mediilor aritmetice

Media aritmetică este considerată drept cea mai reprezentativă valoare pentru impactul factorilor esențiali de influență asupra unui fenomen ce se manifestă în condiții de normalitate. Adeseori, fenomenele sunt comparate şi încadrate doar pe baza acestei valori tipice. Totuşi, o analiză care nu foloseşte decât media aritmetică are dezavantajul că lasă nesesizat aspectul omogenității manifestării fenomenelor colective. De exemplu, două grupe de studenți pot să fie caracterizate în raport cu rezultatul la un examen printr-o aceeaşi notă medie egală cu şapte, obținută însă în condiții diferite. Să presupunem că la prima grupă toți studenții au obținut nota şapte, ceea ce înseamnă o omogenitate perfectă. În schimb, să presupunem pentru a doua grupă că jumătate din efectiv a obținut nota zece în timp ce cealaltă jumătate a obținut nota patru, ceea ce înseamnă o dispersare semnificativă a valorilor. În primul caz, media aritmetică se confundă cu notele, fiind, astfel, foarte reprezentativă pentru acestea. În al doilea caz, notele sunt destul de îndepărtate de media aritmetică, ceea ce face ca aceasta să fie mai puțin reprezentativă pentru studenții grupei. Acest exemplu a vizat valori organizate în serii simple. Pentru distribuțiile de frecvențe, situația este ceva mai complexă întrucât trebuie luată în considerare atât dispersarea valorilor din cadrul fiecărei grupe cât şi dispersarea centrelor intervalelor de variație. Cu cât valorile din cadrul unei grupe sunt mai dispersate, cu atât centrul de interval este mai puțin reprezentativ pentru acestea. De asemenea, o dispersare semnificativă a centrelor intervalelor de variație face ca media aritmetică a distribuției heterograde să fie mai puțin apropiată de aceste valori.

În aprecierea reprezentativității unei medii aritmetice pentru o distribuție heterogradă poate fi luată în considerare şi forma acesteia din urmă. Astfel, la distribuțiile în formă de clopot se consideră că media aritmetică, situată în intervalul cu cea mai mare frecvență, are un grad mare de reprezentativitate pentru valorile seriei. În schimb, pentru distribuțiile în formă de J sau de U, media aritmetică, amplasată nu neapărat într-un interval de frecvență maximă, are, de regulă, un grad redus de reprezentativitate.

4.2.2. Media geometrică

Media geometrică este o mărime folosită pentru a caracteriza aspectele esențiale ale unui fenomen ale cărui efecte pot fi asimilate unei progresii geometrice. Astfel de situații apar îndeosebi în cazul evoluțiilor schimburilor comerciale internaționale pentru anumite perioade, a vânzărilor unor produse în faza de lansare, a unor fenomene demografice etc. Se consideră că la astfel de evoluții media geometrică poate surprinde, uneori chiar într-o măsură mai mare față de media aritmetică, aspectele esențiale.

Media geometrică a unei serii simple notată cu $\overline{X}_{g_0}$, este dată de formula:

$$\overline{X}_{g_0} = \sqrt[N]{\prod_{i=1}^{N} X_i} \qquad (4.3.)$$

În practică, atunci când N este foarte mare, extragerea unei rădăcini de un asemenea ordin poate fi destul de complicată. Din acest motiv, adeseori se preferă logaritmarea relației (4.3.) care devine:

$$\ln\left(\overline{X}_{g_0}\right) = \ln\left(\sqrt[N]{\prod_{i=1}^{N} X_i}\right) = \frac{1}{N}\ln\left(\prod_{i=1}^{N} X_i\right) = \frac{1}{N}\sum_{i=1}^{K_X} \ln x_i \qquad (4.4.)$$

Pentru o distribuție heterogradă, dacă se consideră că toate unitățile dintr-o grupă au o valoare egală cu centrul intervalului de variație, media geometrică, notată cu $\overline{X}_g$, este dată de relația:

$$\overline{X}_g = \sqrt[\sum_{i=1}^{K_X} n_i^X]{\prod_{i=1}^{K_X} X_i'^{\,n_i^x}} \qquad (4.5.)$$

Logaritmând această valoare, din aceleaşi considerente pentru care se logaritmează şi media geometrică a unei serii simple, rezultă:

$$\ln\left(\overline{X}_g\right) = \ln\left(\sqrt[\sum_{i=1}^{K_X} n_i^X]{\prod_{i=1}^{K_X} X_i'^{\,n_i^x}}\right) = \frac{1}{\sum_{i=1}^{K_X} n_i^X}\ln\left(\prod X_i'^{\,n_i^x}\right) = \frac{1}{\sum_{i=1}^{K_X} n_i^X}\sum_{i=1}^{K_X}\left(n_i^X \cdot \ln X_i'\right) \qquad (4.6.)$$

4.2.3. Media armonică

Media armonică este un indicator folosit pentru a descrie fenomene ale căror efecte pot fi asimilate unei funcții hiperbolice.

Pentru o serie simplă, media armonică, notată cu $\overline{X}_{h_0}$, este dată de relația:

$$\overline{X}_{h_0} = \frac{N}{\sum_{i=1}^{N} \frac{1}{X_i}} \tag{4.7.}$$

Media armonică a unei distribuții heterograde, notată cu $\overline{X}_h$, poate fi calculată prin formula:

$$\overline{X}_{h_0} = \frac{\sum_{i=1}^{K_X} n_i^X}{\sum_{i=1}^{K_X} \frac{1}{X_i^{'}} \cdot n_i^X} \tag{4.8.}$$

4.2.4. Medii de ordin superior

O medie de ordin superior este indicată pentru a caracteriza aspectele esențiale ale unor fenomene ale căror efecte pot fi asimilate unor funcții polinomiale.

Pentru o serie simplă, o medie de ordin *p*, notată cu $\overline{X}_{p_0}$, este dată de relația:

$$\overline{X}_{p_0} = \sqrt[p]{\frac{\sum X_i^p}{N}} \tag{4.9.}$$

Media de ordin *p* a unei distribuții heterograde, notată cu $\overline{X}_p$, poate fi calculată prin formula:

$$\overline{X}_p = \sqrt[p]{\frac{\sum_{i=1}^{K_X} X_i^p \cdot n_i^X}{\sum_{i=1}^{K_X} n_i^X}} \tag{4.10}$$

La fel ca în cazul mediilor geometrice, uneori, pentru simplificarea calculelor, se procedează la logaritmarea formulelor mediilor de ordin superior.

4.3. Valoarea mediană

O valoare mediană (numită uneori, mai simplu, doar *mediană*) este o mărime ce ocupă locul central într-o serie statistică ordonată împărțind-o în două grupe de frecvențe egale.

4.3.1. Determinarea valorii mediane

Modalitățile de determinare a valorii mediane se diferențiază în raport cu tipul seriei: simplă sau distribuție heterogradă.

4.3.1.1. Calculul valorii mediane pentru serii simple

În cazul unei serii simple ordonate, valoarea mediană, notată cu Me^x_o, este reprezentată, aşa cum reiese din definiția acestei mărimi, de termenul (sau termenii) care ocupă locul central. Atunci când seria are un număr impar de unități, valoarea mediană este uşor de determinat, întrucât un singur termen deține poziția centrală. În schimb, atunci când seria are un număr par de unități, în mijlocul acesteia se vor afla doi termeni, iar valoarea mediană va fi dată de media aritmetică a acestora.

Exemplul 4.2. Se dau două serii simple:

- prima serie cuprinde numerele: 50; 42; 48; 38; 41;
- a doua serie cuprinde numerele:47; 61; 63; 62; 34; 37.

Se cere să se determine valorile mediane ale celor două serii.

Rezolvare: Prin ordonare, prima serie devine: 38; 41; 42; 48; 50. Fiind o serie cu un număr impar de termeni, valoarea mediană este reprezentată de termenul aflat în centru, adică $Me^x_o = 42$. Tot prin ordonare, a doua serie devine: 34; 37; 47; 61; 62; 63. Fiind o serie cu un număr par de termeni, valoarea mediană este dată de media aritmetică a celor doi termeni ce ocupă poziția centrală, adică:

$$M_{l_0}^{y} = \frac{y_3 + y_4}{2} = \frac{47+61}{2} = 54\,.$$

4.3.1.2. Determinarea valorii mediane pentru distribuții heterograde

Pentru distribuțiile heterograde, valorile mediane pot fi determinate prin două modalități:

- pe cale analitică;
- pe cale grafică.

4.3.1.2.1. Calculul analitic al valorii mediane a unei distribuții heterograde

La o distribuție heterogradă determinarea valorii mediane, notată cu M_e^x, presupune parcurgerea următorului algoritm:

Pasul 1. Se calculează o mărime numită *unitatea mediană a seriei*, notată cu $U^{Me}{}_x$, prin formula:

$$U_x^{M_e} = \frac{\left(\sum_{i=1}^{K_x} n_i^x\right) + 1}{2} \qquad (4.11.)$$

Pasul 2. Se calculează, pentru fiecare grupă, o mărime numită *frecvența absolută cumulată*, notată cu N_{x_i} prin adunarea, la frecvența absolută a grupei, a frecvențelor absolute ale grupelor anterioare:

$$N_{x_i} = \sum_{j=1}^{i} n_j^x$$

(4.12.)

Pasul 3. Se stabileşte intervalul de variație în care se găseşte valoarea mediană, numit *interval median*, care corespunde primei grupe pentru care frecvența absolută cumulată este mai mare decât unitatea mediană;

Pasul 4. Se calculează valoarea mediană prin formula:

$$M_e^x = x_{M_e - 1} + d_x^{M_e} \cdot \frac{U^{M_e} - N_{x_{M_{e-1}}}}{n_{M_e}^x} \qquad (4.13.)$$

unde:

$x_{M_e - 1}$ este limita inferioară a intervalului median;

$d_x^{M_e}$ este lungimea intervalului median;

$N_{x_{M_{e-1}}}$ este frecvența absolută cumulată a intervalului anterior intervalului median;

$n_{M_e}^x$ este frecvența absolută a intervalului median.

Exemplul 4.3. Se cere să se calculeze pe cale analitică valoarea mediană a seriei prezentate în tabelul 4.1.

Rezolvare: În tabelul 4.3. sunt prezentate valorile frecvențelor absolute cumulate.

Unitatea mediană reprezintă:

$$U_x^{M_e} = \frac{\left(\sum_{i=1}^{K_x} n_i^x\right)+1}{2} = \frac{200+1}{2} = 100{,}5$$

Prima grupă pentru care $N_{x_i} > U_x^{M_e}$ corespunde intervalului [700 , 900) care a fost, astfel, desemnat drept interval median.

Tabelul 4.3. Calcule intermediare pentru determinarea valorii mediane

Nr. crt.	Grupe de venituri salariale [RON/lună]	Frecvență absolută (n_i^x)	Frecvență absolută cumulată (N_i)
(0)	(1)	(2)	(3)
1	[300 – 500)	20	20
2	[500 – 700)	50	70
3	[700 – 900)	80	150
4	[900 – 1.100)	40	190
5	[1.100 – 1.300)	10	200
6	Total	200	×
7	Simbol pentru total	$\sum_{i=1}^{K_x} n_i^x$	×

Se determină, în continuare, valoarea mediană:

$$M_e^x = x_{M_e - 1} + d_x^{M_e} \cdot \frac{U^{M_e} - N_{x_{M_e - 1}}}{n_{M_e}^x} = 700 + 200\frac{100{,}5-70}{80} = 776{,}25$$

RON/lună

4.3.1.2.2. Determinarea pe cale grafică a valorii mediane pentru o distribuție heterogradă

Pentru determinarea valorilor mediane ale distribuțiilor heterograde sunt folosite în practică mai multe tehnici grafice. În acest subcapitol vom prezenta doar una dintre acestea, utilizată destul de frecvent datorită simplității sale. Tehnica are la bază următorul algoritm:

Pasul 1. Într-un sistem de coordonate carteziene sunt prezentate intervalele de variație, pe axa absciselor, şi frecvențele absolute cumulate, pe axa ordonatelor.

Pasul 2. Sunt trasate puncte care reflectă relația dintre intervalele de variație şi frecvențele absolute cumulate astfel:

- limitei superioare a unui interval de variație îi corespunde frecvența absolută cumulată a grupei;
- limitei inferioare a aceluiaşi interval de variație îi corespunde frecvența absolută cumulată a grupei anterioare (face excepție prima grupă, la care limitei inferioare a intervalului de variație îi va corespunde o valoare nulă pe axa ordonatelor).

Pasul 3. Punctele trasate anterior sunt unite printr-o linie dreaptă poligonală, rezultând astfel o reprezentare grafică numită *ogivă*.

Pasul 4. Se trasează o linie dreaptă perpendiculară pe axa ordonatelor în dreptul coordonatei care reprezintă jumătate din frecvența absolută cumulată a ultimei grupe.

Pasul 5. La intersecția dintre ogivă şi linia dreaptă trasată anterior se coboară o perpendiculară pe axa absciselor, pe care o va intersecta într-un punct ce corespunde valorii mediane a seriei.

Exemplul 4.4. Se cere să se determine pe cale grafică valoarea mediană a seriei prezentate în tabelul 4.1.

Rezolvare: Pe baza intervalelor de variație şi a frecvențelor absolute cumulate, acestea din urmă determinate în tabelul 4.3., a fost desenată, în figura 4.1., ogiva distribuției heterograde. A fost trasată apoi o linie dreaptă perpendiculară pe axa ordonatelor, în dreptul coordonatei 100, care reprezintă jumătate din frecvența absolută cumulată a ultimei grupe. Această linie dreaptă a intersectat ogiva în punctul notat cu *M*, de la care s-a coborât o perpendiculară pe axa absciselor pe care a întâlnit-o într-un punct ce corespunde valorii mediane.

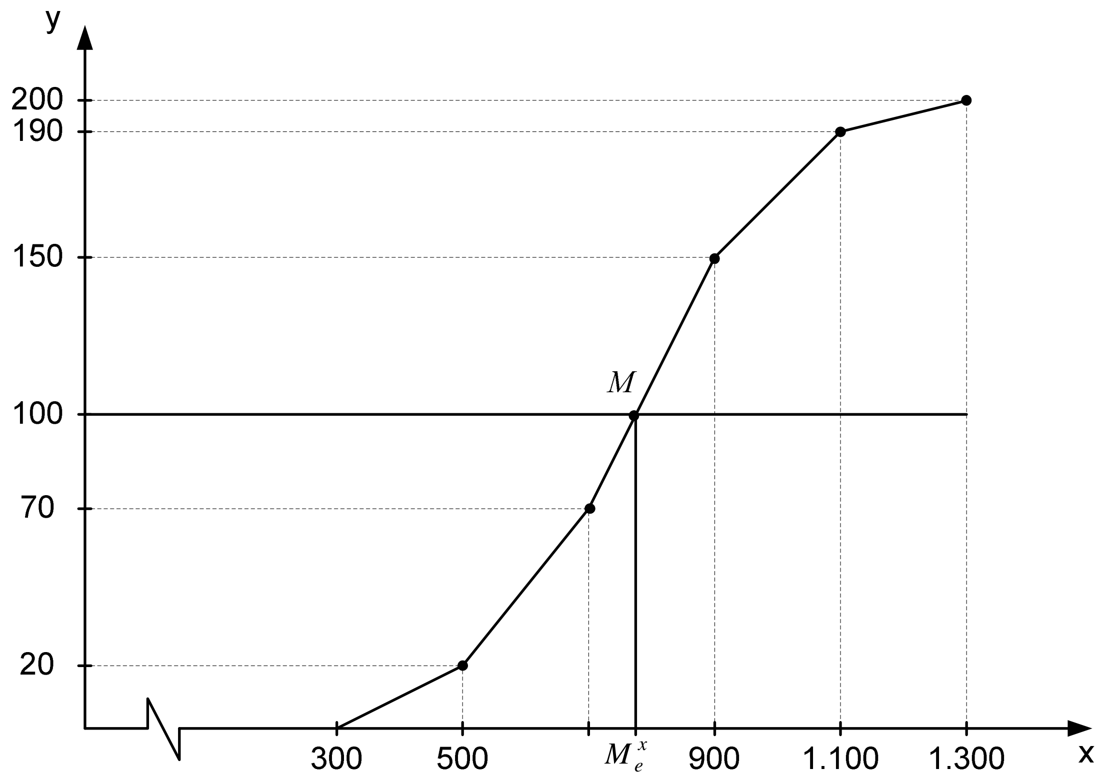


Fig. 4.1. Determinarea pe cale grafică a valorii mediane a unei distribuții heterograde

4.3.2. Utilizarea valorilor mediane în caracterizarea fenomenelor colective

O mărime care împarte o serie statistică ordonată în două grupe de frecvențe egale are semnificația unui nivel mijlociu pentru ansamblul valorilor seriei. Cu toate acestea, mediana reflectă, în comparație cu media aritmetică, într-o măsură mult mai mică trăsăturile esențiale ale fenomenelor colective. În consecință, valoarea mediană este folosită mai degrabă pentru a completa caracterizările făcute prin intermediul valorilor medii, mai ales când acestea nu sunt foarte reprezentative pentru fenomenele studiate.

O valoare mediană este foarte apropiată de media aritmetică atunci când seria statistică este dispusă relativ simetric. În cazul unei simetrii perfecte, media aritmetică împarte în două seria ordonată, confundându-se, în fapt, cu valoarea mediană. După cum se va vedea într-un capitol ulterior, relația dintre valoarea mediană şi media aritmetică este utilizată în aprecierea gradului de reprezentativitate al valorilor tipice.

4.4. Modul unei distribuții heterograde

Modul unei distribuții heterograde (numit şi *dominantă*) este o mărime care exprimă valoarea cu cea mai mare frecvență din cadrul seriei.

4.4.1. Determinarea modului unei distribuții heterograde

Se consideră că modul unei distribuții heterograde trebuie să se afle în interiorul unui interval cu frecvența mai mare decât cea a intervalelor învecinate. Un astfel de interval este numit *interval modal*. În raport cu situația intervalelor modale se pot delimita trei tipuri de distribuții heterograde:

- *serii unimodale*, care au doar câte un interval modal (fig. 4.2.a.);
- *serii plurimodale cu un singur interval modal principal*, care au mai multe intervale modale însă dintre acestea doar unul, numit *principal*, are frecvența absolută maximă, celelalte intervale modale fiind numite *secundare* (fig. 4.3.b.);
- *serii plurimodale cu mai multe intervale modale principale*, care au mai multe intervale modale cu frecvența absolută maximă (fig. 4.2.c.).

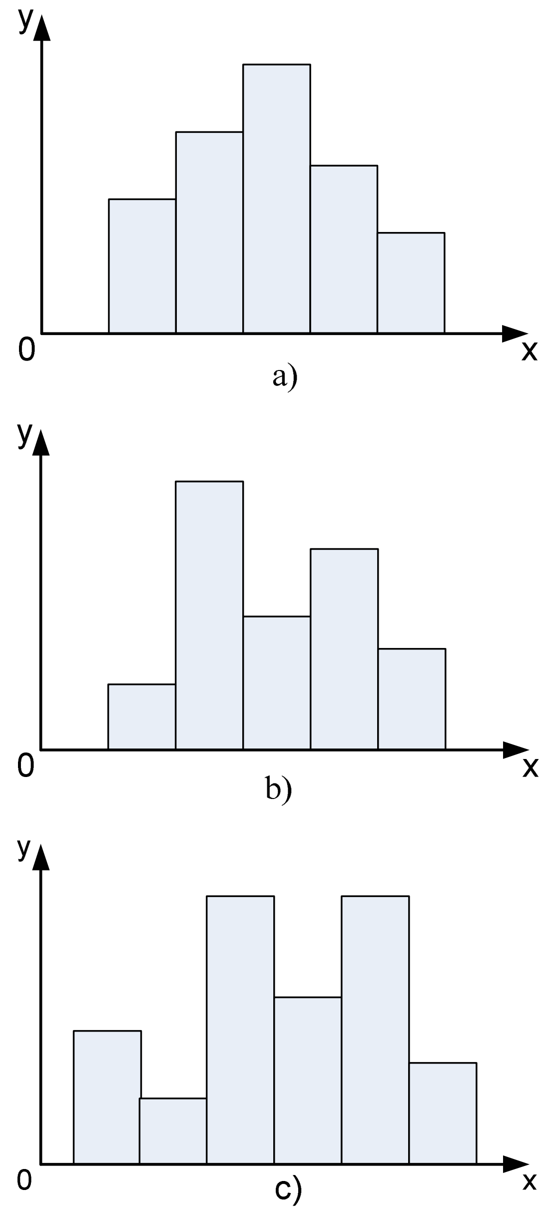


Fig. 4.2. Histograme ale unor tipuri de distribuții heterograde

a) serie unimodală;
b) serie plurimodală cu un singur interval modal principal;
c) serie plurimodală cu mai multe intervale modale principale

În lucrările din cadrul statisticii matematice pot fi întâlnite mai multe puncte de vedere asupra abordării seriilor cu mai multe intervale modale. După unele dintre acestea, rigoarea unei analize statistice solicită ca într-o serie să nu fie decât un singur interval modal. Pentru a se ajunge la aceasta, seriile cu mai multe intervale modale pot fi transformate prin diferite procedee: schimbarea numărului de grupe, trecerea la intervale de variație inegale ş.a.m.d. După alte opinii, analiza seriilor statistice se poate face şi cu mai multe valori ale modului.

În determinarea modului unei distribuții heterograde se pot folosi două modalități:

- prin calcul analitic;
- prin tehnici grafice.

4.4.1.1. Calculul analitic al modului unei distribuții heterograde

Pentru calculul analitic al modului unei distribuții heterograde poate fi aplicat următorul algoritm:

Pasul 1. Se stabileşte intervalul modal pentru care se va calcula modul;

Pasul 2. Se determină diferența dintre frecvența absolută a intervalului modal şi frecvența absolută a intervalului anterior intervalului modal, notată cu Δ_1 (atunci când intervalul modal corespunde primei grupe, se poate considera că aceasta este precedată de o grupă cu frecvența nulă);

Pasul 3. Se determină diferența dintre frecvența absolută a intervalului modal şi frecvența absolută a intervalului ulterior intervalului modal, notată cu Δ_2 (atunci când intervalul modal corespunde ultimei grupe se poate considera că aceasta este urmată de o grupă cu frecvența absolută nulă);

Pasul 4. Se calculează valoarea modului prin formula:

$$M_o^x = x_{M_0-1} + d_x^{M_0} \frac{\Delta_1}{\Delta_1 + \Delta_2} \qquad (4.14.)$$

unde:

x_{M_0-1} este limita inferioară a intervalului modal; $d_x^{M_0}$ este lungimea intervalului modal

Exemplul 4.5. Se cere să se determine, prin calcul analitic, modul seriei statistice prezentată în tabelul 4.1.

Rezolvare: Seria prezentată în tabelul 4.1. este unimodală, iar frecvența maximă corespunde intervalului [700 , 900). Cele două diferențe vor avea valorile:

$\Delta_1 = n_3^x - n_2^x = 80 - 50 = 30$; $\Delta_2 = n_3^x - n_4^x = 80 - 40 = 40$.

Aplicând relația (4.14.) rezultă valoarea modului seriei:

$$M_o^x = x_{M_0 - 1} + d_x^{M_0} \frac{\Delta_1}{\Delta_1 + \Delta_2} = 700 + 200 \frac{30}{30 + 40} = 785{,}71 \text{ RON/lună.}$$

4.4.1.2. Determinarea pe cale grafică a modului unei distribuții heterograde

În acest subcapitol va fi prezentată o tehnică grafică de determinare a modului preferată în practică datorită simplității sale. Această tehnică este descrisă de următorul algoritm:

Pasul 1. Se desenează histograma distribuției heterograde;

Pasul 2. Sunt stabilite, pe baza histogramei, intervalele modale;

Pasul 3. Pentru dreptunghiul care corespunde unui interval modal sunt trasate două linii drepte:

- una din colțul din dreapta sus al dreptunghiului intervalului modal până la colțul din dreapta sus al dreptunghiului intervalului anterior intervalului modal;
- alta din colțul din stânga sus al intervalului modal până la colțul din stânga sus al intervalului ulterior intervalului modal;

Pasul 4. La intersecția celor două linii drepte trasate anterior se coboară o perpendiculară pe axa absciselor, pe care o va intersecta în punctul ce corespunde valorii modului.

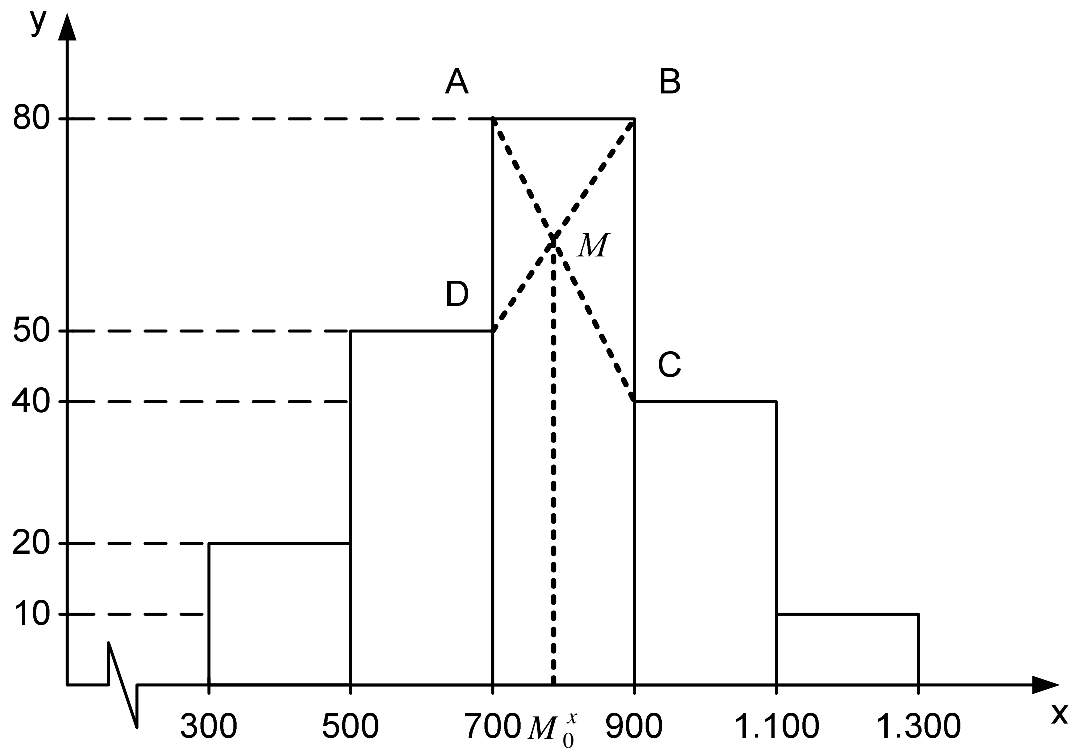


Fig. 4.3. Determinarea pe cale grafică a modului unei distribuții heterograde

Exemplul 4.6. Se cere să se determine pe cale grafică modul seriei prezentate în tabelul 4.1.

Rezolvare: În figura 4.3. este prezentată histograma seriei. Se poate observa că seria are un singur interval modal, ce corespunde dreptunghiului cu cea mai mare latură verticală. Din colțurile acestuia A şi B se trasează două linii drepte către colțurile dreptunghiurilor învecinate, respectiv D. La intersecția celor două linii se află punctul M din care se coboară o perpendiculară către axa absciselor pe care o va intersecta într-un punct ce corespunde modului seriei.

4.4.2. Utilizarea modului în caracterizarea fenomenelor colective

Rolul pe care modul unei serii statistice îl are în caracterizarea fenomenelor studiate derivă din legătura, prezentată anterior, dintre frecvențe şi probabilități. Valoarea cu cea mai mare frecvență are semnificația rezultatului cel mai probabil al unui fenomen, de care trebuie să se țină seama în cercetările statistice. Totuşi, aşa cum se întâmplă şi cu valoarea mediană, în comparaţie cu media aritmetică, modul reflectă într-o măsură mult mai mică trăsăturile esenţiale ale fenomenelor studiate. Şi tot la fel ca în cazul valorii mediane, modul unei serii este folosit mai mult pentru a completa caracterizările făcute pe baza valorilor medii, în special când acestea nu sunt foarte reprezentative.

Relaţia dintre un mod al unei distribuţii de frecvenţe şi media aritmetică a acesteia trebuie analizată diferenţiat, în raport cu numărul şi tipul intervalelor modale. Astfel, la seriile unimodale, valoarea modului este apropiată de cea a mediei aritmetice atunci când unicul interval modal este situat în centrul intervalului de valori, iar seria este dispusă simetric în raport cu acesta (în cazul unei simetrii perfecte, valoarea modului ajunge chiar să se confunde cu cea a mediei aritmetice). Petru seriile plurimodale cu un singur interval modal principal, valoarea modului din acesta este de asemenea apropiată de cea a mediei aritmetice atunci când intervalul modal principal este situat în centrul seriei care are o dispunere simetrică (şi în acest caz, dacă simetria este perfectă, valoarea modului ajunge să se confunde cu cea a mediei aritmetice). În ce priveşte seriile plurimodale cu mai multe intervale modale principale, relaţia dintre valorile modurilor şi media aritmetică este ceva mai complexă şi trebuie analizată pe baza aspectelor concrete ale distribuţiilor de frecvenţe. Pentru acest tip de serii poate fi menţionat, ca un caz particular, distribuţia în formă de U, la care media aritmetică este egal depărtată faţă de cele două valori ale modului.

La fel ca în cazul valorii mediane, comparaţiile dintre valoarea unui mod şi cea a mediei aritmetice servesc în evaluarea simetriei unei serii statistice, aspect care va fi abordat într-un capitol ulterior.

Capitolul 5 - Dispersia seriilor statistice

5.1. Coordonate ale studiului dispersiei seriilor statistice

În capitolele anterioare s-a menţionat că valorile tipice ale unei serii statistice sunt cu atât mai puţin reprezentative cu cât împrăştierea (sau *dispersia*) seriei este mai mare. Astfel, dispersia unei serii devine un indicator important, cu toată că nu singurul, al reprezentativităţii valorilor tipice.

O cercetare statistică riguroasă îşi propune ca în afară de a studia reprezentativitatea valorilor tipice în termeni generali sau intuitivi, să transpună acest aspect într-o formă cuantificabilă, care să permită comparaţiile şi clasificările. Din acest motiv, în cercetările statistice este practic inerentă determinarea unor mărimi numerice care exprimă dispersia seriilor. În general, aceste mărimi sunt calculate pe baza diferenţelor (*abaterilor*) valorilor unei serii faţă de anumite valori tipice, în special faţă de media aritmetică.

La o distribuţie de frecvenţe reprezentativitatea valorilor tipice este influenţată, aşa cum s-a menţionat în capitolul anterior, nu doar de dispersia centrelor de interval ci şi de reprezentativitatea pe care acestea, la rândul lor, o au în raport cu valorile din grupe. Din acest motiv, studiul reprezentativităţii unei valori tipice pentru o distribuţie de frecvenţe poate cuprinde şi evaluarea dispersiei valorilor din fiecare grupă.

5.2. Indicatori ai dispersiilor seriilor statistice

În acest subcapitol vor fi prezentate succint cinci mărimi folosite destul de frecvent în practică pentru evaluarea dispersiei:

- abaterea medie liniară;
- varianța;
- abaterea medie pătratică;
- coeficientul de variație în raport cu abaterea medie liniară;
- coeficientul de variație în raport cu abaterea medie pătratică.

5.2.1. Abaterea medie liniară

Abaterea medie liniară este un indicator care exprimă nivelul mediu al diferențelor (*abaterilor*) dintre valorile unei serii şi o valoare tipică a acesteia. De regulă abaterile sunt stabilite în raport cu media aritmetică a seriei; ceva mai rar sunt calculate şi în funcție de valoarea mediană.

Media abaterilor față de o valoare tipică nu poate fi exprimată pe baza simplei însumări a acestora întrucât diferențele pozitive şi cele negative s-ar anula reciproc (se poate chiar demonstra că în cazul unei serii simple suma diferențelor față de media aritmetică este nulă). Din acest motiv sunt folosite valorile absolute ale acestor diferențe. În raport cu tipul seriilor statistice se pot delimita două modalități de determinare a abaterilor medii liniare:

a) pentru seriile simple;
b) pentru distribuțiile heterograde.

a) Calculul abaterii medii liniare a unei **serii simple**, are la bază formula:

$$\bar{d}_{x_0} = \frac{\sum_{i=1}^{N} |x_i - \bar{x}_o|}{N} \qquad (5.1.)$$

în care:

$\bar{d}_{x_0}$ este abaterea medie liniară a unei serii simple în raport cu o caracteristică x; N este numărul de unități statistice ale seriei; x_i este valoarea caracteristicii x pentru o unitate statistică i; $\bar{x}_0$ este media aritmetică a seriei.

Exemplul 5.1. În tabelul 5.1. este prezentată o serie statistică simplă care descrie productivitatea medie a muncii pentru un grup de şase angajaţi ai unei firme. Se cere să se calculeze abaterea medie liniară a seriei.

Tabelul 5.1. Productivitatea medie a muncii pentru un grup de angajaţi

Nr. crt.	Productivitatea medie a muncii (x_i) [RON/lună]
(0)	(1)
1	700
2	700
3	630
4	870
5	620
6	680

Rezolvare: În tabelul 5.2. sunt prezentate valorile intermediare utilizate în calculul abaterii medii liniare.

Determinarea abaterii medii liniare necesită, mai întâi, calculul mediei aritmetice:

$$\overline{x}_0 = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{4.200}{6} = 700 \text{ RON/lună}$$

Pe baza valorilor intermediare calculate prin tabelul 5.1. poate fi determinată abaterea medie liniară a seriei:

$$\overline{d}_{x_0} = \frac{\sum_{i=1}^{N} |x_i - x_0|}{N} = \frac{340}{6} = 56{,}7 \text{ RON/lună,}$$

ceea ce înseamnă că, în medie, valorile seriei simple diferă cu 56,7 RON/lună faţă de media aritmetică.

Tabelul 5.2. Valori intermediare utilizate în calculul abaterii medii liniare a unei serii simple

RON/lună

Nr. crt.	x_i	Abatere față de media aritmetică $(x_i - \overline{x}_o)$	$\lvert x_i - \overline{x}_o \rvert$
(0)	(1)	(3)	(4)
1	700	–	–
2	700	–	–
3	630	- 70	70
4	870	170	170
5	620	- 80	80
6	680	- 20	20
Total	4.200	×	340
Simbol pentru total	$\sum_{i=1}^{n} x_i$	×	$\sum_{i=1}^{N} \lvert x_i - x_0 \rvert$

b) Determinarea **abaterii medii liniare a unei distribuții heterograde**, are la bază relația:

$$\overline{d}_x = \frac{\sum_{i=1}^{K_x} \lvert x_i^{'} - \overline{x} \rvert \cdot n_i^x}{\sum_{i=1}^{K_x} n_i^x} \qquad (5.2.)$$

în care:

- $\overline{d}_x$ este abaterea medie liniară a distribuției heterograde;
- K_x este numărul de grupe formate în raport cu caracteristica x;
- $x_i^{'}$ este centrul intervalului de variație al unei grupe i;
- $\overline{x}$ este media aritmetică a distribuției heterograde în raport cu caracteristica x;
- n_i^x este frecvența absolută a grupei i.

Exemplul 5.2. În tabelul 5.3. este prezentată o distribuție heterogradă care descrie repartizarea unui grup de întreprinderi în raport cu cifra de afaceri. Se cere să se calculeze abaterea medie liniară a seriei.

Tabelul 5.3. Repartizarea unui grup de întreprinderi în raport cu cifra de afaceri

Nr. crt.	Interval de variație [mil. euro]	Frecvență absolută $\left(n_i^x\right)$
(0)	(1)	(2)
1	[2 ; 6)	15
2	[6 ; 10)	25
3	[10 ; 14)	30
4	[14 ; 18)	20
5	[18 ; 22)	10

Rezolvare: În tabelul 5.4. sunt prezentate valorile intermediare care servesc în determinarea abaterii medii liniare. Ca şi în exemplul precedent, determinarea abaterii medii liniare demarează cu calculul mediei aritmetice: $\bar{x} = \frac{\sum_{i=1}^{K_x} x' \cdot n_i^x}{\sum_{i=1}^{K_x} n_i^x} = \frac{1.140}{100} = 11{,}4$ mil. euro

Tabelul 5.4. Valori intermediare utilizate în calculul abaterii medii liniare a unei distribuții heterograde

Nr. crt.	Interval de variație [mil. euro]	Frecvență absolută $\left(n_i^x\right)$	Centru de interval $\left(x_i^{'}\right)$ [mil. euro]	$\left(x_i^{'} \cdot n_i^x\right)$	Abatere față de media aritmetică $\left(x_i^{'} - \bar{x}\right)$	$\left\lvert x_i^{'} - \bar{x}\right\rvert \cdot n_i^x$
(0)	(1)	(2)	(3)	(4) = (3) × (2)	(5)	(6) = |(5)|×(2)
1	[2 ; 6)	15	4	60	- 7,4	111
2	[6 ; 10)	25	8	200	- 3,4	85
3	[10 ; 14)	30	12	360	0,6	18
4	[14 ; 18)	20	16	320	4,6	92
5	[18 ; 22)	10	20	200	8,6	86
6	Total	100	×	1.140	×	392
7	Simbol pentru total	$\sum_{i=1}^{K_x} n_i^x$	×	$\sum_{i=1}^{K_x} x_i^{'} \cdot n_i^x$	×	$\sum_{i=1}^{K_x}\left(\left\lvert x_i^{'} - \bar{x}\right\rvert \cdot n_i^x\right)$

În raport cu valorile intermediare calculate prin tabelul 5.4. se determină abaterea medie liniară a distribuției heterograde:

$$\overline{d}_x = \frac{\sum_{i=1}^{K_x} \left| x_i^{'} - \overline{x} \right| \cdot n_i^x}{\sum_{i=1}^{K_x} n_i^x} = \frac{392}{100} = 3{,}92 \text{ mil. euro,}$$

ceea ce înseamnă că, în medie, cifra de afaceri a unei firme diferă cu 3,92 milioane de euro față de cifra de afaceri medie.

Abaterea medie liniară a unei serii poate lua, după cum se poate observa din formulele sale de calcul, doar valori pozitive. Cu cât valoarea sa este mai mare cu atât seria este mai dispersată, iar media sa aritmetică este mai puțin reprezentativă. Totuşi, faptul că această mărime nu îmbracă o formă relativă induce unele dificultăți în comparațiile dintre seriile statistice sau în clasificarea acestora în raport cu dispersia.

5.2.2. Varianța

Varianța unei serii este o mărime care exprimă nivelul mediu al pătratelor diferențelor dintre valorile seriei şi media aritmetică a acesteia. Prin utilizarea pătratelor diferențelor nu mai este posibilă anularea reciprocă a acestora, astfel încât nu mai este necesară folosirea valorilor absolute. La fel ca în cazul abaterii medii liniare, calculul varianței se diferențiază, în raport cu tipurile de serii statistice, în două forme:

a) pentru seriile simple;
b) pentru distribuțiile heterograde.

a) Calculul **varianței unei serii simple** are la bază formula:

$$\sigma_{x_0}^2 = \frac{\sum_{i=1}^{N} \left(x_i - \overline{x}_0 \right)^2}{N} \qquad (5.3.)$$

unde $\sigma_{x_0}^2$ este varianța seriei simple.

Exemplul 5.3. Se cere să se calculeze varianța seriei simple prezentate în tabelul 5.1.

Rezolvare: În exemplul 5.1. a fost calculată deja media aritmetică a seriei simple $\overline{x}_0$ = 700 RON/lună. Pe baza acesteia sunt calculate valorile intermediare pentru determinarea varianței, care sunt prezentate în tabelul 5.5. În raport cu acestea rezultă o valoare a varianței:

$$\sigma_{x_0}^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x}_0)^2}{N} = \frac{40.600}{6} = 6.766,7(\text{RON/lună})^2,$$ ceea ce înseamnă că diferența la pătrat a valorilor seriei față de media aritmetică are un nivel mediu de 6.766,7 (RON/lună)^2.

Tabelul 5.5. Valori intermediare utilizate în calculul varianței unei serii simple

Nr. crt.	x_i [RON/lună]	$x_i - \bar{x}_0$ [RON/lună]	$(x_i - \bar{x}_0)^2$ [RON^2/lună^2]
(0)	(1)	(2)	(3) = (2)^2
1	700	–	
2	700	–	
3	630	- 70	4 900
4	870	170	28 900
5	620	- 80	6 400
6	680	- 20	400
Total	4 200	×	40 600
Simbol pentru total	$\sum_{i=1}^{n} x_i$	×	$\sum_{i=1}^{N}(x_i - \bar{x}_0)^2$

b) Determinarea **varianței unei distribuții heterograde** se bazează pe relația:

$$\sigma_x^2 = \frac{\sum_{i=1}^{K_x}\left(x_i^{'} - \bar{x}\right)^2 \cdot n_i^x}{\sum_{i=1}^{K_x} n_i^x} \tag{5.4.}$$

unde σ_x^2 este varianța distribuției heterograde.

Exemplul 5.4. Se cere să se calculeze varianța distribuției heterograde prezentată în tabelul 5.3.

Rezolvare: Şi în acest caz, vom profita de faptul că într-un exemplu anterior a fost calculată media aritmetică a distribuției heterograde, $\bar{x} = 11{,}4$ mil. euro.

Tabelul 5.6. Valori intermediare utilizate în calculul varianței unei distribuții heterograde

Nr. crt.	Interval de variație [mil. euro]	$x_i^{'} - \bar{x}$ [mil. euro]	n_i^x	$\left(x_i^{'} - \bar{x}\right)^2$ [(mil. euro)^2]	$\left(x_i^{'} - \bar{x}\right)^2 \cdot n_i^x$ [(mil. euro)^2]
(0)	(1)	(3)	(2)	(4) = (2)^2	(5) = (4) × (3)
1	[2 ; 6)	- 7,4	15	54,76	821,4
2	[6 ; 10)	- 3,4	25	11,56	289,0
3	[10 ; 14)	0,6	30	0,36	10,8
4	[14 ; 18)	4,6	20	21,16	423,2
5	[18 ; 22)	8,6	10	73,96	739,6
6	Total	×	100	×	2.284
7	Simbol pentru total	×	$\sum_{i=1}^{K_x} n_i^x$	×	$\sum_{i=1}^{K_x} \left(x_i^{'} - \bar{x}\right)^2 \cdot n_i^x$

În tabelul 5.6. sunt prezentate valorile intermediare în calculul varianței. În raport cu acestea a rezultat o valoare a varianței:

$$\sigma_x^2 = \frac{\sum_{i=1}^{K_x} \left(x_i^{'} - \bar{x}\right)^2 \cdot n_i^x}{\sum_{i=1}^{K_x} n_i^x} = \frac{2.284}{100} = 22,84 \quad \text{(mil. euro)}^2,$$ ceea ce înseamnă că pătratul diferenței dintre cifra de afaceri a unei firme şi media cifrei de afaceri din cadrul grupului de firme are un nivel mediu de 22,84 (mil. euro)^2.

Din formulele de calcul ale varianței se poate observa că această mărime nu poate lua decât valori pozitive. O serie statistică este cu atât mai dispersată cu cât varianța sa este mai mare.

Modul de calcul al varianței induce unele deosebiri față de abaterea medie liniară în ce priveşte exprimarea dispersiei unei serii statistice. Faptul că se operează cu abateri ridicate la pătrat face ca unitatea de măsură a varianței să fie reprezentată de pătratul unității de măsură a caracteristicii. În plus, aceeaşi ridicare la pătrat face ca abaterile mari să contribuie la valoarea varianței în proporții mult mai mari decât abaterile mici. În aceste condiții, varianța exprimă într-o măsură mai mare față de abaterea medie liniară amploarea dispersiei unei serii statistice.

La fel ca în cazul abaterii medii liniare, faptul că varianța are o formă absolută cauzează unele dificultăți în comparațiile dintre seriile statistice sau în clasificarea acestora pe baza dispersiei.

5.2.3. Abaterea medie pătratică

Abaterea medie pătratică are semnificația unei medii de ordinul doi (numită şi *medie pătratică*) a diferențelor dintre valorile unei serii statistice şi media aritmetică a acesteia. În fapt, abaterea medie pătratică poate fi obținută, atât pentru seriile simple cât şi pentru distribuțiile heterograde, extrăgând rădăcina pătrată din valoarea varianței. La seriile simple, abaterea medie pătrată, notată cu σ_{x_0}, este dată de relația:

$$\sigma_{x_0} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x}_0)^2}{N}} = \sqrt{\sigma_{x_0}^2} \qquad (5.5.)$$

Pentru o distribuție heterogradă, abaterea medie pătratică este notată cu σ_x şi poate fi calculată prin formula:

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{K_x}\left(x_i' - \bar{x}\right)^2 \cdot n_i^x}{\sum_{i=1}^{K_x} n_i^x}} = \sqrt{\sigma_x^2} \qquad (5.6.)$$

Exemplul 5.5. Se cere să se calculeze abaterile medii pătratice ale seriilor prezentate în tabelele 5.1. şi 5.3.

Rezolvare: Pentru ambele serii, determinarea abaterii medii pătratice este simplă în condițiile în care în exemplele anterioare au fost calculate varianțele.

Pentru seria simplă a rezultat o abatere medie pătratică: $\sigma_{x_0} = \sqrt{\sigma_{x_0}^2} = \sqrt{6.766,7} = 82,26$ RON/lună

Pentru distribuția heterogradă, s-a obținut o abatere medie pătratică:

$\sigma_x = \sqrt{\sigma_x^2} = \sqrt{22,84} = 4,78$ mil. euro.

Formulele de calcul asociate abaterii medii pătratice indică faptul că această mărime nu poate avea decât valori pozitive. Cu cât o serie statistică este mai dispersată, cu atât abaterea medie pătratică a acesteia va fi mai mare.

Media pătratică este, în mod obligatoriu, mai mare sau egală față de media aritmetică, ceea ce face ca întotdeauna abaterea medie pătratică a unei serii să fie mai mare sau egală față de abaterea medie liniară a seriei. La fel ca în cazul varianței, abaterile mari contribuie la valoarea abaterii medii pătratice într-o proporție mult mai mare decât abaterile mici. În consecință, abaterea medie pătratică exprimă, în comparație cu abaterea medie liniară, într-o măsură mult mai mare amploarea dispersiei unei serii statistice. Abaterea medie pătratică se deosebeşte de varianță prin faptul că este exprimată în unitatea de măsură a caracteristicii, ceea ce face mai facilă aprecierea nivelului abaterilor. La fel ca şi abaterea medie liniară sau varianța, abaterea medie pătratică este o mărime absolută, ceea ce face foarte dificilă comparația dintre seriile statistice sau clasificarea acestora din perspectiva dispersiei.

5.2.4. Coeficientul de variație în raport cu abaterea medie liniară

Coeficientul de variație în raport cu abaterea medie liniară este o mărime relativă, în formă procentuală, obținută prin raportarea abaterii medii liniare la media aritmetică în valoare absolută. Pentru o serie simplă, coeficientul de variație în raport cu abaterea medie liniară, notat cu $CV_x^{\bar{d}_0}$, este dat de formula:

$$CV_x^{\bar{d}_0} = \frac{\bar{d}_{x_0}}{|\bar{x}_0|} \times 100 \qquad (5.7.)$$

Coeficientul de variație în raport cu abaterea medie liniară al unei distribuții heterograde, notat cu $CV_x^{\bar{d}}$, poate fi calculat prin formula:

$$CV_x^{\bar{d}} = \frac{\bar{d}_x}{|\bar{x}|} \times 100 \qquad (5.8.)$$

Evident, o astfel de mărime nu poate avea decât valori pozitive, iar seria este cu atât mai dispersată cu cât valoarea este mai mare.

Calitatea de mărime relativă facilitează utilizarea acestui indicator în comparațiile şi clasificările seriilor statistice din perspectiva dispersiei. Astfel, se apreciază că o valoare mai mare de

30% indică o serie cu omogenitate redusă pentru care media aritmetică nu este prea reprezentativă.

Exemplul 5.6. Se cere să se aprecieze reprezentativitatea mediilor aritmetice, pe baza coeficientului de variație în raport cu abaterea medie liniară, pentru seriile prezentate în tabelele 5.1. şi 5.3.
Rezolvare: Calculul celor două valori este destul de facil, în condițiile în care atât mediile aritmetice cât şi abaterile medii liniare au fost determinate în exemple anterioare.

Coeficientul de variație în raport cu abaterea medie liniară al seriei simple are valoarea:

$$CV_x^{\bar{d}_0} = \frac{\bar{d}_{x_0}}{|\bar{x}_0|} \times 100 = \frac{56{,}7}{700} \times 100 = 8{,}1\%$$

ceea ce indică o serie cu omogenitate semnificativă, pentru care media aritmetică este reprezentativă.

Pentru distribuția heterogradă, coeficientul de variație în raport cu abaterea medie liniară are valoarea:

$$CV_x^{\bar{d}} = \frac{\bar{d}_x}{|\bar{x}|} \times 100 = \frac{3{,}92}{11{,}4} \times 100 = 34{,}4\%,$$

ceea ce indică o omogenitate relativ redusă a seriei, pentru care media aritmetică nu este foarte reprezentativă.

5.2.5. Coeficientul de variație în raport cu abaterea medie pătratică

Coeficientul de variație în raport cu abaterea medie pătratică, propus în anul 1896 de către statisticianul Karl Pearson, este o altă mărime relativă, în formă procentuală care măsoară dispersia unei serii statistice. Acest indicator este obținut prin raportarea abaterii medii pătratice la valoarea absolută a mediei aritmetice. Pentru o serie simplă, coeficientul de variație în raport cu abaterea medie pătratică, notat cu $CV_x^{\sigma_0}$, poate fi calculat prin formula:

$$CV_x^{\sigma_0} = \frac{\sigma_{x_0}}{|\bar{x}_0|} \times 100 \qquad (5.9.)$$

Coeficientul de variație în raport cu abaterea medie pătratică al unei distribuții heterograde, notat cu CV_x^{σ} este dat de relația:

$$CV_x^{\sigma} = \frac{\sigma_x}{|\bar{x}|} \times 100 \qquad (5.10)$$

Din formulele de calcul se poate observa că această mărime nu poate avea decât valori pozitive. Cu cât valoarea sa este mai mare cu atât seria este mai dispersată. În condițiile în care abaterea medie pătratică este mai mare sau egală decât abaterea medie liniară şi coeficientul de variație în raport cu abaterea medie pătratică va fi întotdeauna mai mare sau cel mult egal față de coeficientul de variație în raport cu abaterea medie liniară.

Fiind o mărime relativă, coeficientul de variație în raport cu abaterea medie pătratică este utilizat frecvent în comparațiile şi clasificările seriilor statistice din perspectiva dispersiei. Astfel, se apreciază că atunci când valoarea sa depăşeşte nivelul de 40%, seria statistică este puțin omogenă, iar media sa aritmetică nu este prea reprezentativă. Aprecierea dispersiei pe baza coeficientului de variație în raport cu abaterea medie pătratică este considerată mai riguroasă decât cea realizată prin coeficientul de variație în raport cu abaterea medie liniară în condițiile în care abaterea medie pătratică reflectă amploarea dispersării într-o măsură mai mare decât abaterea medie liniară.

Exemplul 5.7. Se cere să se aprecieze, pe baza coeficientului de variație în raport cu abaterea medie pătratică, reprezentativitatea mediilor aritmetice ale seriilor statistice prezentate în tabelele 5.1. şi 5.3.

Rezolvare: Cele două valori pot fi determinate destul de simplu, pe baza mediilor aritmetice şi a abaterilor medii pătratice calculate în exemplele anterioare.

Pentru seria simplă, coeficientul de variație în raport cu abaterea medie pătratică are valoarea:

$$CV_x^{\sigma_0} = \frac{\sigma_{x_0}}{|\bar{x}_0|} \times 100 = \frac{82,26}{700} \times 100 = 11,75\%$$

ceea ce înseamnă că omogenitatea seriei este semnificativă, iar media aritmetică are o reprezentativitate mare. Pentru distribuția heterogradă se determină un coeficient de variație în raport cu abaterea medie pătratică:

$$CV_x^{\sigma} = \frac{\sigma_x}{|\bar{x}|} \times 100 = \frac{4,78}{11,4} \times 100 = 41,9\%$$

ceea ce indică o slabă omogenitate a seriei şi o reprezentativitate redusă a mediei aritmetice.

Capitolul 6 - Asimetria şi boltirea seriilor statistice

6.1. Conceptul de asimetrie a seriilor statistice

O valoare medie a unei serii statistice exprimă rezultatul factorilor esenţiali de influenţă asupra fenomenului colectiv de influenţă asupra fenomenului colectiv studiat. Abaterile de la medie ale celorlalte valori ale seriei exprimă impactul pe care alţi factori, întâmplători, îl au asupra fenomenului. Atunci când influenţa factorilor întâmplători se produce cu regularitate, valorile seriei sunt dispuse simetric faţă de medie. În schimb, atunci când această influenţă se manifestă în mod neregulat, seria este asimetrică în raport cu media. Studiul asimetriei seriilor statistice are aplicaţii practice îndeosebi în cazul distribuţiilor heterograde, fiind folosit la asocierea cu una dintre formele de abstractizare a seriilor: distribuţia în formă de J, distribuţia în formă de U, distribuţia în formă de clopot etc.

Cel mai adesea sunt folosite asocierile cu o distribuţie în formă de clopot, care reflectă o lege de repartiţie normală ce caracterizează frecvent manifestările fenomenelor colective. După cum se ştie, o astfel de serie este perfect simetrică, astfel încât studiul unei distribuţii heterograde poate servi în evaluarea gradului în care seria diferă de o distribuţie în formă de clopot. În afara distribuţiilor heterograde, cercetarea asimetriei poate fi aplicată şi la seriile simple, mai ales atunci când se încearcă asocierea acestora cu legi de distribuţie normală.

În studiul asimetriei unei serii statistice sunt abordate mai multe aspecte: măsura în care aceasta este îndepărtată de o dispunere simetrică a valorilor, preponderenţa valorilor mai mici sau, dimpotrivă, mai mari faţă de medie etc.

Rigorile unei cercetări statistice impun folosirea unor mărimi numerice prin care aceste aspecte să poată fi cuantificate iar seriile să poată fi comparate şi clasificate.

6.2. Evaluarea asimetriei seriilor statistice

În acest subcapitol vor fi prezentate succint două modalități de evaluare a asimetriei unei serii statistice:

- prin comparația dintre media aritmetică şi valoarea modului;
- prin comparația dintre media aritmetică şi valoarea mediană.

6.2.1. Evaluarea asimetriei prin comparația dintre media aritmetică şi valoarea modului

Cercetarea asimetriei seriilor statistice pe baza comparației dintre media aritmetică şi valoarea modului este indicată îndeosebi în situația distribuțiilor unimodale. În acest caz modul are semnificația celui mai probabil rezultat iar atunci când factorii întâmplători influențează în mod regulat fenomenul studiat simetria seriei statistice se manifestă prin egalitatea dintre mod şi media aritmetică. Când însă factorii întâmplători se manifestă în mod neregulat, asimetria seriei se poate reflecta printr-o valoare a modului diferită față de media aritmetică. Aprecierea asimetriei pe baza comparației dintre media aritmetică şi mod se poate realiza şi pe cale grafică, reprezentându-se seriile statistice prin curbe sau poligoane de frecvențe, pentru care valoarea modului corespunde celui mai înalt punct al graficului (fig. 6.1.). Totuşi, reprezentările grafice nu permit cuantificarea asimetriei, astfel încât este necesară utilizarea unor mărimi numerice, calculate pe baza celor două valori tipice.

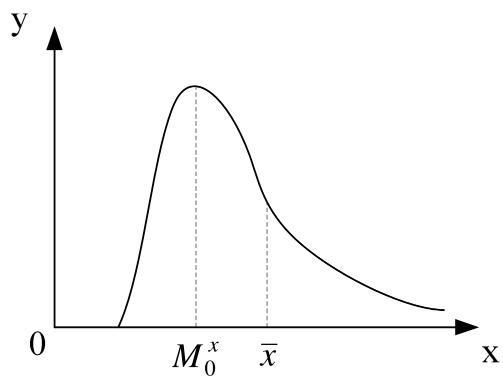
Diferența dintre media aritmetică şi valoarea modului este o mărime absolută, greu de utilizat în comparațiile dintre seriile statistice sau în clasificarea acestora din perspectiva asimetriei. Pentru astfel de situații se recomandă utilizarea unor mărimi relative, aşa cum este *coeficientul de asimetrie în raport cu modul*, propus de Karl Pearson. Acest indicator, notat cu $C_{as_x}^{M_o}$, poate fi obținut raportând la abaterea medie pătratică (atunci când aceasta nu este nulă), diferența dintre media aritmetică şi mod:

$$C_{as_x}^{M_o} = \frac{\overline{X - M_0^x}}{\sigma_x} \quad (6.1.)$$

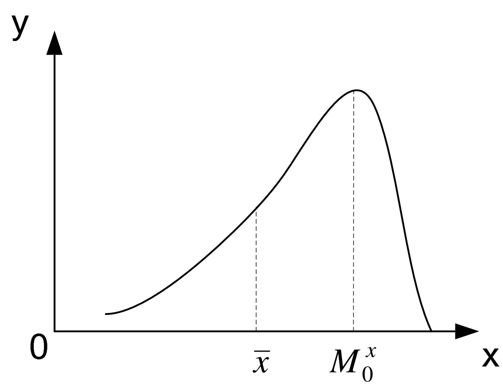
Se poate demonstra că diferența, în valoare absolută, dintre media aritmetică şi mod este cel mult egală cu abaterea medie pătratică a unei serii. Din acest motiv coeficientul de asimetrie al seriei nu poate lua decât valori cuprinse în intervalul [-1; 1].

În condițiile în care abaterea medie pătratică nu poate avea decât valori pozitive rezultă că valoarea coeficientului este pozitivă sau negativă după cum diferența dintre media aritmetică şi valoarea modului este mai mare, respectiv, mai mică decât zero. Astfel spus, când coeficientul este mai mare decât zero seria are asimetrie pozitivă (spre dreapta) iar când este mai mic decât zero asimetria seriei este negativă (spre stânga).

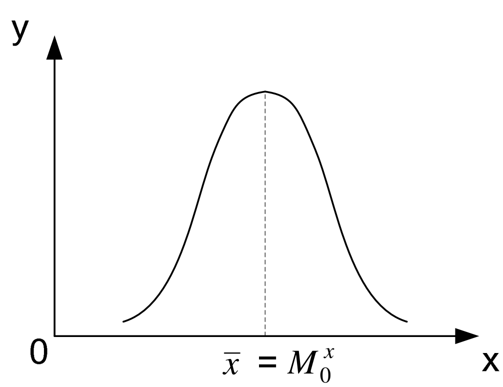
Acest indicator poate fi utilizat şi în cuantificarea intensității asimetriei. Cu cât valorile sale absolute sunt mai apropiate de 1 cu atât asimetria este mai pronunțată. Se obişnuieşte ca intervalul [0; 1] pe care îl ocupă valorile absolute ale coeficientului să fie împărțit în trei intervale de lungimi egale pentru fiecare dintre acestea fiind asociat, în raport cu depărtarea de valoarea 1, un grad de asimetrie: puternică, moderată sau slabă. Astfel, în funcție de valorile coeficientului pot fi apreciate atât sensul cât şi intensitatea asimetriei unei serii (tabelul 6.1.).



a) Serie cu asimetrie pozitivă



b) Serie cu asimetrie negativă



c) Serie simetrică

Fig. 6.1. Reprezentarea prin curbe de frecvențe a relației dintre media aritmetică şi valoarea modului

Tabelul 6.1. Evaluarea asimetriei pe baza valorilor coeficientului de asimetrie în raport cu modul

Nr. crt.	**Valori ale coeficientului de asimetrie în raport cu modul ($C_{as_x}^{M_o}$)**	**Sensul şi intensitatea asimetriei**
1	$-1 \leq C_{as_x}^{M_o} < -\frac{2}{3}$	Negativă puternică
2	$-\frac{2}{3} \leq C_{as_x}^{M_o} < -\frac{1}{3}$	Negativă moderată
3	$-\frac{1}{3} \leq C_{as_x}^{M_o} < 0$	Negativă slabă
4	$C_{as_x}^{M_o} = 0$	Serie simetrică
5	$0 < C_{as_x}^{M_o} \leq \frac{1}{3}$	Pozitivă slabă
6	$\frac{1}{3} < C_{as_x}^{M_o} \leq \frac{2}{3}$	Pozitivă moderată
7	$\frac{2}{3} < C_{as_x}^{M_o} \leq 1$	Pozitivă puternică

Din comparaţia dintre media aritmetică şi valoarea modului unei serii pot rezulta trei situaţii:

- *asimetrie pozitivă* (numită şi *asimetrie de dreapta),* atunci când media aritmetică este mai mare decât modul seriei (fig. 6.1.a)*;*
- *asimetrie negativă* (numită şi *asimetrie de stânga*), atunci când media aritmetică este mai mică decât modul seriei (fig. 6.1.b);
- *simetria*, atunci când media aritmetică este egală cu modul seriei (fig. 6.1.c).

Exemplul 6.1. În tabelul 6.2. este prezentată o distribuţie heterogradă care descrie repartizarea punctelor de desfacere ale unei firme în raport cu vânzările realizate la un sortiment de produs. Se cere să se aprecieze sensul şi intensitatea asimetriei seriei pe baza coeficientului de asimetrie în raport cu modul.

Rezolvare: Determinarea coeficientului de asimetrie în raport cu modul presupune calculul prealabil al mediei aritmetice, al modului şi al abaterii medii pătratice.

Tabelul 6.2. Repartizarea punctelor de desfacere ale unei firme în raport cu vânzările realizate

Nr. crt.	**Interval de variație [mii buc.]**	**Frecvență absolută** (n_i^x)
(0)	(1)	(2)
1	(0 ; 40]	5
2	(40 ; 80]	9
3	(80 ; 120]	15
4	(120 ; 160]	14
5	(160 ; 200]	7

a) Calculul mediei aritmetice

Tabelul 6.3. Valori intermediare utilizate în calculul mediei aritmetice şi abaterii absolute medii pătratice

Nr. crt.	**Interval de variație [mii buc]**	**Frecvență absolută** (n_i^x)	**Centru de interval** ($x_i^{'}$) **[mii buc]**	$x_i^{'} n_i^x$ **[mii buc]**	($x_i^{'}$ - $\overline{X}$) **[mii buc]**	($x_i^{'}$ - $\overline{X}$) n_i^x **[(mii buc)^2]**
(0)	(1)	(2)	(3)	(4) = (3) × (2)	(5)	(6) = (5)^2 × (2)
1	(0 ; 40]	5	20	100	-87,2	38019,2
2	(40 ; 80]	9	60	540	-47,2	20050,6
3	(80 ; 120]	15	100	1500	-7,2	777,6
4	(120; 160]	14	140	1960	32,8	15061,8
5	(160; 200]	7	180	1260	72,8	37098,9
6	Total	50	×	5360	×	111008,1
7	Simbol pentru total	$\sum_{i=1}^{K_x} n_i^x$	×	$\sum_{i=1}^{K_x} x_i^{'} n_i^x$	×	$\sum_{i=1}^{K_x} \left(x_i^{'} - \overline{X}\right) n_i^x$

În tabelul 6.3. sunt prezentate valorile intermediare care servesc în calculul mediei aritmetice a seriei. Pe baza acestora rezultă o valoare a mediei aritmetice:

$$\overline{X}=\frac{\sum_{i=1}^{Kx}x_i^{'}\times n_i^x}{\sum_{i=1}^{Kx}n_i^x}=\frac{5360}{50}=107{,}2 \text{ mii bucăți}$$

b) Calculul valorii modulului seriei

Intervalul modal al seriei (cu frecvența maximă) este (80; 120]. Modul seriei are valoarea:

$$M_0^x=X_{M_0-1}+d_x\times\frac{\Delta_1}{\Delta_1+\Delta_2}=80+40\frac{6}{6+1}=114{,}3 \text{ mii bucăți}$$

c) Calculul abaterii medii pătratice

Valorile intermediare utilizate în calculul abaterii medii pătratice sunt prezentate în tabelul 6.3. Pe baza acestora rezultă o valoare a abaterii medii pătratice:

$$\sigma_x=\sqrt{\frac{\sum_{i=1}^{K_x}\left(x_i^{'}-\overline{X}\right)}{\sum_{i=1}^{K_x}n_i^x}}=\sqrt{\frac{111008{,}1}{50}}=47{,}1 \text{ mii bucăți}$$

d) Determinarea coeficientului de asimetrie în raport cu modul

Coeficientul are valoarea: $C_{as_x}^{M_0}=\frac{\overline{X}-M_0^x}{\sigma_x}=\frac{107{,}2}{114{,}3}=-0{,}15$, ceea ce semnifică o asimetrie negativă slabă.

6.2.2. Evaluarea asimetriei prin comparația dintre media aritmetică şi valoarea mediană

Studiul asimetriei pe baza comparației dintre media aritmetică şi valoarea mediană poate fi realizat atât pentru distribuții heterograde cât şi pentru seriile simple. Valoarea mediană, care împarte o serie ordonată în două grupe de frecvențe egale, se confundă cu media aritmetică atunci când factorii întâmplători influențează fenomenul studiat în mod regulat. Dacă aceşti factori întâmplători nu acționează cu regularitate, atunci asimetria seriei se manifestă printr-o valoare a mediei aritmetice diferită față de valoarea mediană.

Diferențele dintre media aritmetică şi valoarea mediană au semnificații similare diferențelor dintre media aritmetică şi valoarea modului, evocate anterior. O serie are o asimetrie negativă (de stânga) atunci când media aritmetică este mai mică decât valoarea mediană, şi o asimetrie pozitivă (de dreapta), atunci când media aritmetică este mai mare decât valoarea mediană.

Pentru cuantificarea intensităţii asimetriei unei serii statistice poate fi folosită o mărime relativă, numită *coeficient de asimetrie în raport cu mediana*. Acest indicator, notat cu $C_{as_x}^{M_e}$, poate fi calculat (atunci când abaterea medie pătratică a seriei nu este nulă) prin formula:

$$C_{as_x}^{M_e} = \frac{3\left(\overline{X} - M_e^x\right)}{\sigma_x} \qquad (6.2.)$$

Tabelul 6.4. Evaluarea asimetriei pe baza valorilor coeficientului de asimetrie în raport cu mediana

Nr. crt.	**Valori ale coeficientului de asimetrie în raport cu mediana ($C_{as_x}^{M_e}$)**	**Sensul şi intensitatea asimetriei**
1	$-3 \leq C_{as_x}^{M_e} < -2$	Negativă puternică
2	$-2 \leq C_{as_x}^{M_e} < -1$	Negativă moderată
3	$-1 \leq C_{as_x}^{M_e} < 0$	Negativă slabă
4	$C_{as_x}^{M_e} = 0$	Serie simetrică
5	$0 < C_{as_x}^{M_e} \leq 1$	Pozitivă slabă
6	$1 < C_{as_x}^{M_e} \leq 2$	Pozitivă moderată
7	$2 < C_{as_x}^{M_e} \leq 3$	Pozitivă puternică

În condiţiile în care abaterea medie pătratică este mai mare ca zero, valoarea coeficientului este pozitivă sau negativă după cum diferenţa dintre media aritmetică şi valoarea mediană este pozitivă, respectiv, negativă. Rezultă că asimetria este pozitivă atunci când coeficientul este mai mare ca zero şi negativă atunci când coeficientul este mai mic decât zero.

Această mărime poate fi folosită şi pentru cuantificarea intensităţii asimetriei. Cu cât valorile sale absolute sunt mai mari, cu atât asimetria este mai pronunţată. Se poate demonstra că diferenţa, în valoare absolută, dintre media aritmetică şi valoarea mediană este cel mult egală cu abaterea medie pătratică, astfel încât valorile coeficientului se încadrează în intervalul [-3 ; 3].

Se obişnuieşte, la fel ca în cazul mărimii anterioare, ca intervalul [0 ; 3] pe care îl ocupă valorile absolute ale coeficientului, să fie împărţit în trei intervale de lungimi egale iar pentru fiecare dintre acestea să fie asociat, în funcţie de depărtarea faţă de valoarea 3, un grad de asimetrie: puternică, moderată sau slabă.

La fel ca în cazul mărimii precedente, valorile acestui coeficient pot fi folosite pentru a aprecia deopotrivă sensul şi intensitatea asimetriei seriilor statistice (tab. 6.4.).

Exemplul 6.2.: Se cere să se analizeze asimetria seriei din exemplul anterior pe baza coeficientului de asimetrie în raport cu mediana.

Rezolvare: Determinarea coeficientului presupune calculul prealabil al valorii mediane (media aritmetică, şi abaterea medie pătratică au fost calculate în exemplul anterior) . În tabelul 6.5. sunt prezentate calculele intermediare pentru determinarea valorii mediane. Unitatea mediană are valoarea:

$$U_x^{M_e} = \frac{\left(\sum_{i=1}^{K_x} n_i^x\right) + 1}{2} = \frac{50+1}{2} = 25{,}5$$

Drept interval median a fost desemnat intervalul (80 ; 120].

Valoarea mediană reprezintă:

$$M_e^x = X_{M_e - 1} + d_x \frac{U_x^{M_e} - N_{X_{M_e-1}}^C}{n_{M_e}^x} = 80 + 40\frac{25{,}5 - 14}{15} = 110{,}7$$ mii bucăţi

Rezultă o valoare a coeficientului de asimetrie în raport cu mediana:

$$C_{as_x}^{M_e} = \frac{3\left(\overline{X} - M_e^x\right)}{\sigma_x} = \frac{3(107{,}2 - 110{,}7)}{47{,}1} = -0{,}22,$$ ceea ce semnifică o asimetrie slabă.

Tabelul 6.5. Valori intermediare utilizate în calculul valorii mediane

Nr. crt.	**Interval de variație [mii buc]**	**Frecvență absolută** (n_i^x)	**Frecvență absolută cumulată** ($N^C_{X_{M_e-1}}$)
(0)	(1)	(2)	(3)
1	(0 ; 40]	5	5
2	(40 ; 80]	9	14
3	(80 ; 120]	15	29
4	(120 ; 160]	14	43
5	(160 ; 200]	7	50
6	Total	50	×
7	Simbol pentru total	$\sum_{i=1}^{K_x} n_i^x$	×

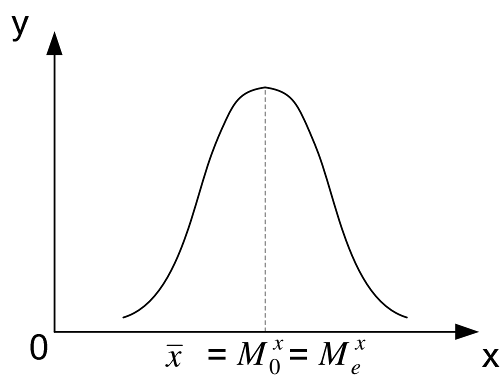
6.3. Boltirea distribuțiilor heterograde

6.3.1. Conceptul de boltire a unei distribuții heterograde

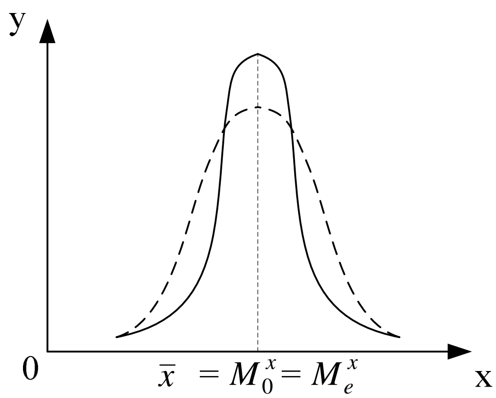
Boltirea (numită şi *kurtosisul*) unei distribuții heterograde este o trăsătură care se referă la aplatizarea curbei asociate seriei. De regulă, acest aspect este folosit în aprecierea gradului în care o serie unimodală se apropie de distribuția normală. În acest scop, se ia drept bază curba specifică unei repartiții normale, definindu-se în raport cu aceasta trei tipuri de distribuții:

- *distribuții mezokurtice*, pentru care curbele de frecvențe sunt asemănătoare, în ceea ce priveşte aplatizarea, unei curbe de distribuție normală (fig. 6.2.a);
- *distribuții leptokurtice,* la care curbele de frecvențe sunt mai ascuțite față de curba unei distribuții normale (fig. 6.2.b);
- *distribuții platykurtice*, pentru care curbele de frecvențe sunt mai turtite decât curba unei distribuții normale (fig. 6.2.c).

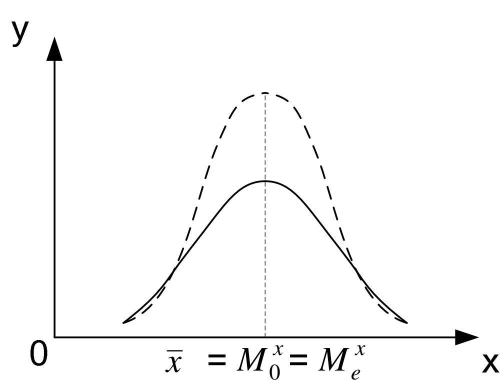
În general, se apreciază boltirea seriilor simetrice sau cu o asimetrie slabă şi relativ omogene, pentru celelalte serii comparația cu o distribuție normală fiind mai puțin relevantă.



a) Distribuţie mezokurtică



b) Distribuție loptokurtică



c) Distribuţie platykurtică

Fig. 6.2. Tipuri de distribuții în raport cu aplatizarea curbelor

Reprezentările grafice ale distribuțiilor, cu toarte că evidențiază deosebirile dintre cele trei tipuri de distribuții, nu permit, totuşi, cuantificarea gradului în care o distribuție se apropie de legea de repartiție normală. Din acest motiv, într-o cercetare statistică se recurge, de regulă, la exprimarea boltirii prin mărimi numerice.

6.3.2. Evaluarea boltirii unei distribuții heterograde

În acest subcapitol, înainte de a trece la prezentarea propriu-zisă a unei mărimi ce caracterizează boltirea, considerăm necesar să definim în prealabil noțiunea de *momente centrate ale distribuțiilor heterograde*. **Momentul centrat de ordin p** al unei distribuții heterograde este o mărime notată cu μ_p şi dată de relația:

$$\mu_p = \frac{\sum_{i=1}^{K_x} \left(x_i^{'} - \overline{x}\right)^p \times n_i^x}{\sum_{i=1}^{K_x} n_i^x} \qquad (6.3.)$$

Pe baza momentelor centrate ale unei distribuții heterograde poate fi determinat un indicator de apreciere a boltirii, numit *coeficientul pearsonian al boltirii*. Această mărime, notată cu β_{x_2} poate fi calculată raportând momentul centrat de ordinul patru la pătratul momentului centrat de ordinul doi (adică varianța seriei):

$$\beta_{x_2} = \frac{\mu_{X_4}}{\mu_{X_2}^2} \qquad (6.4.)$$

Valoarea acestui coeficient are următoarele semnificații:
- pentru $\beta_{x_2} < 3$, distribuția este platykurtică;
- pentru $\beta_{x_2} = 3$, distribuția este mezokurtică;
- pentru $\beta_{x_2} > 3$, distribuția este leptokurtică.

Exemplul 6.3. În tabelul 6.6. este prezentată o distribuție heterogradă care descrie productivitatea orară a muncii la un grup de 100 de angajați au unei firme. Se cere să se aprecieze boltirea seriei.

Tabelul 6.6. Repartizarea angajaților unei firme în funcție de productivitatea orară a muncii

Nr. crt.	**Interval de variație [mii buc]**	**Frecvență absolută** (n_i^x)
(0)	(1)	(2)
1	(2 ; 4]	10
2	(4 ; 6]	25
3	(6 ; 8]	30
4	(8 ; 10]	25
5	(10 ; 12]	10

Rezolvare: Determinarea coeficientului impune calculul prealabil al mediei aritmetice, al varianței şi al momentului centrat de ordinul patru. valorile intermediare utilizate în determinarea acestor mărimi sunt prezentate în tabelele 6.7. şi 6.8.

Tabelul 6.7. Valori intermediare utilizate în calculul mediei aritmetice

Nr. crt.	**Interval de variație [RON/h]**	**Frecvență absolută** (n_i^x)	**Centru de interval** ($x_i^{'}$) **[RON/h]**	$x_i^{'} n_i^x$ **[RON/h]**
(0)	(1)	(2)	(3)	(4) = (3) × (2)
1	(2 ; 4]	10	3	30
2	(4 ; 6]	25	5	125
3	(6 ; 8]	30	7	210
4	(8 ; 10]	25	9	225
5	(10 ; 12]	10	11	110
6	Total	100	×	700
7	Simbol pentru total	$\sum_{i=1}^{K_x} n_i^x$	×	$\sum_{i=1}^{K_x} n_i^x x_i^{'}$

Media aritmetică a seriei are valoarea

$$\overline{X}=\frac{\sum_{i=1}^{K_x} x_i^{'} n_i^x}{\sum_{i=1}^{K_x} n_i^x}=\frac{700}{100}=7 \text{ RON/h.}$$

Momentul centrat de ordinul doi (varianța) reprezintă:

$$\mu_{x_2}=\frac{\sum_{i=1}^{K_x}\left(x_i^{'}-\overline{x}\right)^2 \times n_i^x}{\sum_{i=1}^{K_x} n_i^x}=\frac{520}{100}=5{,}2 \text{ (RON/h)}^2$$

Tabelul 6.8. Valori intermediare utilizate în calculul varianței şi momentului centrat de ordinul patru

Nr. crt.	**Interval de variație [RON/h]**	**Frecvență absolută** (n_i^x)	($x_i^{'}$-n_i^x) **[RON/h]**	($x_i^{'}$-n_i^x)$^{\mathbf{2}}$ n_i^x **[(RON/h)^2]**	($x_i^{'}$-n_i^x)$^{\mathbf{4}}$ n_i^x **[(RON/h)^4]**
(0)	(1)	(2)	(3)	(4) = (3)^2 × (2)	(5) = (3)^4 × (2)
1	(2 ; 4]	10	-4	160	2560
2	(4 ; 6]	25	-2	100	400
3	(6 ; 8]	30	-	-	-
4	(8 ; 10]	25	2	100	400
5	(10 ; 12]	10	4	160	2560
6	Total	100	×	520	5920
7	Simbol pentru total	$\sum_{i=1}^{K_x} n_i^x$	×	$\sum_{i=1}^{K_x}\left(x_i^{'}-\overline{x}\right)^2 n_i^x$	$\sum_{i=1}^{K_x}\left(x_i^{'}-\overline{x}\right)^4 n_i^x$

Momentul centrat de ordinul patru are valoarea:

$$\mu_{x_4}=\frac{\sum_{i=1}^{K_x}\left(x_i^{'}-\overline{x}\right)^4 \times n_i^x}{\sum_{i=1}^{K_x} n_i^x}=\frac{5920}{100}=59{,}2 \text{ (RON/h)}^4$$

Rezultă:

$\beta_{x_2} = \frac{\mu_{X_4}}{\mu_{X_2}^2} = \frac{59{,}2}{(5{,}2)^2} = 2{,}19$, ceea ce indică o distribuție platykurtică.

Capitolul 7 - Legile fenomenelor colective

7.1. Caracteristici ale legilor fenomenelor colective

Unul dintre scopurile majore ale cercetărilor statistice este reprezentat de identificarea legilor ce guvernează fenomenele colective. Pe baza acestora pot fi previzionate rezultatele posibile sau pot fi apreciate influențele unor factori. În cadrul statisticii matematice au fost propuse mai multe tipuri de funcții care reflectă legile ce acționează asupra fenomenelor colective. Aceste funcții nu pot fi însă decât nişte simplificări ale realității întrucât nu iau în calcul decât aspectele considerate esențiale ale fenomenelor studiate. În aceste condiții pot fi definite două forme ale valorilor parametrilor unui fenomen colectiv:

- *valori teoretice*, date de funcțiile matematice prin care sunt reprezentate legile asociate fenomenului;
- *valori empirice*, care reflectă datele statistice culese asupra fenomenului.

Valorile teoretice pot fi interpretate drept rezultate ale factorilor esențiali de influență în timp ce valorile empirice reflectă influența tuturor factorilor: atât a celor esențiali cât şi a celor considerați nerelevanți. Dacă o lege asociată unui fenomen colectiv reflectă în mare măsură realitatea atunci este de aşteptat ca impactul factorilor considerați nerelevanți să nu fie semnificativ, astfel încât valorile teoretice să fie apropiate de cele empirice. În această logică, se poate aprecia că o valoare teoretică este o aproximare a unei valori empirice obținută prin neglijarea efectelor factorilor considerați nerelevanți (din acest motiv, valorile teoretice sunt numite şi *valori ajustate*).

În practică, determinarea funcției care reflectă o lege asociată unui fenomen colectiv se desfăşoară, de regulă, în trei etape:

1) alegerea formei funcției;
2) determinarea parametrilor funcției;
3) evaluarea acurateței valorilor teoretice.

1) Pentru **alegerea formei funcției** se porneşte de la unele aspecte ale seriei statistice care prezintă valorile empirice: omogenitatea, asimetria, boltirea ş.a.m.d. Aceste aspecte pot fi relevate fie prin calcule analitice fie prin reprezentări grafice.

2) Pentru **determinarea parametrilor funcției** se porneşte, de regulă, de la premisa că valorile teoretice ale funcției trebuie să fie cât mai apropiate de valorile empirice. În practică, pentru îndeplinirea acestei condiții sunt folosite câteva procedee matematice de minimizare a diferențelor dintre cele două tipuri de valori.

3) Prin **evaluarea acurateței valorilor teoretice** se apreciază în fapt în ce măsură funcția reflectă manifestarea fenomenului studiat şi, implicit, ce încredere se poate avea în calculele făcute pe baza funcției. De regulă, în această operațiune sunt luate ca reper diferențele dintre valorile teoretice şi cele empirice.

7.2. Distribuția normală

7.2.1. Proprietăți ale distribuției normale

O distribuție normală caracterizează fenomenele ce sunt influențate de mai mulți factori, dintre care nici unul nu are un impact predominant. Se consideră că această trăsătură este comună celor mai multe dintre fenomenele colective desfăşurate în condiții naturale, ceea ce face ca distribuția normală să fie folosită frecvent în cercetările statistice.

Unei serii statistice ideale, ale cărei valori ar urma o distribuție normală, îi poate fi asociată o curbă de frecvențe cu ecuația:

$$y = \frac{1}{\sigma_x \cdot \sqrt{2\pi}} \cdot e^{-\frac{(x-\bar{x})^2}{2\sigma_x^2}} \qquad (7.1.)$$

Din ecuația curbei de frecvențe rezultă mai multe proprietăți. Astfel, curba este simetrică, în formă de clopot, cu un maxim în dreptul mediei aritmetice ($\bar{x}$) în raport cu care valorile scad continuu la stânga şi la dreapta, tinzând asimptotic către axa absciselor (fig. 7.1.). În dreptul coordonatelor de abscise $\bar{x}-\sigma_x$ şi $\bar{x}+\sigma_x$, curba are două puncte de inflexiune. Se poate demonstra că în intervalul $[\bar{x}-\sigma_x\ ;\ \bar{x}+\sigma_x]$ se află concentrată 68,26% din suprafața delimitată de curba de frecvențe, ceea ce indică o omogenitate semnificativă a seriei. În plus, din perspectiva boltirii, curba are semnificația unei distribuții mezokurtice.

Din aceeaşi ecuaţie (7.1.) reiese şi faptul că o distribuţie normală poate fi definită prin doi parametri: media aritmetică $\bar{x}$ şi varianţa σ_x^2.

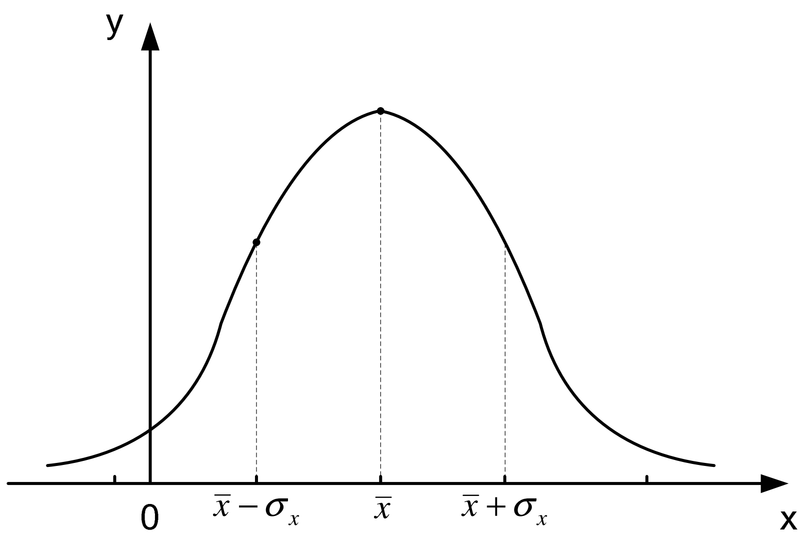


Fig. 7.1. Curba frecvenţelor asociată unei distribuţii normale

7.2.2. Evaluarea probabilităţilor prin distribuţii normale

Ecuaţia curbei frecvenţelor unei serii statistice cu distribuţia normală poate fi folosită pentru atribuirea de probabilităţi în manifestarea unui fenomen colectiv. În acest scop, seria statistică trebuie transpusă într-o variabilă aleatoare de tip continuu, care asociază probabilităţi intervalelor valorice ale seriei prin intermediul frecvenţelor relative. Valorile variabilei aleatoare vor avea aceeaşi medie aritmetică $\bar{x}$ şi aceeaşi varianţă σ_x^2 pe care le are şi seria statistică din care provine. De asemenea, funcţia densităţii probabilistice are ecuaţia curbei frecvenţelor seriei, fiind definită pe intervalul (- ∞ ; + ∞). În aceste condiţii, probabilitatea ca o valoare *X* a variabilei aleatoare să fie mai mare decât un număr *x* poate fi calculată prin formula:

$$P_{(X<x)} = \int_{-\infty}^{x} y_{(x)} dx = \int_{-\infty}^{x} \frac{1}{\sigma_x \cdot \sqrt{2\pi}} \cdot e^{-\frac{(X-\bar{x})}{2\sigma_x}} dx \qquad (7.2.)$$

Calculul integralei din această formulă poate fi destul de dificil, ceea ce a condus la dezvoltarea unor metode mai simple de determinare a probabilităților pentru distribuțiile normale. Cea mai des utilizată dintre acestea are la bază folosirea unei aşa-numite *distribuții normale standard* – un caz particular al distribuțiilor normale care are media aritmetică nulă şi abaterea medie pătratică egală cu 1 (fig. 7.2.).

Transformarea unei distribuții normale oarecare X într-o distribuție normală standard Z are la bază relația:

$$Z = \frac{X - \overline{x}}{\sigma_x} \tag{7.3.}$$

Pentru o distribuție normală standard pot fi stabilite valori tabelate ale probabilității ca valorile distribuției să fie mai mari decât un număr z_i (această probabilitate este proporțională cu suprafața haşurată din figura 7.3.). În tabelul 7.1. sunt prezentate câteva astfel de valori tabelate. Evident, dacă se cunoaşte o astfel de probabilitate, $P(Z > z_i)$ se poate determina şi probabilitatea evenimentului opus:

$$P(Z \le z_i) = 1 - P(Z > z_i) \tag{7.4.}$$

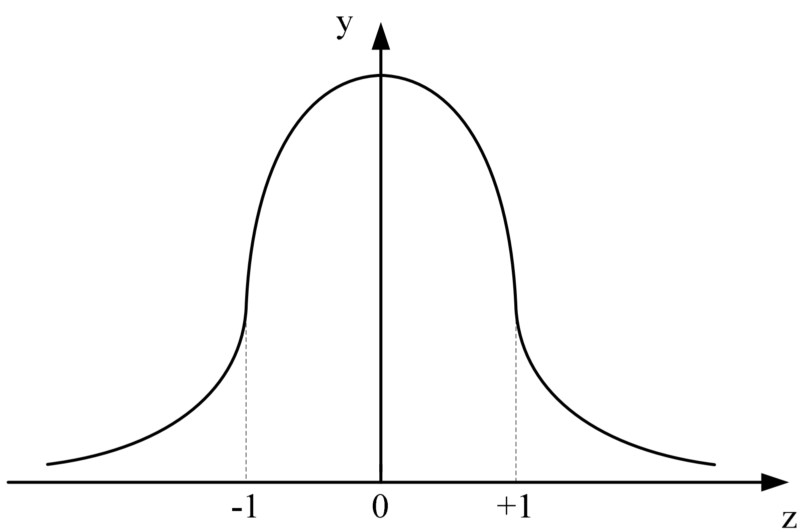


Fig. 7.2. Curba de frecvențe a distribuției normale standard

Se poate demonstra că probabilitatea ca valorile unei distribuții să se afle într-un interval (z_1 ; z_2) este dată de relația:

$$P(z_1 < Z < z_2) = P(Z > z_1) - P(Z > z_2) \tag{7.5.}$$

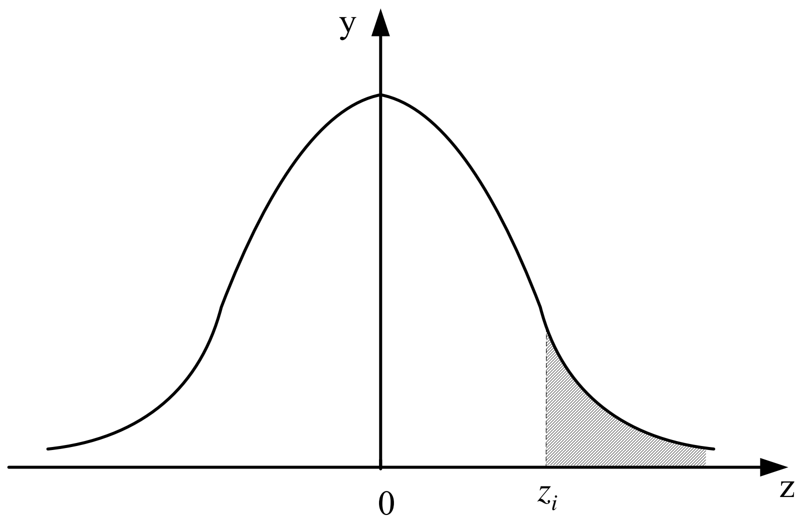


Fig. 7.3. Reprezentarea grafică a probabilității ca valorile unei distribuții normale standard să fie mai mari decât un număr z

Simetria graficului distribuției normale standard față de punctul de coordonată zero pe abscisă face ca suprafața delimitată la dreapta de un număr pozitiv z_i să fie egală cu suprafața delimitată la stânga de un număr negativ, egal cu primul în valoare absolută (fig. 7.4.). Dacă se ia în considerare relația dintre aceste suprafețe şi probabilitățile asociate distribuției normale standard rezultă:

$$P(Z > z_i) = P(Z < -z_i) \qquad (7.6.)$$

Tabelul 7.1. Valori tabelate ale probabilităților specifice unei distribuții normale standard

z_i	0	0,25	0,5	0,75	1
$P(Z > z_i)$	0,5000	0,4013	0,3085	0,2266	0,1587

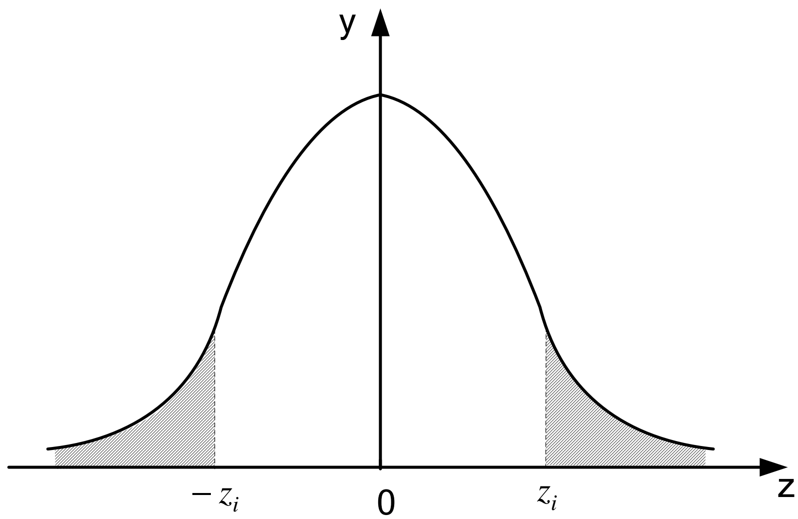


Fig. 7.4. Reprezentarea grafică a probabilităților $P(Z > z_i)$ şi $P(Z < -z_i)$

Estimarea probabilităților pe baza distribuțiilor normale este folosită destul de frecvent în practică pentru previziunea fenomenelor cărora le poate fi asociată o astfel de lege. În acest scop este necesară cunoaşterea celor doi parametri ce definesc o distribuție normală: media aritmetică şi varianța.

Exemplul 7.1. Managerii unei firme evaluează rentabilitatea unui sortiment de produs. Pe baza prețului şi a costurilor a fost identificat un prag de rentabilitate la nivelul vânzărilor de 80 mii bucăți, sub care realizarea sortimentului de produs devine nerentabilă. Din datele culese asupra cererii potențiale a rezultat că în anul viitor vânzările vor urma o distribuție normală cu media aritmetică de 90 mii bucăți şi varianța de 400 (mii bucăți)^2. Se cere să se estimeze probabilitatea ca realizarea sortimentului de produs să se soldeze cu pierderi.

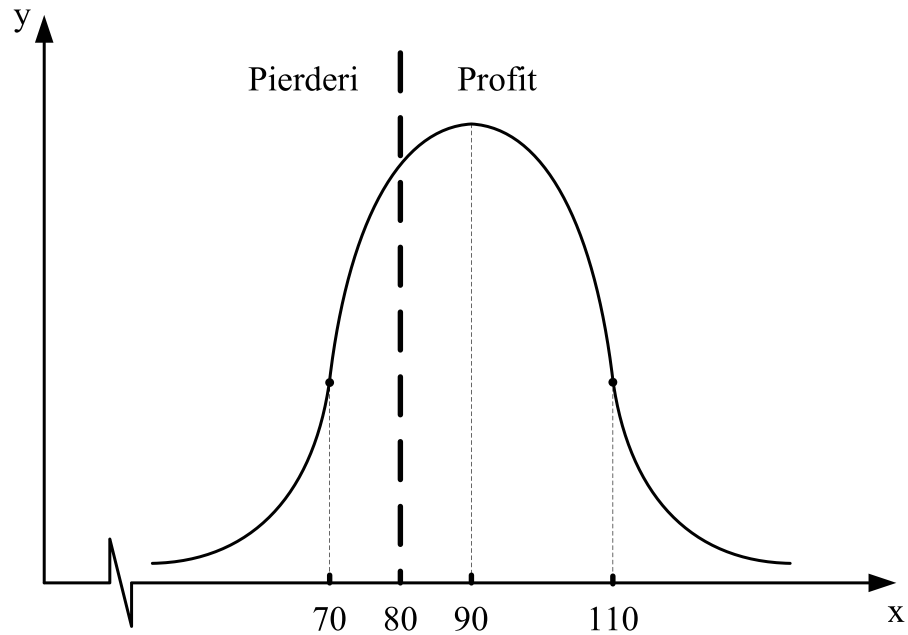


Fig. 7.5. Distribuția normală a vânzărilor unui sortiment de produs

Rezolvare: A estima probabilitatea ca realizarea produsului să se soldeze cu pierderi înseamnă, în fapt, a calcula probabilitatea ca nivelul vânzărilor să fie mai mic decât 80 mii bucăți (fig. 7.5.). Pentru aceasta este necesară determinarea, în prealabil, a abaterii medii pătratice:

$$\sigma_x = \sqrt{\sigma_x^2} = \sqrt{400} = 20 \text{ mii bucăți}$$

Calculul probabilității pe baza valorilor tabelate necesită trecerea la o distribuție normală standard prin transformarea:

$$z_1 = \frac{x_1 - \bar{x}}{\sigma_x} = \frac{80 - 90}{20} = -0{,}5$$

Probabilitatea ca vânzările să fie mai mici de 80 mii bucăți este echivalentă, pentru distribuția normală standard, cu probabilitatea ca valoarea variabilei aleatoare Z să fie mai mică decât $z_1 = -0{,}5$. După cum se poate remarca în figura 7.6., această probabilitate este de fapt egală cu probabilitatea ca valorile distribuției normale standard să fie mai mari decât $z_2 = 0{,}5$. Valoarea tabelată a acesteia este 0,3085 ceea ce indică că o probabilitate de 30,85% ca realizarea sortimentului de produs să se soldeze cu pierderi.

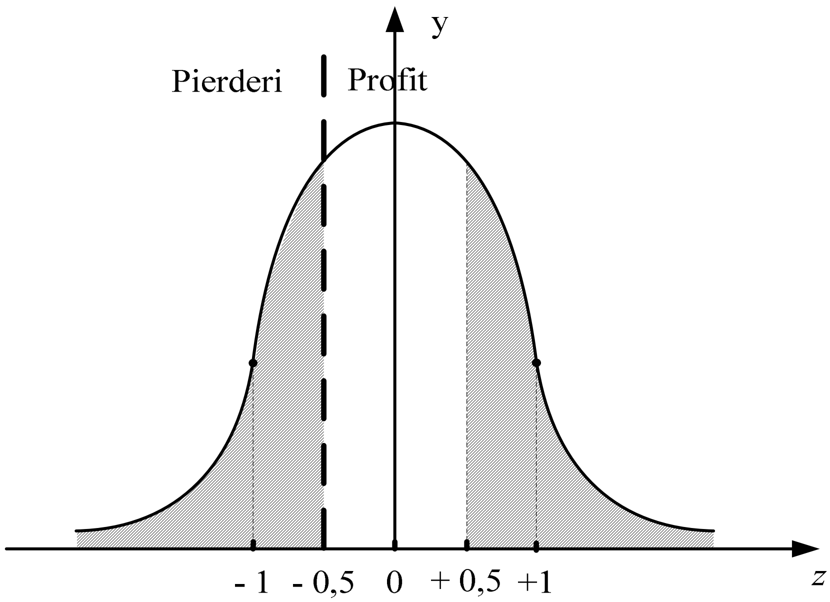


Fig. 7.6. Distribuția normală standard a vânzărilor unui sortiment de produs

Capitolul 8 - Cercetarea statistică prin sondaj

8.1. Coordonate ale cercetării statistice prin sondaj

Într-un capitol anterior a fost prezentat sondajul drept o modalitate de culegere a datelor statistice ce vizează doar o parte (numită *eşantion*) din populaţia studiată. În acest caz, valorile mărimilor ce caracterizează populaţia nu pot fi cunoscute cu certitudine ci sunt doar estimate pe baza valorilor determinate pentru eşantion. Trecerea de la valorile certe ale parametrilor unui eşantion la valorile probabile ale parametrilor populaţiei este cunoscută sub denumirea de *inferenţă statistică*.

O cercetare statistică riguroasă presupune cunoaşterea gradului de încredere ce se poate avea în valorile estimate ale parametrilor ce caracterizează populaţia studiată. Din acest motiv, estimările sunt transpuse adeseori sub forma unor distribuţii probabilistice.

Despre valorile parametrilor calculaţi pentru un eşantion se spune că au calitatea de *estimatori ai valorilor parametrilor populaţiei*, ceea ce înseamnă că pot servi în estimarea acestora. Un estimator este numit *nedeplasat* atunci când valoarea sa este egală cu media aritmetică a distribuţiei probabilistice asupra parametrului asociat populaţiei. Atunci când cele două valori diferă, estimatorul este numit *deplasat*. Drept parametri de caracterizare a unei populaţii sunt folosiţi diferiţi indicatori statistici, dintre care se remarcă prin frecvenţa utilizării media aritmetică (notată cu $\bar{x}_s$ în cazul sondajului şi cu μ_s în cazul populaţiei) şi proporţia unei caracteristici în ansamblul populaţiei (notată cu p_e în cazul eşantionului şi cu p_p în cazul populaţiei). Pentru cele două mărimi, valorile determinate pentru eşantioane pot fi considerate drept estimatori nedeplasaţi pentru valorile probabile ale parametrilor populaţiei.

Un aspect important al inferenţei statistice este reprezentat de cuantificarea acurateţei estimărilor. Măsura preciziei unei cercetări statistice prin sondaj poate fi stabilită exact prin intermediul unui indicator numit *eroare efectivă de inferenţă*, notat cu e_{ef} şi dat de relaţia:

$$e_{ef} = \theta - \hat{\theta} \tag{8.1.}$$

în care:

- θ este valoarea reală a unui parametru ce caracterizează o populaţie statistică;

- $\hat{\theta}$ este valoarea estimată a parametrului pe baza datelor culese prin sondaj.

Din nefericire, cel mai adesea eroarea efectivă de inferență nu poate fi calculată întrucât valoarea reală a parametrului ce caracterizează populația este necunoscută (dacă ar fi cunoscută nu ar mai fi nevoie de sondaj). În aceste condiții, eroarea efectivă de inferență poate fi doar estimată. În evaluarea acesteia trebuie luați în considerare câțiva factori care o pot influența:

1. reprezentativitatea eşantionului pentru populația statistică din care provine;
2. volumul eşantionului;
3. dispersia populației studiate.

1. Un eşantion este considerat **reprezentativ** atunci când structura sa este asemănătoare cu aceea a populației din care provine. Şansele ca o valoare estimată prin sondaj să fie apropiată de valoarea reală sunt cu atât mai mari cu cât eşantionul utilizat este mai reprezentativ. În situația, oarecum ideală, în care valorile unei caracteristici au aceleaşi proporții pentru eşantionul folosit şi pentru populația studiată, parametrul estimat este chiar egal cu parametrul real al populației.

2. **Volumul unui eşantion** este o mărime, notată cu n, care reprezintă numărul de unități statistice conținut de eşantion. În principiu, atunci când volumul unui eşantion creşte, sporesc şi şansele ca valoarea estimată a unui parametru să fie apropiată de cea reală. În cazul extrem, în care numărul de unități statistice al eşantionului ar fi egal cu numărul unităților statistice ale populației (în acest caz sondajul s-ar transforma însă într-un recensământ) ar exista certitudinea că valoarea estimată este egală cu cea reală. Volumul unui eşantion poate fi luat în considerare şi prin prisma ponderii pe care o deține în volumul populației. Se consideră că acuratețea estimării este cu atât mai mare cu cât această pondere este mai mare.

3. **Dispersia populației studiate** poate cauza valori mari ale erorii efective de sondaj. Altfel spus, cu cât populația studiată este mai omogenă, cu atât sunt mai mari şansele ca valorile estimate să fie apropiate de cele reale. În situația extremă în care toate unitățile populației statistice au aceeaşi valoare putem fi siguri că, indiferent cum este alcătuit eşantionul (acesta poate fi constituit chiar dintr-o singură unitate) valoarea estimată este egală cu valoarea reală.

În raport cu cei trei factori pot fi stabilite distribuții probabilistice asupra valorilor erorilor efective de estimare. Pe baza acestora se pot determina aşa numite *intervale de încredere*, care sunt intervale în interiorul cărora putem aprecia, cu probabilități cunoscute, că se află valori reale ale parametrilor populației studiate. Probabilitatea ca valoarea unui parametru să se afle într-un interval de încredere este numită *nivel de încredere*. Unele proprietăți ale distribuțiilor probabilistice fac ca adeseori în practică să se prefere determinarea nivelului de încredere pe baza probabilității ca valoarea parametrului să nu se afle în intervalul de încredere. Această probabilitate, numită *nivel de semnificație*, este notată cu α în timp ce nivelul de încredere, care corespunde unui eveniment opus, este notat cu $1 - \alpha$.

Atunci când în cadrul inferenței statistice sunt utilizați estimatori nedeplasați, valorile acestora pot fi stabilite, pentru simplificarea calculelor probabilistice, în centrul intervalelor de încredere. Limitele unui interval de încredere se vor afla, în acest caz, la o distanță egală de estimator. Această distanță, notată cu e_1^α şi numită *eroare limită de inferență*, este în fapt o estimare, pentru un nivel de semnificație α, a erorii efective de inferență. În aceste condiții, intervalul de încredere are forma $[\hat{\theta} - e_1^\alpha; \hat{\theta} + e_1^\alpha]$ iar probabilitatea ca valoarea reală a parametrului populației să se afle în acest interval este egală cu nivelul de încredere:

$$P\left(\hat{\theta} - e_1^\alpha \leq \theta \leq \hat{\theta} + e_1^\alpha\right) = 1 - \alpha \qquad (8.2.)$$

Eroarea limită de inferență, care reprezintă un indiciu al acurateței estimării poate fi evaluată pe baza celor trei factori care influențează eroarea efectivă de inferență: volumul eşantionului, reprezentativitatea acestuia şi dispersia populației. În situația, destul de frecventă în practică, în care dispersia populației nu este cunoscută, aceasta poate fi estimată pe baza dispersiei eşantionului. Cunoscând impactul acestor factori se poate alcătui un eşantion astfel încât acuratețea inferenței să se situeze deasupra unui nivel minim acceptabil.

Adeseori în practică este mai util ca în loc de a se stabili un interval de încredere pentru un parametru să se determine probabilitatea ca valoarea acestuia să fie mai mică sau mai mare decât un anumit nivel. În acest scop pot fi folosite proprietățile distribuției de probabilități asociată inferenței statistice.

8.2. Tipologia sondajelor statistice

Sondajele folosite în cercetările statistice îmbracă forme foarte variate, în raport cu scopurile urmărite şi cu resursele disponibile. În acest subcapitol vor fi prezentate succint câteva din tipurile de sondaje, relevante din perspectiva inferenței statistice, grupate în raport cu două criterii:

a) volumul eşantionului;
b) procedeul de alcătuire a eşantionului.

a) În funcție de **volumul eşantionului** se diferențiază două tipuri de sondaje:

a_1) **sondaje de volum mare**, la care eşantioanele au un volum mai mare de 30 de unități statistice;

a_2) **sondaje de volum redus**, ale căror eşantioane au un volum de cel mult 30 de unități statistice.

Se consideră că estimările realizate pe baza sondajelor de volum mare au o acuratețe superioară celor care utilizează sondaje de volum redus. În schimb, sondajele de volum redus sunt, de regulă, mai uşor de organizat şi mai puțin costisitoare față de cele de volum mare.

b) În raport cu **procedeul de alcătuire a eşantionului** se delimitează trei tipuri de sondaje:

b_1) **sondaje aleatoare**, la care unitățile statistice ale eşantioanelor sunt alese în mod întâmplător;

b_2) **sondaje dirijate**, la care unitățile statistice sunt stabilite în funcție de trăsăturile populației studiate, relevante în raport cu scopul cercetării statistice;

b_3) **sondaje mixte**, care sunt combinații ale sondajelor întâmplătoare şi ale sondajelor dirijate (de exemplu, o populație poate fi împărțită, în raport cu trăsăturile sale, în mai multe grupe, iar pentru fiecare dintre acestea este alcătuit, în mod întâmplător, câte un eşantion).

Se consideră că sondajele dirijate sau mixte asigură, în comparație cu sondajele aleatoare, un grad mai înalt de reprezentativitate a eşantioanelor, ceea ce conduce la o acuratețe mai mare a inferenței statistice. Totuşi, alcătuirea eşantioanelor în raport cu trăsăturile relevante ale populației (care nu sunt întotdeauna uşor de identificat şi de evaluat) poate induce o complexitate deosebită cercetării prin sondaj.

8.3. Inferența statistică pentru sondajele de volum mare

8.3.1. Fundamentele teoretice ale inferenței sondajelor de volum mare

Inferența statistică în cazul sondajelor de volum mare are la bază aşa-numita *teoremă limită centrală*. Aceasta stipulează că dacă dintr-o populație statistică se constituie un număr suficient de mare de eşantioane de volum n atunci media aritmetică a acestora are o distribuție normală sau, cel puțin, apropiată de cea normală, în două situații: dacă şi populației îi poate fi asociată o lege de distribuție normală, sau dacă n tinde la infinit. Media aritmetică a distribuției normale a eşantioanelor va fi egală cu media aritmetică a populației statistice, iar abaterea medie pătratică (numită şi *eroarea standard*) notată cu σ_s, poate fi calculată prin relația:

$$\sigma_s = \frac{\sigma_p}{\sqrt{n}} \qquad (8.3.)$$

unde σ_p este abaterea medie pătratică a populației studiate.

Condiția de infinitate a volumului eşantionului este atenuată în practică, unde se consideră că este suficient ca sondajele să fie de volum mare (adică n să fie mai mare decât 30) pentru ca media aritmetică a eşantioanelor să urmeze o distribuție aproximativ normală.

8.3.2. Determinarea intervalelor de încredere asupra mediei aritmetice

Pentru determinarea intervalelor de încredere asupra mediei aritmetice sunt folosite variate procedee, care se diferențiază în raport cu condițiile concrete în care se aplică. În acest subcapitol vor fi prezentate succint modalitățile de stabilire a intervalelor de încredere pentru trei situații:

a) în condițiile cunoaşterii dispersiei populației;
b) în condițiile în care dispersia populației nu este cunoscută ci doar estimată;

c) în condițiile în care eşantionul are o pondere semnificativă în ansamblul populației.

8.3.2.1. Determinarea intervalelor de încredere în condițiile cunoaşterii dispersiei populației

În situația în care dispersia populației este cunoscută, intervalele de încredere pot fi stabilite pe baza proprietăților unei distribuții normale cu media aritmetică μ_s şi abaterea medie pătratică σ_s (fig. 8.1.)

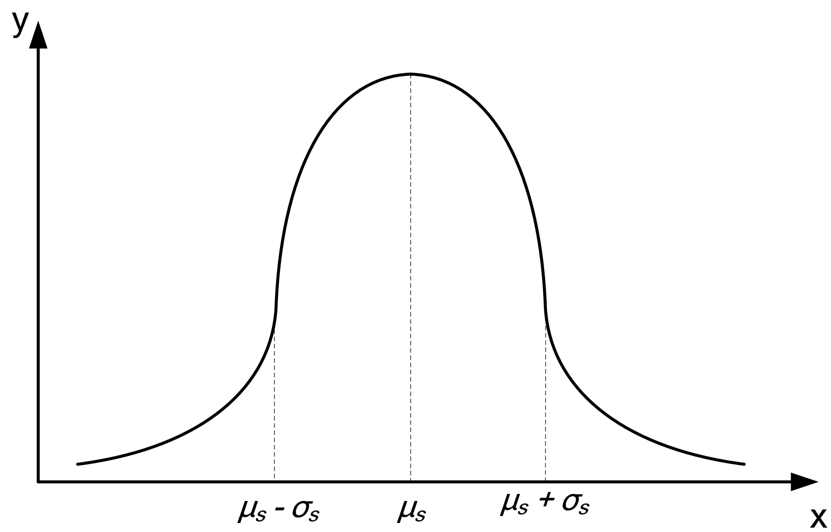


Fig. 8.1. Distribuția normală a mediilor aritmetice ale eşantioanelor

Una dintre aceste proprietăți facilitează calculul suprafețelor delimitate de graficul distribuției normale şi de linii verticale trasate la distanțe egale de media aritmetică a distribuției. O astfel de suprafață, care reprezintă în fapt probabilitatea, notată cu 1 – σ (nu întâmplător se foloseşte simbolul asociat nivelului de încredere) ca media aritmetică a unui eşantion să se găsească într-un interval de valori ce are în centru media aritmetică a populației, este dată de relația:

$$1-\alpha = P\left(\mu_s - z_i^{\alpha} \cdot \sigma_s \leq \overline{x} \leq \mu_s + z_i^{\alpha} \cdot \sigma_s\right) \quad (8.4.)$$

unde z_i^{α} este o mărime numită coeficient de încredere.

Valorile mărimii z_i^{α} pot fi determinate pe baza proprietăților distribuției normale, ceea ce simplifică foarte mult calculele probabilistice. Aşa cum s-a menționat în capitolul anterior, 68,26%

din suprafaţa delimitată de graficul distribuţiei normale se află în intervalul $\left[\bar{x}-\sigma_x; x+\sigma_x\right]$, ceea ce înseamnă că pentru z_i^α = 1 vom avea:

$$P\left(\mu_s-\sigma_s \leq \bar{x} \leq \mu_s+\sigma_s\right) = 0{,}6826$$

Acest nivel de probabilitate nu oferă însă o siguranţă prea mare pentru inferenţa statistică. În practică, în cadrul estimărilor se operează de regulă cu niveluri de probabilitate mai mari de 90%, în special cu valorile de 95% şi 99%. Probabilităţii de 95% îi corespunde o valoare z_i^α = 1,96, ceea ce înseamnă că:

$$P\left(\mu_s-1{,}96\cdot\sigma_s \leq \bar{x} \leq \mu_s+1{,}96\cdot\sigma_s\right) = 0{,}95$$

De asemenea, probabilităţii de 99% îi corespunde o valoare z_i^α = 2,576, de unde rezultă:

$$P\left(\mu_s-2{,}576\cdot\sigma_s \leq \bar{x} \leq \mu_s+2{,}576\cdot\sigma_s\right) = 0{,}99$$

Relaţia (8.4.) permite calculul probabilităţii ca media aritmetică a unui eşantion să se afle în interiorul unui interval stabilit pe baza mediei aritmetice a populaţiei. Inferenţa statistică vizează însă mai degrabă stabilirea probabilităţii ca media aritmetică a populaţiei să se afle într-un interval de valori determinat pe baza mediei aritmetice a unui eşantion. În acest scop, relaţia (8.4) este modificată pe baza următoarelor transformări:

- inegalitatea $\mu_s - z_i^\alpha\cdot\sigma_s \leq \bar{x}$ este echivalentă cu inegalitatea $\mu_s \leq \bar{x} + z_i^\alpha\cdot\sigma_s$;
- inegalitatea $\bar{x} \leq \mu_s + z_i^\alpha\cdot\sigma_s$ este echivalentă cu inegalitatea $\mu_s \geq \bar{x} - z_i^\alpha\cdot\sigma_s$

Rezultă astfel relaţia care stă la baza determinării unui interval de încredere pentru un nivel de semnificaţie dat:

$$P\left(\bar{x} - z_i^\alpha\cdot\sigma_s \leq \mu_s \leq \bar{x} + z_i^\alpha\cdot\sigma_s\right) = 1-\alpha$$

sau:

$$P\left(\bar{x} - z_i^\alpha\cdot\frac{\sigma_p}{\sqrt{n}} \leq \mu_s \leq \bar{x} + z_i^\alpha\cdot\frac{\sigma_p}{\sqrt{n}}\right) = 1-\alpha \qquad (8.5.)$$

Relaţia (8.5.) poate fi considerată drept un caz particular al relaţiei (8.2.) de stabilire a nivelului de încredere a unui parametru, în care valoarea estimată este reprezentată de media aritmetică a eşantionului, valoarea reală este reprezentată de media aritmetică a populaţiei, iar eroarea limită de inferenţă este dată de produsul dintre coeficientul de încredere z_i^{α} şi abaterea medie pătratică a distribuției σ_s:

$$e_1^{\alpha} = z_i^{\alpha} \cdot \sigma_s \qquad (8.6.)$$

Determinarea intervalelor de încredere pe baza relației (8.5.) este destul de simplă în condițiile în care pot fi utilizate valori cunoscute ale coeficientului de încredere.

Exemplul 8.1. În cadrul unei firme producătoare de componente electronice se estimează durata medie de funcționare în regim intensiv a unui sortiment de produs. Testele au fost întreprinse asupra unui eşantion de 100 de produse rezultând pentru acestea o durată medie de funcționare de 400 de ore. Cunoscându-se că pentru întreaga producție a firmei abaterea medie pătratică a duratei de funcționare reprezintă 200 ore se cere să se determine intervalul de încredere pentru media aritmetică cu un nivel de încredere de 68,26%.

Rezolvare: În raport cu abaterea medie pătratică a populației şi cu volumul eşantionului se calculează valoarea erorii standard:

$$\sigma_s = \frac{\sigma_p}{\sqrt{n}} = \frac{20}{\sqrt{100}} = \frac{200}{10} = 20\,\text{h}$$

Nivelului de încredere 1 – α = 0,6826 îi corespunde o valoare tabelată

z_i^{α} = 1, de unde rezultă:

$$P(400 - 1\times 20 \le \mu_s \le 400 + 20\times 20) = 0{,}6826$$

ceea ce înseamnă că poate fi stabilit intervalul de încredere [380 ; 420] în care se află, cu o probabilitate de 68,26% durata medie de funcționare pentru toate produsele realizate de firmă.

În practică, situațiile în care se cunosc dispersiile populațiilor cercetate prin sondaj sunt destul de rare (pentru a fi cunoscută dispersia ar fi necesar să se cunoască şi media aritmetică astfel încât sondajul ar fi inutil). Din acest motiv, procedeele de determinare a intervalelor de încredere în condițiile cunoaşterii populației studiate au mai mult o semnificație teoretică.

8.3.2.2. Determinarea intervalelor de încredere pe baza dispersiei estimate

Atunci când nu se cunoaşte dispersia populației studiate, aceasta trebuie estimată pe baza dispersiei eşantionului. Drept estimator al abaterii medii pătratice a populației σ_p poate fi utilizată o mărime numită *abatere medie pătratică de sondaj*, notată cu S şi care poate fi calculată pe baza valorilor din eşantion prin formula:

$$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x}_s)^2}{n-1}} \qquad (8.7.)$$

Valoarea abaterii medii pătratice de sondaj este obţinută împărţind suma pătratelor abaterilor faţă de media aritmetică la $n - 1$ şi nu la numărul total de unităţi aşa cum se întâmplă în cazul abaterii medii pătratice a unei serii simple. Explicaţia vine din faptul că s-a constatat că valoarea astfel calculată este un estimator mai bun decât abaterea medie pătratică a valorilor eşantionului.

Procedeul determinării intervalelor de încredere pe baza estimărilor asupra dispersiei populaţiei este asemănător celui utilizat atunci când se cunoaşte dispersia reală a populaţiei, cu deosebirea că în formulele de calcul abaterea medie pătratică a populaţiei este înlocuită cu estimatorul acesteia, adică abaterea medie pătratică de sondaj:

$$P\left(\overline{x} - z_i^{\alpha} \cdot \frac{S}{\sqrt{n}} \leq \mu_s \leq \overline{x} + z_i^{\alpha} \cdot \frac{S}{\sqrt{n}} \right) \qquad (8.8.)$$

Exemplul 8.2. Pentru fundamentarea unei decizii asupra înfiinţării unei reţele de comercializare a produselor cosmetice s-a întreprins un studiu asupra cererii potenţiale din zonă. În acest scop s-a întreprins un sondaj asupra unui eşantion de 160 de persoane. Cheltuielile lunare pentru produsele cosmetice ale acestor persoane au o medie aritmetică de 30 RON şi o abatere medie pătratică de sondaj de 10 RON. Se cere să se determine cu o probabilitate de 99%, intervalul de încredere al cheltuielilor medii pentru întreaga populaţie din zonă.

Rezolvare: Nivelului de încredere $\alpha - 1 = 0,99$ îi corespunde un coeficient de încredere $z_i^{\alpha} = 2,576$. Rezultă:

$$P\left(\overline{x} - z_i^{\alpha} \cdot \frac{S}{\sqrt{n}} \leq \mu_s \leq \overline{x} + z_i^{\alpha} \cdot \frac{S}{\sqrt{n}} \right) = 1 - \alpha, \text{ adică,}$$

$$P\left(30 - 2,576 \times \frac{10}{\sqrt{160}} \leq \mu_s \leq 30 + 2,576 \times \frac{10}{\sqrt{160}} \right) = 0,99$$

ceea ce înseamnă că media aritmetică a cheltuielilor pentru întreaga populaţie din zonă se află, cu o probabilitate de 99%, în intervalul [27,96 ; 32,04].

În situația în care eşantionul ia forma unei distribuții heterograde, abaterea medie pătratică de sondaj poate fi calculată prin formula:

$$S = \sqrt{\frac{\sum_{i=1}^{k_x} \left(x_i^{'} - \bar{x}\right)^2 \cdot n_i^x}{\left(\sum_{i=1}^{k_x} n_i^x\right) - 1}} = \sqrt{\frac{\sum_{i=1}^{k_x} \left(x_i^{'} - \bar{x}\right)^2 \cdot n_i^x}{n - 1}} \qquad (8.9.)$$

Exemplul 8.3. S-a întreprins un studiu asupra situației materiale a consumatorilor unui sortiment de produs. În acest scop s-a recurs la un eşantion de 170 de persoane, grupat în raport cu venitul mediu lunar (tabelul 8.1.). Se cere să se determine, pe baza acestui eşantion, intervalul de încredere în care se situează, cu o probabilitate de 95 %, media aritmetică a veniturilor tuturor consumatorilor.

Tabelul 8.1. Distribuție heterogradă asociată unui eşantion

Nr. crt.	Interval de variație [RON]	Frecvență absolută $\left(n_i^x\right)$
(0)	(1)	(2)
1	[300 ; 500)	20
2	[500 ; 700)	30
3	[700 ; 900)	60
4	[900 ; 1.100)	40
5	[1.100 ; 1.300)	20

Rezolvare: În tabelul 8.2. sunt prezentate valorile intermediare utilizate în calculul abaterii medii pătratice de sondaj.

Media aritmetică a eşantionului are valoarea:

$$\bar{x}_s = \frac{\sum_{i=1}^{k_x} x_i^{'} \cdot n_i^x}{\sum_{i=1}^{k_x} n_i^x} = \frac{138.000}{170} = 811{,}76 \text{ RON}$$

Tabelul 8.2. Valori intermediare folosite în calculul abaterii medii pătratice de sondaj

Nr. crt.	Interval de variație [RON]	n_i^x	Centru de interval $x_i^{'}$ [RON]	$x_i^{'} \cdot n_i^x$ [RON]	$x_i^{'} \cdot \overline{x}_s$ [RON]	$\left(x_i^{'} \cdot \overline{x}_s\right)^2 \cdot n_i^x$ [RON^2]
(0)	(1)	(2)	(3)	(4)	(5)	(6) = (5)^2 × (2)
1	[300; 500)	20	400	8.000	– 411,76	3.390.926
2	[500 ; 700)	30	600	18.000	– 211,76	1.345.269
3	[700 ; 900)	60	800	48.000	– 11,76	8.298
4	[900 ; 1.100)	40	1.000	40.000	188,24	1.417.372
5	[1.100 ; 1.300)	20	1.200	24.000	388,24	3.014.606
6	Total	170	×	138.000	×	9.176.471
7	Simbol pentru total	$\sum_{i=1}^{k_x} n_i^x$	×	$\sum_{i=1}^{k_x} x_i^{'} \cdot n_i^x$	×	$\sum_{i=1}^{k_x} \left(x_i^{'} \cdot \overline{x}_s\right)^2 \cdot n_i^x$

Abaterea medie pătratică de sondaj reprezintă:

$$S = \sqrt{\frac{\sum_{i=1}^{k_x} \left(x_i^{'} - \overline{x}_s\right) \cdot n_i^x}{\left(\sum_{i=1}^{k_x} n_i^x\right) - 1}} = \sqrt{\frac{9.176.471}{170 - 1}} = 233 \text{ RON}$$

Nivelului de încredere α – 1=0,95 îi corespunde un coeficient de încredere z_i^{α} =0 1,96. Rezultă:

$$P\left(\overline{x} - z_i^{\alpha} \cdot \frac{S}{\sqrt{n}} \le \mu_s \le \overline{x} + z_i^{\alpha} \cdot \frac{S}{\sqrt{n}}\right) = 1 - \alpha, \text{ adică,}$$

$$P\left(811{,}76 - 1{,}96 \times \frac{233}{\sqrt{170}} \le \mu_s \le 811{,}76 + 1{,}96 \times \frac{233}{\sqrt{170}}\right) = 0{,}95$$

ceea ce înseamnă că media aritmetică a veniturilor lunare pentru toți consumatorii se află, cu o probabilitate de 95%, în intervalul [776,74 ; 846,78].

8.3.2.3. Determinarea intervalelor de încredere atunci când eşantionul are o pondere semnificativă în cadrul populației

De regulă, din considerente de eficiență, eşantioanele utilizate în sondaje au o pondere foarte mică în totalul populației studiate. Pentru aceste sondaje volumul populației statistice nu este inclus în calculele de inferență statistică întrucât este considerat infinit în raport cu volumul eşantionului.

În practică, sunt folosite uneori şi sondaje la care eşantionul are o pondere semnificativă în totalul populației. Se consideră că la aceste sondaje eşantionul are o reprezentativitate deosebită, ceea ce conduce la creşterea acurateței şi la reducerea erorii efective de inferență. Din acest motiv, în stabilirea intervalelor de încredere se obişnuieşte ca eroarea standard să fie corectată cu un aşa-numit *factor de corecție pentru populația finită*, o mărime notată cu FC_{fin} şi dată de formula:

$$FC_{fin} = \sqrt{1-\frac{n}{N}}$$

(8.10)

unde N este volumul populației studiate.

Relația de determinare a intervalului de încredere devine, în aceste condiții:

$$P\left(\overline{X}_s - z_i^{\alpha}\frac{s}{\sqrt{n}}\sqrt{1-\frac{n}{N}} \le \mu_s \le \overline{X}_s - z_i^{\alpha}\frac{s}{\sqrt{n}}\sqrt{1-\frac{n}{N}}\right) = 1-\alpha$$

(8.11)

Exemplul 8.4. În cadrul unei firme s-a efectuat un sondaj prin care s-au studiat performanțele vânzărilor unui sortiment de produs. Eşantionul folosit în acest scop a inclus 40 din cele 160 de puncte de desfacere ale firmei. S-a determinat pentru acestea o medie aritmetică a vânzărilor lunare de 1500 bucăți şi o abatere medie pătratică de sondaj de 300 bucăți. Se cere să se determine, cu o probabilitate de 99%, intervalul de încredere al mediei aritmetice a vânzărilor pentru toate punctele de desfacere.

Rezolvare: Unui nivel de încredere $\alpha - 1 = 0{,}99$ îi corespunde un coeficient de încredere $z_i^{\alpha} = 2{,}576$. rezultă:

$$P\left(\overline{X}_s - z_i^{\alpha}\frac{s}{\sqrt{n}}\sqrt{1-\frac{n}{N}} \le \mu_s \le \overline{X}_s - z_i^{\alpha}\frac{s}{\sqrt{n}}\sqrt{1-\frac{n}{N}}\right) = 1-\alpha,$$

adică

$$P\left(1500 - 2{,}576\frac{300}{\sqrt{40}}\sqrt{1-\frac{40}{160}} \le \mu_s \le 1500 - 2{,}576\frac{300}{\sqrt{40}}\sqrt{1-\frac{40}{160}}\right) = 0{,}99$$

de unde reiese că media aritmetică a vânzărilor lunare pentru toate punctele de desfacere ale firmei se află, cu o probabilitate de 99%, în intervalul [1394;1616].

8.3.3. Determinarea volumului unui eşantion

Acurateţea unui sondaj, reprezentată prin eroarea efectivă de inferenţă,depinde, aşa cum s-a văzut, de mai mulţi factori, dintre care, de regulă, cel mai uşor de controlat este volumul eşantionului. Din acest motiv, adeseori în practică se obişnuieşte ca volumul unui eşantion să fie stabilit astfel încât eroarea de sondaj să nu depăşească un nivel maxim acceptabil (se are în vedere şi faptul că cu cât volumul eşantionului este mai mare cu atât costul sondajului este mai mare iar dificultăţile de organizare sporesc).

Procedeul de determinare a volumului unui eşantion are la bază formula care exprimă dependenţa erorii limită de inferenţă faţă de volumul eşantionului. În situaţia în care nu se cunoaşte dispersia populaţiei se poate aprecia, din formula intervalului de încredere, că eroarea limită de inferenţă este dată de relaţia:

$$e_i^{\alpha} = z_i^{\alpha} \cdot \frac{s}{\sqrt{n}} \tag{8.12}$$

de unde rezultă:

$$n = \left(\frac{z_i^{\alpha} \cdot s}{e_i^{\alpha}}\right)^2 \tag{8.13}$$

Pentru un nivel maxim admisibil al erorii limită de inferenţă se poate determina, prin transformarea relaţiei (8.13) volumul minim al eşantionului:

$$n \geq \left(\frac{z_i^{\alpha} \cdot s}{e_i^{\alpha}} \right)^2 \qquad (8.14)$$

În determinarea volumului eşantionului pe baza inegalităţii (8.14) apare o dificultate dată de faptul că abaterea medie pătratică de sondaj nu poate fi calculată dacă nu se cunoaşte volumul eşantionului. În practică, această dificultate este surmontată estimându-se abaterea medie pătratică de sondaj pe baza experienţei dată de sondaje efectuate în trecut sau prin studii preliminare ale populaţiei.

Exemplul 8.5. Se cere să se determine volumul eşantionului pentru un sondaj care are ca obiect estimarea cheltuielilor lunare cu publicitatea ale corporaţiilor dintr-o ramură industrială cunoscând următoarele caracteristici ale sondajului:

- eroarea limită de inferenţă trebuie să fie de cel mult 1000 RON;
- intervalului de încredere îi este asociată o probabilitate de 95%;
- a fost estimată o abatere medie pătratică de sondaj de 10.000 RON.

Rezolvare: Probabilităţii de 95% îi corespunde un coeficient de încredere $z_i^{\alpha} = 1{,}96$

Rezultă pentru volumul eşantionului:

$$n \geq \left(\frac{z_i^{\alpha} \cdot s}{l_i^{\alpha}} \right)^2 = \left(\frac{1{,}96 \cdot 10000}{1000} \right) = 384{,}16$$

Prin rotunjire, se obţine n = 385.

8.3.4. Estimări asupra proporţiilor

Uneori, o populaţie statistică este descrisă prin proporţia unităţilor care posedă o caracteristică certă. Astfel de situaţii apar în special în cazul unor caracteristici calitative.

Se consideră că teorema limită centrală, care este formulată pentru inferenţa statistică asupra mediei aritmetice poate fi adaptată pentru proporţia unei caracteristici. Distribuţia probabilistică a acesteia poate fi astfel aproximată printr-o distribuţie normală în situaţia unui număr semnificativ de sondaje de volum mare. În aceste condiţii, formulele de calcul pentru inferenţa asupra proporţiilor sunt similare celor determinate pentru inferenţa mediei aritmetice dacă se fac următoarele înlocuiri:

- media aritmetică a eşantionului $\overline{X}_s$ este înlocuită cu proporţia caracteristicii din eşantion, notată cu p_e;
- media aritmetică a populaţiei μ_s este înlocuită cu proporţia caracteristicii în ansamblul populaţiei, notată cu p_p;
- abaterea medie pătratică de sondaj s este înlocuită cu o mărime numită *abaterea medie pătratică a proporţiilor*, notată cu s_p şi dată de relaţia:

$$s_p = \sqrt{p_e(100 - p_e)} \qquad (8.15)$$

Cu aceste echivalări, formula de determinare a unui interval de încredere asupra proporţiei devine:

$$P\left(p_e - z_i^{\alpha}\sqrt{\frac{p_e(100 - p_e)}{n}} \le p_s \le p_e - z_i^{\alpha}\sqrt{\frac{p_e(100 - p_e)}{n}} \right) = 1 - \alpha \qquad (8.16)$$

Exemplul 8.6. În cadrul unei firme s-a realizat un sondaj pentru a se estima proporţia rebuturilor la un sortiment de produs. S-a constatat că din eşantionul de 200 de bucăţi testate 16 erau defecte. Se cere să se determine, cu un nivel de semnificaţie de 12%, intervalul de încredere al proporţiei rebuturilor pentru întreaga producţie:

Rezolvare: În cadrul eşantionului, proporţia produselor defecte reprezintă:

$$p_e = \frac{\text{numar de bucati rebutate}}{\text{volumul esantionului}} \times 100 = \frac{16}{200} \times 100 = 8\%$$

Nivelului de semnificaţie de 1% (sau, altfel spus, a nivelului de încredere de 99%) îi corespunde un coeficient de încredere $z_i^{\alpha} = 2{,}576$. Rezultă:

$$P\left(p_e - z_i^{\alpha}\sqrt{\frac{p_e(100 - p_e)}{n}} \le p_s \le p_e - z_i^{\alpha}\sqrt{\frac{p_e(100 - p_e)}{n}} \right) = 1 - \alpha$$

, adică:

$$P\left(8 - 2{,}756\sqrt{\frac{8(100 - 8)}{200}} \le p_s \le 8 - 2{,}576\sqrt{\frac{8(100 - 8)}{200}} \right) = 1 - 0{,}01$$

ceea ce înseamnă că proporţia rebuturilor, pentru întreaga producţie se află, cu o probabilitate de 99%, în intervalul [3,06; 12,94].

Formula de determinare a volumului unui eşantion dată pentru inferenţa asupra mediilor aritmetice poate fi adaptată, pe baza relațiilor de echivalență menționate anterior, la inferența asupra proporțiilor astfel:

$$n \geq \left(\frac{z_i^{\alpha}}{l_i^{\alpha}} \right) \times p_e \times (100 - p_e) \qquad (8.17)$$

Întrucât proporția unităților din eşantion care posedă o anumită caracteristică nu poate fi cunoscută în momentul stabilirii volumului eşantionului, aceasta trebuie estimată pe baza unor sondaje anterioare sau a studiului preliminar al populației.

Exemplul 8.7. Se cere să se stabilească volumul eşantionului pentru un sondaj care vizează estimarea proporției facturilor incorect completate ale unei firme: Se cunosc:

- proporția facturilor incorecte, estimată preliminar pe baza sondajelor precedente, reprezintă 15%;
- eroarea limită de inferență trebuie să fie de cel mult 4%;
- intervalului de încredere asupra proporției facturilor incorecte îi este asociată o probabilitate de 95%.

Rezolvare: Nivelul de încredere de 95% îi corespunde un coeficient de încredere $z_i^{\alpha} = 1{,}96$. Rezultă:

$$n \geq \left(\frac{z_i^{\alpha}}{\ell_i^{\alpha}} \right)^2 \times p_e \times (100 - p_e) = \left(\frac{1{,}96}{4} \right) \times 15 \times (100 - 15) = 306{,}1$$

Rotunjindu-se prin majorare valoarea calculată, se obține $n = 307$.

8.4. Inferența statistică asupra sondajelor de volum redus

În comparație cu sondajele de volum mare, sondajele de volum redus sunt, de regulă, mai puțin costisitoare însă oferă o acuratețe inferioară. Acest ultim aspect face ca în principiu sondajele de volum redus să nu fie recomandate pentru cercetările statistice. Totuşi, uneori în practică pot să apară situații în care sondajele de volum redus sunt preferate celor de volum mare: atunci când nu există posibilitatea alcătuirii unui eşantion de volum mare, când sondajele de volum mare ar fi mult prea costisitoare ş.a.m.d.

Teorema limită centrală stipulează că inferența statistică poate fi descrisă de o distribuție normală chiar şi pentru sondajele de volum redus, cu condiția ca populația studiată să urmeze tot o distribuție normală. Într-un astfel de caz estimările pot fi realizate prin procedee similare celor utilizate pentru sondajele de volum mare. Din nefericire însă, cel mai adesea în practică nu sunt disponibile suficiente date pentru a se aprecia dacă populația studiată se supune unei legi de distribuție normală, ceea ce impune folosirea altor tipuri de distribuții probabilistice.

Se consideră că acuratețea inferioară pe care sondajele de volum redus o au în comparație cu sondajele de volum mare este cauzată de faptul că un eşantion de mici dimensiuni nu reflectă corespunzător dispersia populației studiate. În general, cu cât eşantionul este mai mic, cu atât sporesc şansele ca dispersia acestuia să fie mai mică în comparație cu dispersia populației.

În aceste condiții, inferența statistică a sondajelor de volum redus se poate realiza luându-se drept bază procedeele de inferență pentru sondajele de volum mare. Aceste tehnici trebuie însă adaptate pentru a se lua în considerare faptul că eşantioanele de volum redus reflectă într-o măsură mai mică dispersia populației studiate. În acest scop se foloseşte un tip de distribuție probabilistică, numit *distribuția t*, asemănător cu o distribuție normală (are un grafic simetric, în formă de clopot, însă mai aplatizat decât cel specific unei distribuții normale) dar care face ca pentru o aceeaşi abatere medie pătratică de sondaj şi acelaşi nivel de încredere să corespundă o eroare limită de inferență mai mare (fig. 8.2.).

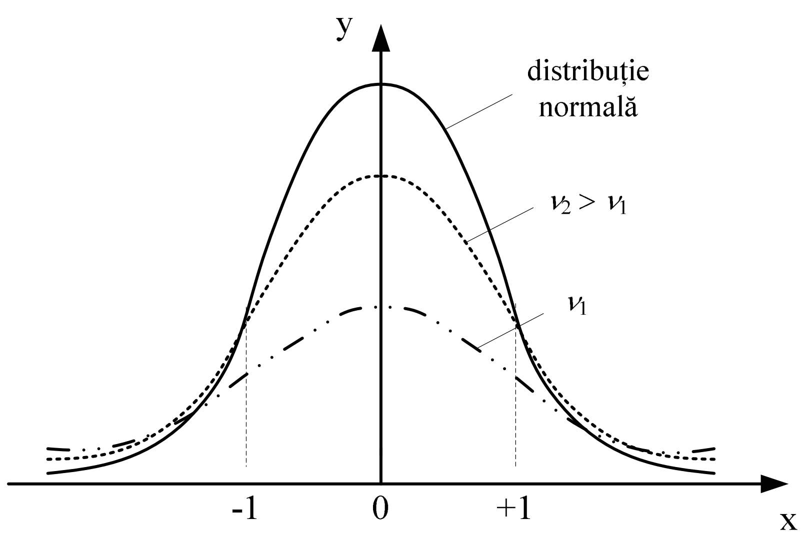


Fig. 8.2. Distribuția normală şi distribuții de tip t

În fapt, există mai multe forme ale distribuției în raport cu reflectarea în cadrul eşantionului a dispersiei populației studiate. Drept criteriu de diferențiere poate fi folosit un indicator numit *număr de grade de libertate*, notat cu ν, care este dat de numărul de unități statistice independente folosite pentru estimarea unui parametru. În cazul estimării dispersiei pe baza sondajelor de volum redus se consideră că numărul de grade de libertate poate fi obținut scăzând o unitate din volumul eşantionului. Justificarea vine din faptul că indicatorii dispersiei folosiți în estimare sunt calculați pe baza abaterilor față de media aritmetică. Întrucât suma algebrică a acestora este întotdeauna nulă rezultă că valoarea abaterii unei unități față de medie poate fi dedusă din celelalte. Se poate deci concluziona că numărul de grade de libertate asociat estimării dispersiei pe baza unui eşantion de volum redus este dat de relația:

$$\nu = n - 1 \qquad (8.14)$$

unde n este volumul eşantionului.

Cu cât numărul de grade de libertate este mai mare, cu atât dispersia populației este reflectată mai semnificativ în cadrul eşantionului iar distribuția *t* este mai apropiată de distribuția normală (fig. 8.2).

Proprietățile distribuțiilor de tip t facilitează unele calcule probabilistice. Poate fi astfel cunoscută probabilitatea, reprezentată prin suprafața haşurată din figura 8.3., ca valorile distribuției să depăşească un anumit punct critic. În funcție de numărul de grade de libertate şi proporția, notată cu *q*, pe care suprafața o are în totalul ariei delimitate de graficul distribuției, se pot stabili valori tabelate, notate cu t_q^{ν}, care exprimă poziția punctului critic (tabelul 8.3.).

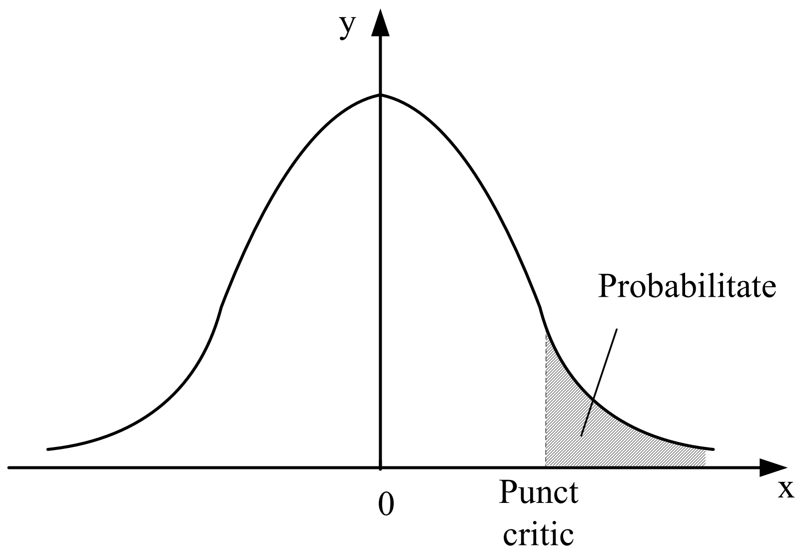


Fig. 8.3. Reprezentarea probabilității ca valorile unei distribuții de tip t să depășească un punct critic

Tabelul 8.3. Valori tabelate pentru mărimea t_q^{ν}

Probabilitate / **Număr de grade de libertate**	**0,10**	**0,05**	**0,025**	**0,01**	**0,005**
4	1,533	2,132	2,776	3,747	4,604
5	1,476	2,015	2,571	3,365	4,032
6	1,440	1,943	2,447	3,143	3,707
23	1,319	1,714	2,069	2,500	2,807
24	1,318	1,711	2,064	2,492	2,797

Cu un raționament similar celui folosit în cazul inferenței asupra sondajelor de volum mare, se poate determina formula de stabilire a intervalelor de încredere pentru sondajele de volum redus:

$$P\left(\overline{X}_s - t_q^{\nu} \cdot \frac{s}{\sqrt{n}} \le \mu_s \le \overline{X}_s + t_q^{\nu} \cdot \frac{s}{\sqrt{n}}\right) = 1 - \alpha \qquad (8.19)$$

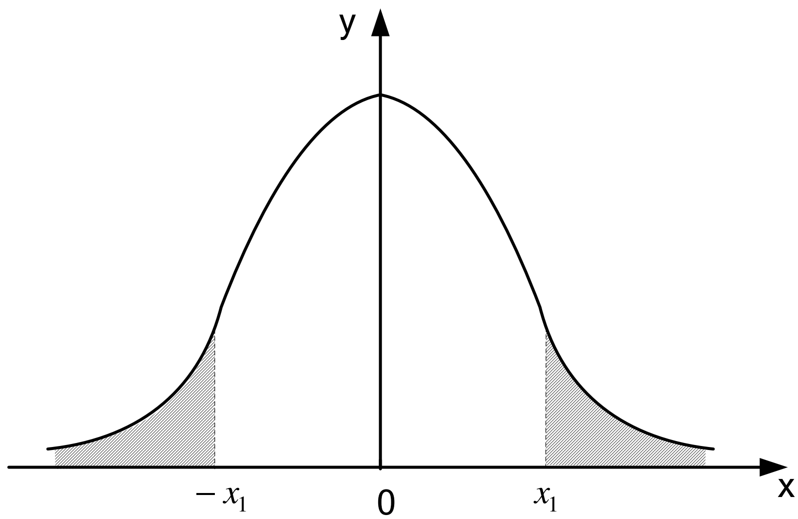


Figura 8.4. Probabilități asociate unui interval de încredere

Proporția q se stabileşte luându-se în considerare faptul că într-o distribuție t standardizată intervalul de încredere este dispus simetric în raport cu valoarea nulă a mediei aritmetice (figura 8.4.) astfel încât probabilitatea ca valorile distribuției să nu fie cuprinse într-un interval $[-x_1,\ x_1]$ reprezintă de fapt dublul probabilității ca valorile distribuției să fie mari decât valoarea x_1. Altfel spus:

$$q = \frac{\alpha}{2} \tag{8.20}$$

Exemplul 8.7. : Managerii unei firme de transport şi-au propus să estimeze costul mediu anual al întreținerii unui autocamion. În acest scop au fost selectate cinci maşini pentru care au fost obținute datele prezentate în tabelul 8.4. Se cere să se determine, cu o probabilitate de 95%, intervalul de încredere al costului mediu anual de întreținere pentru ansamblul parcului de autocamioane al firmei.

Tabelul 8.4. Costuri anuale de întreținere pentru un grup de cinci autocamioane

Nr. crt.	Cost anual de întreținere (x_i)[RON]
1	60
2	80
3	80
4	70
5	70

Rezolvare: În tabelul 8.5. sunt prezentate datele intermediare folosite în calculul mediei aritmetice şi a abaterii medii pătratice de sondaj.

Tabelul 8.5. Date intermediare utilizate în calculul mediei aritmetice şi a abaterii medii pătratice de sondaj

Nr. crt.	X_i [RON]	$X_i - \overline{X}$ [RON]	$\left(X_i - \overline{X}\right)^2$ [RON^2]
(0)	(1)	(2)	(3) = (2)^2
1	60	-12	144
2	80	8	64
3	80	8	64
4	70	2	4
5	70	2	4
Total	360	×	280
Simbol pentru total	$\sum_{1=i}^{n} X_i$	×	$\sum_{1=i}^{n}\left(X_i - \overline{X}\right)$

Media aritmetică a eşantionului are valoarea:

$$\overline{X}_s = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{360}{5} = 72 \text{ RON}$$

Abaterea medie pătratică de sondaj reprezintă:

$$S = \sqrt{\frac{\sum_{n-1}^{n}\left(X_i - \overline{X}\right)}{n-1}} = \sqrt{\frac{280}{5,1}} = 8,37 \text{ RON}$$

Numărul de grade de libertate pentru acest sondaj are valoarea $\nu = n - 1 = 5 - 1 = 4$

În raport cu nivelul de semnificație $\alpha = 0,05$ se determină: $q = \frac{\alpha}{2} = \frac{0,05}{2} = 0,025$

Din tabelul 8.3. se extrage pentru $\nu = 4$ şi q = 0,025 o valoare $t_q^\nu = 2,776$

Rezultă: $P\left(\overline{X}_s - t_q^\nu \cdot \frac{s}{\sqrt{n}} \le \mu_s \le \overline{X}_s + t_q^\nu \cdot \frac{s}{\sqrt{n}}\right) = 1 - \alpha$, adică

$$P\left(72-2{,}776\cdot\frac{8{,}37}{\sqrt{5}}\leq\mu_s\leq72+2{,}776\cdot\frac{8{,}37}{\sqrt{5}}\right)=0{,}95$$

ceea ce înseamnă că media aritmetică pentru ansamblul populației studiate se situează, cu o probabilitate de 95%, în intervalul [61,6 ; 82,4].

8.5. Verificarea ipotezelor statistice prin sondaje

Uneori, sondajele statistice sunt utilizate pentru a verifica anumite aprecieri preliminare asupra populației studiate. În acest scop sunt formulate două ipoteze:

1) o ipoteză care îmbracă forma aprecierii inițiale, numită *ipoteza nulă* şi notată cu H_0;
2) o ipoteză care reprezintă opusul aprecierii inițiale, numită *ipoteza alternativă* şi notată cu H_A.

În condițiile în care caracteristicile populației studiate prin sondaj nu pot fi cunoscute cu certitudine, confirmarea sau infirmarea ipotezei nule trebuie să se facă în termeni probabilistici, pe baza unui nivel de semnificație. În acest scop pot fi folosite unele proprietăți ale tipurilor de distribuții probabilistice specifice tipurilor de sondaje folosite.

În practică, pentru verificarea procedeelor statistice sunt utilizate diverse procedee. În acest subcapitol, vom prezenta un algoritm de verificare, prin sondaje de volum mare, a ipotezelor asupra mediei aritmetice a unei populații. După cum se ştie, pentru sondajele de volum mare probabilitățile pot fi stabilite prin intermediul unei distribuții normale standard, în care suprafața ce reprezintă nivelul de semnificație poate fi împărțită în două arii dispuse simetric (fig. 8.6.).

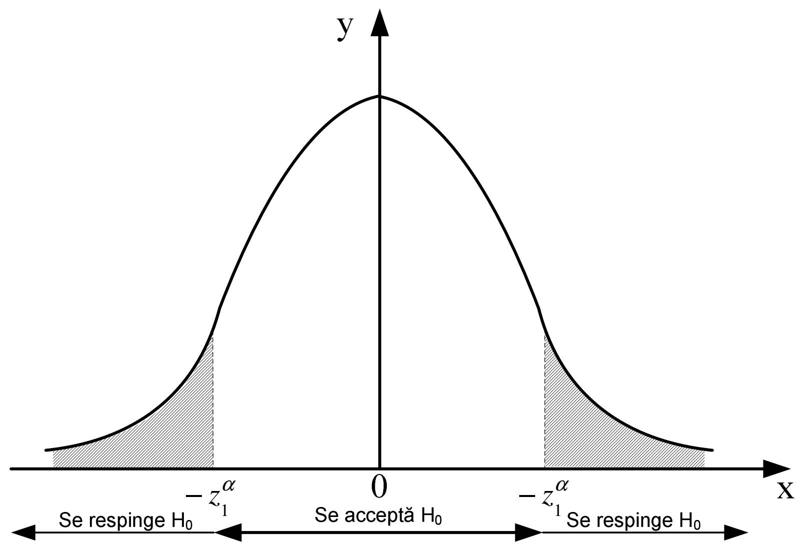


Fig. 8.6. Verificarea unei ipoteze statistice printr-un sondaj de volum mare

Domeniul din graficul distribuției asociat acceptării ipotezei nule are limitele date de coordonatele $-z_1^{\alpha}$ şi $+z_1^{\alpha}$, suprafața sa reprezentând astfel nivelul de încredere $1-\alpha$.

Algoritmul de verificare a ipotezei asupra mediei aritmetice cuprinde mai multe etape:

Pasul 1 - Se stabilesc cele două ipoteze:

- ipoteza nulă, $H_0 : \mu_s = \mu_0$, unde μ_0 este valoarea atribuită inițial mediei aritmetice a populației;
- ipoteza alternativă, $H_A : \mu_s \neq \mu_0$.

Pasul 2 – Se stabileşte un nivel de semnificație acceptabil pentru verificarea ipotezei nule.

Pasul 3 - Se determină, în funcție de nivelul de semnificație, valoarea tabelată z_i^{α}, numită, în acest caz, *valoare critică*.

Pasul 4 - Se determină media aritmetică a eşantionului $\bar{x}_s$ şi abaterea medie pătratică de sondaj *s*.

Pasul 5 - Se calculează o mărime numită *valoarea testului statistic z*, prin relația:

$$z = \frac{\bar{x}_s - \mu_0}{s}\sqrt{n} \qquad (8.20)$$

Pasul 6 – Se compară valoarea testului statistic *z* cu valorile $-z_i^{\alpha}$ şi $+z_i^{\alpha}$ rezultând una din următoarele concluzii:

- dacă z aparține intervalului $\left[-z_i^{\alpha};+z_i^{\alpha}\right]$ se acceptă ipoteza nulă;
- dacă z nu aparține intervalului $\left[-z_i^{\alpha};+z_i^{\alpha}\right]$ se respinge ipoteza nulă.

Exemplul 8.8. Managerii departamentului de marketing al unei firme ce produce dulciuri consideră că în medie un client cheltuieşte săptămânal pentru ciocolată suma de 7 RON. Pentru a se verifica această ipoteză a fost efectuat un sondaj pe un eşantion de 40 de clienți, determinându-se o medie aritmetică de 6,8 RON şi o abatere medie pătratică de sondaj de 0,8 RON. Se cere să se identifice concluziile verificării pentru un nivel de semnificație de 5%.

Rezolvare: Ipoteza nulă este dată de aprecierea inițială a managerilor, adică:

$H_0 : \mu_s = \mu_0 = 7$ RON

Ipoteza alternativă este reprezentată de opusul ipotezei nule:

$H_A : \mu_s \neq 7$ RON

Petru un nivel de semnificație α = 5% (sau un nivel de încredere de 95%) se stabileşte o valoare tabelată z_i^{α} = 1,96. Valoarea testului statistic z reprezintă: $z = \frac{\bar{x}_s - \mu_0}{s}\sqrt{n} = \frac{6,8-7}{0,8} \times \sqrt{40} = -1,581$

Deoarece -1,581 aparține intervalului [– 1,96 ; + 1,96] se poate considera că, pentru un nivel de semnificație de 5 %, ipoteza nulă poate fi acceptată.

Capitolul 9 - Analiza statistică a legăturilor dintre variabile

9.1. Coordonate ale analizei statistice a legăturilor dintre variabile

În cadrul cercetărilor statistice, termenul de variabilă desemnează o colecție de date organizate în raport cu o caracteristică a populației studiate. Pentru a fi completă, o cercetare statistică presupune atât studiul separat al fiecărei variabile (altfel spus, al fiecărui aspect esențial al fenomenului cercetat) cât şi o abordare a legăturilor semnificative care există între variabile. Aceste legături pot fi transpuse în relații de tip cauză-efect, folosite în elaborarea modelelor care descriu mecanismele fenomenelor cercetate.

În cadrul modelării fenomenelor colective sunt utilizate două tipuri de variabile:

- *variabile independente*, care descriu factorii de influență asupra fenomenelor modelate;
- *variabile dependente*, care descriu efectele acțiunii factorilor de influență.

Într-o cercetare statistică, analiza legăturilor dintre variabile, numită şi *analiză a corelației* vizează mai multe aspecte:

a) identificarea legăturilor relevante dintre variabile;

b) stabilirea formelor sub care se manifestă aceste legături;

c) evaluarea intensității legăturilor dintre variabile.

a) **Identificarea legăturilor relevante dintre variabile** se bazează pe studiul evoluției în paralel a unei variabile dependente şi a uneia sau mai multor variabile independente. Atunci când schimbările unei variabile sunt însoțite de modificări importante ale altei variabile se poate emite ipoteza unei legături relevante. Trebuie avut însă în vedere faptul că modificările concomitente a două variabile nu sunt neapărat rezultatul unei legături semnificative. Uneori, simultan cu factorii de influență studiați se produce şi acțiunea altor factori, pe care nu i-am luat în considerare dar care au un impact determinant asupra fenomenului cercetat. Coincidența acțiunii ne face să atribuim toate efectele factorilor pe care i-am considerat relevanți când, de fapt, acestea s-au datorat în primul rând factorilor pe care i-am neglijat. Un alt aspect care poate spori complexitatea identificării factorilor legăturilor dintre variabile este dat de faptul că influența unor factori asupra fenomenelor cercetate se produce cu întârziere.

b) **Formele stabilite pentru legăturile dintre variabile** sunt deosebit de importante din perspectiva aplicării modelelor ce descriu mecanismele fenomenelor cercetate. Se recomandă ca legăturii dintre o variabilă dependentă şi una sau mai multe variabile independente să îi fie atribuită forma unei funcții matematice ale cărei parametri să poată fi determinați. Funcțiile matematice folosite în acest scop pot fi clasificate în raport cu două criterii:

b_1) numărul de variabile independente;

b_2) tipul ecuației matematice.

b_1) În raport cu **numărul de variabile independente**, funcțiile matematice utilizate pot fi împărțite în două categorii:

- funcții cu o singură variabilă independentă, de forma $y = f(x)$;
- funcții cu mai multe variabile independente, de forma $y = f(x_1, x_2, \dots, x_n)$.

În practică, funcțiile cu mai multe variabile independente, cu toate că pot conferi o rigoare deosebită cercetării, sunt adeseori evitate ca urmare a complexității deosebite pe care o induc analizei statistice. În schimb, funcțiile cu o singură variabilă independentă, sunt folosite, datorită simplității pe care o conferă, chiar şi atunci când nu aduc o rigoare prea mare modelelor.

b_2) În raport cu **tipul ecuației matematice** se pot delimita două categorii ale funcțiilor folosite în analiza legăturilor dintre variabile:

- funcții liniare, date de o ecuație liniară;
- funcții neliniare, cu o ecuație matematică mai complexă: parabolice, hiperbolice, logaritmice, exponențiale etc.

Din aceleaşi considerente de simplitate, în practică, funcțiile liniare sunt folosite mult mai frecvent decât funcțiile neliniare.

Un aspect important în cazul legăturilor cu o singură variabilă independentă este reprezentat de corespondența dintre direcțiile în care evoluează variabila dependentă şi cea independentă. Din această perspectivă se pot delimita două tipuri de legături între variabile:

- *legături directe*, în care cele două variabile evoluează în acelaşi sens;
- *legături inverse*, în care variabilele evoluează în sensuri opuse.

c) **Evaluarea intensității legăturilor dintre variabile** are rolul de apreciere a impactului pe care factorii de influență reprezentaţi prin variabilele independente îl au asupra aspectului reprezentat printr-o variabilă dependentă. Cu cât legătura este mai intensă cu atât influența acestor factori este mai determinantă. Evaluarea intensității legăturilor dintre variabile oferă, totodată, un indiciu asupra impactului unor factori care nu au fost reprezentați prin variabile independente, ceea ce permite aprecierea reprezentativității relațiilor de tip cauză efect.

9.2. Tehnici grafice de caracterizare a legăturilor dintre variabile

Tehnicile grafice de caracterizare a legăturilor dintre variabile, numite şi *corelograme*, sunt simplu de aplicat şi pot oferi indicii asupra unor aspecte importante ale legăturilor dintre variabile. În general, tehnicile grafice se bazează pe reprezentarea valorilor variabilelor în sisteme de coordonate carteziene. Din perspectiva seriilor statistice prin care sunt descrise variabilele se diferențiază două tipuri de corelograme:

- *corelograme pentru seriile simple*, care constau în reprezentări prin puncte ce au drept coordonate valorile variabilelor;
- *corelograme pentru distribuții heterograde*, care presupun reprezentarea subgrupelor prin dreptunghiuri ce corespund intervalelor de variație, în interiorul fiecărui dreptunghi fiind trasat un număr de puncte egal cu frecvența absolută a grupei (atunci când frecvențele absolute sunt foarte mari, în locul punctelor se pot trasa figuri geometrice cu suprafețele proporționale cu frecvențele).

Pe baza corelogramelor se pot face aprecieri asupra unor caracteristici ale legăturilor dintre variabile:

a) forma funcției matematice adecvată pentru exprimarea unei legături;

b) sensul legăturii dintre variabile;

c) intensitatea legăturii dintre variabile.

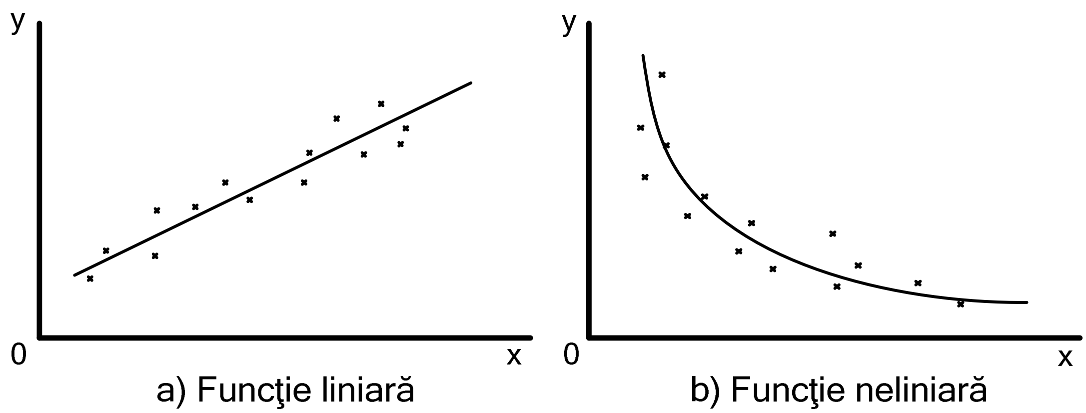


Fig. 9.1. Alegerea, pe cale grafică, a funcției matematice asociată legăturii dintre variabile

a) **Forma funcției matematice** utilizată în exprimarea unei legături poate fi aleasă prin tehnici grafice, folosindu-se condiția ca graficul funcției să fie cât mai apropiat de reprezentările valorilor variabilei. Chiar dacă nu pot conduce neapărat la determinarea parametrilor funcției, corelogramele facilitează, cel puțin, alegerea între o funcție liniară şi una neliniară (fig. 9.1.)

b) **Sensul legăturii dintre două variabile** poate fi apreciat destul de facil prin intermediul corelogramelor, care relevă creşterea sau descreşterea variabilei dependente odată cu creşterea variabilei independente. În figura 9.2. sunt prezentate reprezentările grafice ale două tipuri de legături: o legătură directă, la care creşterii variabilei independente îi corespunde o creştere a variabilei dependente (fig. 9.2.a) şi o legătură inversă, pentru care creşterea variabilei independente determină scăderea variabilei dependente (fig. 9.2.b).

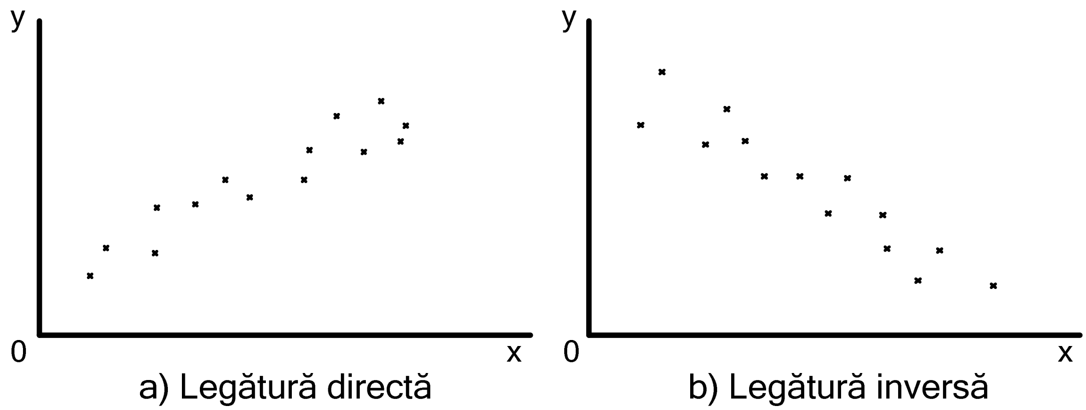


Fig. 9.2. Aprecierea, pe cale grafică, a sensului legăturilor dintre variabile

c) **Intensitatea legăturii dintre variabile** poate fi apreciată, pe cale grafică, pe baza concentrării punctelor ce reprezintă valorile variabilelor şi a apropierii acestora de graficul funcţiei ce exprimă legătura dintre variabile. În figura 9.3. sunt prezentate două legături între variabile: o legătură de intensitate maximă (numită şi *legătură deterministă*) în care punctele se găsesc pe graficul funcţiei asociate legăturii (fig. 9.3.a) şi o legătură de intensitate foarte slabă, în care punctele nu pot fi asociate unei funcţii (fig. 9.3.b).

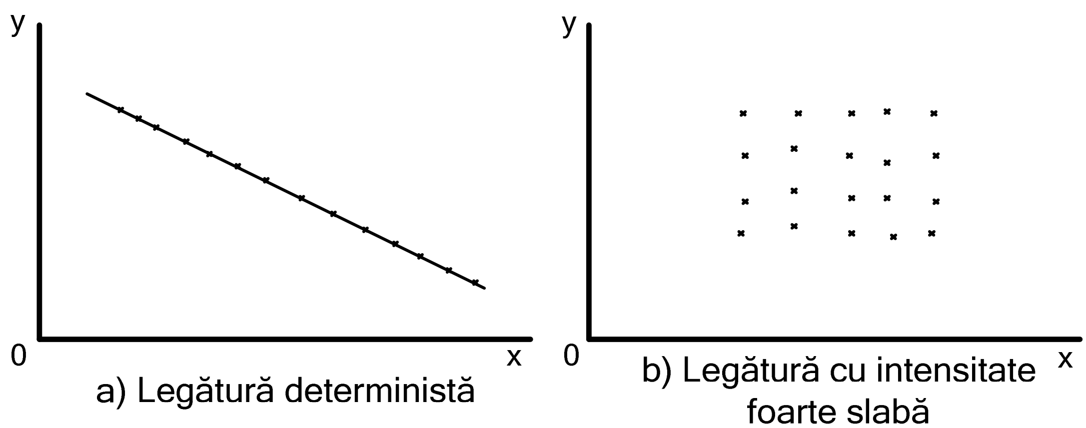


Fig. 9.1. Aprecierea, pe cale grafică, a intensităţii legăturii dintre două variabile

Exemplul 9.1. În tabelul 9.1. sunt prezentate rezultatele unei cercetări asupra legăturii dintre cheltuielile anuale pentru publicitate şi volumul desfacerilor, întreprinsă asupra unui eşantion de cinci firme dintr-o ramură industrială. Se cere să se caracterizeze, pe baza reprezentării grafice, legătura dintre cele două variabile.

Rezolvare: În figura 9.4. este prezentată corelograma asociată legăturii dintre cele două variabile. Pe baza acesteia se poate alege pentru exprimarea legăturii o funcţie liniară. De asemenea, se poate aprecia că legătura dintre cele două variabile este directă (cu o singură excepţie, cu cât cheltuielile pentru publicitate sunt mai mari, cu atât volumul desfacerilor este mai mare) şi de intensitate semnificativă (punctele ce reprezintă valorile variabilelor sunt destul de apropiate de graficul funcţiei liniare).

Tabelul 9.1. Valorile cheltuielilor pentru publicitate şi ale volumului desfacerilor pentru un grup de cinci firme

Nr. crt.	Cheltuieli pentru publicitate [mil. RON]	Volumul desfacerilor [mii buc.]
(0)	(1)	(2)
1	0,2	0,5
2	0,8	1,1
3	0,5	0,7
4	0,6	0,9
5	0,4	0,8

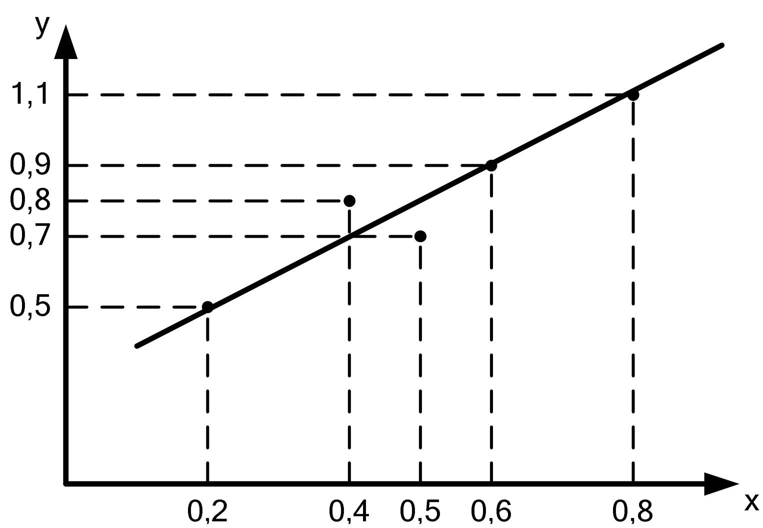


Fig. 9.4. Reprezentarea grafică a relaţiei dintre cheltuielile pentru publicitate şi volumul desfacerilor

Pentru distribuțiile heterograde, caracterizarea relațiilor dintre variabile pe cale grafică este ceva mai dificilă față de seriile simple întrucât, în acest caz punctele au mai degrabă semnificația unor frecvențe decât a unor valori. În consecință, funcția matematică se alege astfel încât graficul ei să fie cât mai apropiat de dreptunghiurile cu concentrații mari de puncte. Sensul şi intensitatea legăturii sunt apreciate, de asemenea, pe baza concentrațiilor de puncte din dreptunghiuri.

Exemplul 9.2. În tabelul 9.2. este prezentată o distribuție heterogradă care descrie vechimea în muncă şi numărul mediu zilnic de rebuturi, pentru un grup de 25 de angajați ai unei firme. Se cere să se caracterizeze pe cale grafică legătura dintre cele două variabile.

Tabelul 9.2. Distribuție heterogradă asupra vechimii în muncă şi numărul de rebuturi

Vechimea în muncă [ani] / Număr mediu zilnic de rebuturi [buc]	(0 ; 4]	(4 ; 8]	(8 ; 12]	(12 ; 16]	(16 ; 20]
(1,0 ; 1,1]	–	–	–	1	3
(1,1 ; 1,2]	1	1	1	2	–
(1,2 ; 1,3]	2	1	5	2	–
(1,3 ; 1,4]	2	1	1	–	–
(1,4 ; 1,5]	2	–	–	–	–

Rezolvare: În figura 9.5. este prezentată corelograma asociată distribuției heterograde. Relația dintre cele două variabile poate fi descrisă atât printr-o funcție liniară cât şi printr-o funcție neliniară. Din considerente de simplitate s-a optat pentru o funcție liniară al cărei grafic să fie apropiat de dreptunghiurile cu concentrații maxime. Din aceeaşi figură se poate deduce că legătura dintre cele două variabile este inversă (cu cât vechimea este mai mare cu atât numărul rebuturilor este mai mic) iar unele valori sunt destul de îndepărtate de grafic, ceea ce înseamnă că legătura nu este foarte intensă.

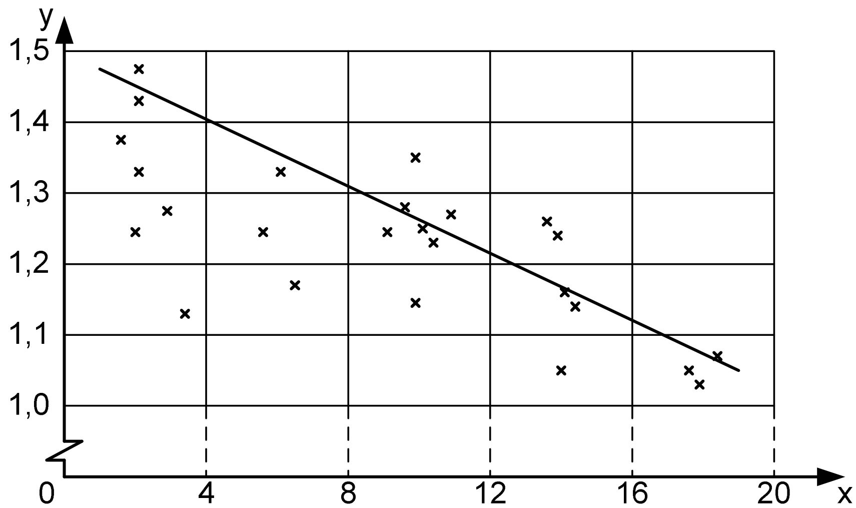


Fig. 9.5. Corelograma distribuției heterograde

Pe lângă avantajul simplității în aplicare, tehnicile grafice de caracterizare a legăturilor dintre variabile au şi dezavantajul unei rigori reduse, în condițiile în care nu pot conduce la cuantificarea aspectelor esențiale ale relațiilor. În plus folosirea lor este limitată, practic, la legăturile cu o singură variabilă independentă, pentru relațiile cu mai multe variabile independente aplicarea fiind foarte complexă.

9.3 Analiza legăturilor dintre variabile prin intermediul regresiei

9.3.1. Conceptul de regresie

Termenul de *regresie* are semnificația de studiu al legăturilor dintre variabile prin intermediul unor funcții matematice numite *funcții de regresie*. Valorile acestora, numite *valori teoretice* sunt aproximări ale valorilor variabilelor dependente, care sunt numite *valori empirice*. Se consideră că o valoare teoretică este rezultatul exclusiv al factorilor de influență exprimați prin variabilele independente în timp ce o valoare empirică este rezultatul tuturor factorilor de influență care acționează, la momentul înregistrării, asupra fenomenului studiat. Această situație se transpune într-o formă matematică astfel:

$$y_i = f(x_i) + \varepsilon_t = y_{x_i} + \varepsilon_t \qquad (9.1.)$$

unde:

- y_i este valoarea empirică a variabilei independente y obţinută în condiţiile i;
- f este funcţia de regresie asociată legăturii dintre variabila dependentă y şi variabila independentă (sau variabilele independente exprimate vectorial) x;
- x_i este o valoare numerică ce exprimă manifestarea în condiţiile i a factorilor de influenţă reprezentaţi prin variabila independentă (sau variabilele independente);
- ε_t este un termen numit *variabilă reziduală*, care exprimă efectele pe care le au asupra variabile dependente factorii de influenţă care nu au fost exprimaţi prin variabilele independente;
- este valoarea teoretică a variabilei dependente în condiţiile i, care se obţine atribuind argumentului funcţiei de regresie valoarea x_i (altfel spus, $y_{x_i} = f_{(xi)}$).

Parametrii unei funcţii de regresie pot rezulta din condiţia ca pentru ansamblul observărilor statistice, care dau circumstanţele de manifestare a fenomenului studiat, diferenţele dintre valorile teoretice şi cele empirice să fie cât mai mici (fig. 9.6.).

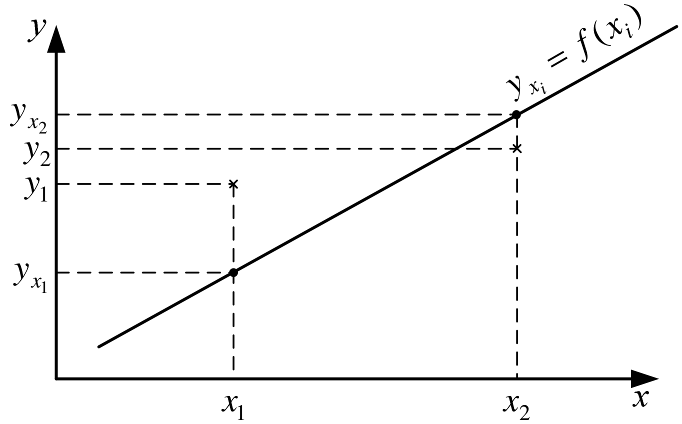


Fig. 9.6. Reprezentarea grafică a valorilor empirice şi a valorilor teoretice

Această condiţie poate fi transpusă într-o expresie matematică în mai multe moduri:
- minimizând suma valorilor absolute ale diferenţelor dintre valorile teoretice şi cele empirice (se folosesc valorile absolute pentru ca diferenţele pozitive să nu le anuleze pe cele negative);
- minimizând suma pătratelor diferenţelor dintre valorile teoretice şi cele empirice (prin ridicare la pătrat toţi termenii sumei devin

pozitivi ceea ce înlătură posibilitatea anulării reciproce a valorilor pozitive şi a celor negative).

În practică, se preferă de regulă a doua modalitate, care îmbracă forma unui procedeu numit *metoda celor mai mici pătrate* şi are la bază minimizarea funcției:

$$S(a_0, a_1, \ldots, a_n) = \sum_{i=1}^{N} [f(x_i) - y_i]^2 = \sum_{i=1}^{N} (y_{x_i} - y_i)^2 \qquad (9.2.)$$

unde a_0, a_1, …, a_n sunt parametrii funcției de regresie $f(x_i)$ care constituie argumente pentru funcția S. Funcția S fiind o funcție de mai multe variabile, minimizarea sa poate fi realizată pe baza ecuațiilor lui Fermat:

$$\begin{cases} \dfrac{\partial S}{\partial a_0} = 0 \\ \dfrac{\partial S}{\partial a_1} = 0 \\ \ldots\ldots\ldots\ldots \\ \dfrac{\partial S}{\partial a_n} = 0 \end{cases} \qquad (9.3.)$$

care conduc, în final, la rezolvarea unui sistem cu n ecuații.

Funcțiile de regresie au aplicații importante în practică. Pe baza acestora se pot face previziuni asupra efectelor posibile ale acțiunii unor factori de influență, atribuind diferite valori variabilelor independente şi calculând valorile teoretice ale variabilelor dependente. În raport cu valorile variabilelor independente folosite, se pot delimita două forme ale previziunii pe baza funcțiilor de regresie:

- **interpolarea**, când valorile variabilelor independente se află în interiorul intervalului de valori obținut prin observări statistice;
- **extrapolarea**, când valorile variabilelor independente se află în afara intervalului de valori obținut prin observații statistice.

Se consideră că în general acuratețea previziunilor prin interpolare este superioară acurateței previziunilor prin extrapolare întrucât pentru valorile variabilelor independente din afara intervalului obținut prin observații statistice fenomenul ar putea urma alte mecanisme decât cele descrise prin funcția de regresie.

În practică sunt folosite diferite forme ale regresiei, pentru a căror clasificare pot fi utilizate mai multe criterii:

a) numărul de variabile independente;

b) ecuația funcției de regresie;

c) forma seriei statistice care descrie variabilele utilizate.

a) În funcție de **numărul de variabile independente**, regresiile pot fi împărțite în două categorii:

a_1) **regresii unifactoriale**, la care se utilizează o singură variabilă independentă;

a_2) **regresii multifactoriale**, la care se utilizează mai multe variabile independente.

b) În raport cu **ecuația funcției de regresie**, se pot delimita două forme de regresie:

b_1) **regresii liniare**, la care se folosesc funcții cu ecuații liniare;

b_2) **regresii neliniare**, la care se folosesc funcții cu ecuații neliniare.

c) În funcție de **forma seriei statistice care descrie variabilele**, regresiile pot fi grupate în două categorii:

c_1) **regresii pentru serii simple**;

c_2) **regresii pentru distribuții heterograde**.

9.3.2. Regresii unifactoriale

Metodele regresiilor unifactoriale se diferențiază în raport cu forma ecuațiilor funcțiilor de regresie:

- metode pentru regresii unifactoriale liniare;
- metode pentru regresii unifactoriale neliniare.

9.3.2.1. Regresii unifactoriale liniare

Procedeele regresiei unifactoriale liniare prezintă anumite particularități în funcție de forma seriei statistice care descrie variabilele, ceea ce justifică împărțirea în două categorii:

- procedee ale regresiei unifactoriale liniare pentru serii simple;
- procedee ale regresiei unifactoriale liniare ale distribuțiilor heterograde.

9.3.2.1.1. Regresii unifactoriale liniare pentru seriile simple

Regresia unifactorială liniară pentru seriile simple are la bază adaptarea formulelor metodei celor mai mici pătrate pentru o funcție liniară cu un singur argument:

$$y_{x_i} = a + bx_i \tag{9.4.}$$

În acest caz, funcția care exprimă suma pătratelor diferențelor dintre valorile teoretice şi valorile empirice îmbracă forma:

$$S_{(a,b)} = \sum_{i=1}^{N} \left(y_{x_i} - y_i\right)^2 = \sum_{i=1}^{N} \left(a + b_{x_i} - y_i\right)^2 \tag{9.5}$$

Determinarea valorilor parametrilor a şi b pentru care funcția *S* are un minim presupune rezolvarea ecuațiilor lui Fermat:

$$\begin{cases} \dfrac{\partial S}{\partial a} = 0 \\ \dfrac{\partial S}{\partial b} = 0 \end{cases} \tag{9.6}$$

Derivata parțială a funcției *S* în raport cu argumentul *a* are expresia:

$$\frac{\partial S}{\partial a} = \sum_{i=1}^{N} \frac{\partial\left[\left(a + bx_i - y_i\right)\right]}{\partial a} = \sum_{i=1}^{N} \left[2 \frac{\partial\left(a + bx_i - y_i\right)}{\partial a} \cdot \left(a + bx_i - y_i\right)\right] =$$

$$= \sum_{i=1}^{N} \left[2 \times 1 \times \left(a + bx_i - y_i\right)\right] = 2\left(N \cdot a + b\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} y_i\right) \tag{9.7.}$$

În raport cu argumentul *b*, derivata parțială a funcției *S* are expresia:

$$\frac{\partial S}{\partial b} = \sum_{i=1}^{N} \frac{\partial\left[\left(a + bx_i - y_i\right)\right]}{\partial b} = \sum_{i=1}^{N} \left[2 \frac{\partial\left(a + bx_i - y_i\right)}{\partial b} \cdot \left(a + bx_i - y_i\right)\right] =$$

$$= \sum_{i=1}^{N} \left[2 \times X_i \times \left(a + bx_i - y_i\right)\right] = 2\left(a\sum_{i=1}^{N} x_i + b\sum_{i=1}^{N} x_i^2 - \sum_{i=1}^{N} x_i y_i\right)$$

Introducând expresiile derivatelor parțiale în ecuațiile lui Fermat obținem:

$$\begin{cases} \dfrac{\partial S}{\partial a} = \sum\limits_{i=1}^{N} 2\left(N \cdot a + b\sum\limits_{i=1}^{N} x_i - \sum\limits_{i=1}^{N} y_i\right) = 0 \\ \dfrac{\partial S}{\partial a} = \sum\limits_{i=1}^{N} 2\left(a\sum\limits_{i=1}^{N} x_i + b\sum\limits_{i=1}^{N} x_i^2 - \sum\limits_{i=1}^{N} x_i y_i\right) = 0 \end{cases} \tag{9.9}$$

Rezultă un sistem de ecuații prin care pot fi determinați parametrii a şi b ai funcției de regresie:

$$\begin{cases} N \cdot a + b\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i \\ a\sum_{i=1}^{N} x_i + b\sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} x_i y_i \end{cases} \tag{9.10}$$

Exemplul 9.3. Se cere să se determine parametrii unei funcții de regresie care să exprime dependența volumului desfacerilor față de cheltuielile pentru publicitate pe baza seriei simple prezentate în tabelul 9.1.

Rezolvare: În această aplicație vom nota cu x_i cheltuielile pentru publicitate şi cu y_i volumul desfacerilor.

Tabelul 9.3. Valori intermediare utilizate în calculul parametrilor funcției de regresie pentru o serie statistică simplă

Nr. crt.	x_i [mil. RON]	y_i [mii buc.]	x_i^2 [(mil. RON)^2]	y_i^2 [(mii buc.)^2]	$x_i \cdot y_i$ [mil. RON × mii buc.]
(0)	(1)	(2)	(3)	(4)	(5)
1	0,2	0,5	0,04	0,25	0,10
2	0,8	1,1	0,64	1,21	0,88
3	0,5	0,7	0,25	0,49	0,35
4	0,6	0,9	0,36	0,81	0,54
5	0,4	0,8	0,16	0,64	0,32
Total	2,5	4,0	1,45	3,40	2,19
Simbol pentru total	$\sum_{i=1}^{N} x_i$	$\sum_{i=1}^{N} y_i$	$\sum_{i=1}^{N} x_i^2$	$\sum_{i=1}^{N} y_i^2$	$\sum_{i=1}^{N} x_i y_i$

În tabelul 9.3. sunt prezentate valorile intermediare pe baza cărora poate fi constituit sistemul de ecuații prin care pot fi determinate valorile parametrilor funcției liniare:

$$\begin{cases} N \cdot a + b\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} y_i \\ a\sum_{i=1}^{N} x_i + b\sum_{i=1}^{N} x_i^2 - \sum_{i=1}^{N} x_i y_i \end{cases}$$

adică:

$$\begin{cases} 5a + 2{,}5b = 4{,}0 \\ 2{,}5a + 1{,}45b = 2{,}19 \end{cases}$$

Prin rezolvarea sistemului rezultă: a = 0,325 mii buc., b = 0,95 mil. buc/RON
În consecință, funcția de regresie are formula: $y_{x_i} = 0{,}325 + 0{,}95 \cdot x_i$

Pe baza unei funcții de regresie liniară pot fi previzionate, destul de facil, efectele factorilor de influență. De asemenea, funcțiile liniare de regresie sunt aplicate frecvent în cadrul simulărilor în care se determină modul în care trebuie acționat asupra factorilor controlabili (exprimați prin variabile independente) astfel încât să se obțină anumite efecte.

Exemplul 9.4. Pe baza funcției de regresie liniară determinată în exemplul anterior se cere să se previzioneze care ar fi valorile volumului vânzărilor în situația în care cheltuielile pentru publicitate ar lua două valori: 0,7 mil. RON şi 0,9 mil. RON.

Se cere, de asemenea, să se aprecieze care, ar trebui să fie valoarea cheltuielilor de publicitate pentru ca volumul vânzărilor să reprezinte 0,6 mii bucăți.

Rezolvare: Previziunea în raport cu prima valoare a cheltuielilor de publicitate este o interpolare iar valoarea prognozată a volumului vânzărilor reprezintă:

$$\hat{y}_{x_i(0,7)} = 0{,}325 + 0{,}95 \times x_i = 0{,}325 + 0{,}95 \times 0{,}7 = 0{,}99 \text{ mii buc.}$$

Previziunea în raport cu a doua valoare a cheltuielilor de publicitate este o extrapolare (cea ce înseamnă că acuratețea sa este inferioară față de prima previziune) iar valoarea prognozată a volumului vânzărilor reprezintă:

$$\hat{y}_{x_i(0,9)} = 0{,}325 + 0{,}95 \times x_i = 0{,}325 + 0{,}95 \times 0{,}9 = 1{,}18 \text{ mii buc}$$

În ce priveşte determinarea nivelului cheltuielilor de publicitate pentru care volumul vânzărilor ar reprezenta 0,6 mii bucăți (valoare notată ca y_{opt}) este necesar să se rezolve ecuația:

$\hat{y}_{x_{opt}} = 0{,}325 + 0{,}95 \times x_i$ adică 0,6 = 0,325 + 0,95 × $\hat{x}_i$ de unde rezultă $\hat{x}_i = 0{,}2895$ mil. RON.

9.3.2.1.2. Regresii unifactoriale liniare pentru distribuțiile heterograde

Parametrii regresiilor pentru distribuții heterograde pot fi determinați prin raționamente similare celor folosite în cazul seriilor simple. Dacă se consideră că unitățile din fiecare subgrupă au valorile pentru cele două caracteristice egale cu centrele intervalelor de variație, se ajunge la următoarele relații de echivalență:

- expresia N este echivalentă cu expresia $\sum_{i=1}^{Kx}\sum_{j=1}^{Ky} n_{ij}^{xy} = \sum_{i=1}^{Kx} n_i^x = \sum_{j=1}^{Ky} n_j^j$, unde n_{ij}^{xy} este frecvența subgrupei cu numărul i după caracteristica x şi cu numărul de ordine j după caracteristica y (după cum se ştie, frecvența absolută a unei grupe este egală cu suma frecvențelor absolute ale subgrupelor componente):

$$n_i^x = \sum_{j=1}^{Ky} n_{ij}^{xy} \qquad (9.11)$$

şi

$$n_j^x = \sum_{i=1}^{Ky} n_{ij}^{xy} \qquad (9.12)$$

de unde rezultă:

$$\sum_{i=1}^{Kx} n_i^x = \sum_{i=1}^{Kx}\sum_{j=1}^{Ky} n_{ij}^{xy} = \sum_{j=1}^{Kx}\sum_{i=1}^{Ky} n_{ij}^{xy} = \sum_{j=1}^{Ky} n_j^y \qquad (9.13)$$

- expresia $\sum_{i=1}^{N} x_i$ este echivalentă cu expresia $\sum_{i=1}^{K_x} x_i^{'} \cdot n_i^x$;
- expresia $\sum_{i=1}^{N} y_i$ este echivalentă cu expresia $\sum_{j=1}^{K_y} y^{'} \cdot n_j^y$;
- expresia $\sum_{i=1}^{N} x_i^2$ este echivalentă cu expresia $\sum_{i=1}^{K_x} x_i^{'2} \cdot n_i^x$;
- expresia $\sum_{i=1}^{N} y_i^2$ este echivalentă cu expresia $\sum_{i=1}^{K_x} y_i^2 \cdot n_j^y$;
- expresia $\sum_{i=1}^{N} x_i y_i$ este echivalentă cu expresia $\sum_{i=1}^{K_x}\sum_{j=1}^{kY} x_i^{'} \cdot y_i^{'} \cdot n_{ij}^{xy} =$

$$=\sum_{i=1}^{K_y} y_j^{'} \sum_{j=1}^{kY} x_i^{'} \cdot n_{ij}^{xy} = \sum_{i=1}^{K_x} x_i^{'} \sum_{j=1}^{Ky} y_j^{'} \cdot n_{ij}^{xy}$$

În raport cu aceste relații de echivalență se obțin pentru ecuațiile lui Fermat expresiile:

$$\begin{cases} a\sum_{i=1}^{K_y} n_{ij}^{xy} + b\sum_{i=1}^{Kx} x_i^{'} n_i^{x} = \sum_{j=1}^{Ky} y_j^{'} n_j^{y} \\ a\sum_{i=1}^{K_x} x_i^{'} n^{x} + b\sum_{i=1}^{Kx} x_i^{'2} n^{x} = \sum_{i=1}^{Kx}\sum_{j=1}^{Ky} x_i^{'} y_j^{'} n_j^{xy} \end{cases} \tag{9.14}$$

Exemplul 9.5.: Se cere să se determine parametrii unei funcții de regresie care să exprime dependența numărului mediu zilnic de rebuturi al unui angajat față de vechimea acestuia în muncă, pe baza distribuției heterograde prezentată în tabelul 9.2. Se cere, de asemenea, ca pe baza funcției de regresie să se aprecieze care ar fi numărul mediu zilnic de rebuturi al unui angajat cu vechimea în muncă de cinci ani.

Tabelul 9.4. Valori intermediare utilizate pentru calculul parametrilor funcției de regresie pentru o distribuție heterogradă

x_{i-1} - x_i [ani] / y_{j-1} - y_j [buc.]	(0 ; 4]	(4 ; 8]	(8 ; 12]	(12 ; 16]	(16 ; 20]	Total	Simbol pentru total	y'_j [buc]	$y'_j \cdot n_j^y$ [buc]	$y'^2_j \cdot n_j^y$ [buc^2]	$y'_j \cdot \sum x'_i \cdot n_{ij}^{xy}$ [buc × ani]
(1,0 ; 1,1]	–	–	–	1	3	4	n_1^y	1,05	4,20	4,4100	71,4
(1,1 ; 1,2]	1	1	1	2	–	5	n_2^y	1,15	5,75	6,6125	52,9
(1,2 ; 1,3]	2	1	5	2	–	10	n_3^y	1,25	12,50	15,6250	110,0
(1,3 ; 1,4]	2	1	1	–	–	4	n_4^y	1,35	5,40	7,2900	27,0
(1,4 ; 1,5]	2	–	–	–	–	2	n_5^y	1,45	2,90	4,2050	5,8
Total	7	3	7	5	3	25	$\sum_{j=1}^{Ky} n_j^y$	×	30,75	38,1425	267,1
Simbol pentru total	n_1^x	n_2^x	n_3^x	n_4^x	n_5^x	$\sum_{i=1}^{Kx} n_i^x$	×	×	$\sum_{j=1}^{Ky} y'_j \cdot n_j^y$	$\sum_{j=1}^{Ky} y'^2_j \cdot n_j^y$	$\sum_{i=1}^{Kx} y'_j \sum_{j=1}^{Ky} x'_i \cdot n_{ij}^{xy}$
x'_i [ani]	2	6	10	14	18	×	×				
$x'_i \times n_i^x$ [ani]	14	18	70	70	54	226	$\sum_{i=1}^{Kx} x'_i n_i^x$				
$x'^2_i \times n_i^x$ [ani]	28	108	700	980	972	2.788	$\sum_{i=1}^{Kx} x'^2_i n_i^x$				
$x'_i \cdot \sum_{j=1}^{Ky} y'_j \cdot n_{ij}^{xy}$ [ani × buc]	18,5	22,5	87,5	81,9	56,7	267,1	$\sum_{i=1}^{Kx} x'_i \sum_{j=1}^{Ky} y'_j \cdot n_{ij}^{xy}$				

Rezolvare: În tabelul 9.4. sunt prezentate valorile intermediare care conduc la determinarea parametrilor funcției de regresie. Înlocuind aceste valori în ecuațiile lui Fermat se obține:

$$\begin{cases} 25a + 226b = 30{,}75 \\ 226a + 2.788b = 267{,}1 \end{cases}$$

Prin rezolvarea sistemului rezultă: $\begin{cases} b = -0{,}0146 \text{ bucăți/an} \\ a = 1{,}362 \text{ bucați} \end{cases}$

adică: $y_{x_i} = 1{,}362 - 0{,}0146 x_i$.

Pe baza funcției de regresie se poate aprecia că pentru un angajat cu vechimea în muncă de cinci ani, numărul mediu zilnic de rebuturi ar avea valoarea

$\hat{y}_{x_i}(5) = 1{,}362 - 0{,}0146 \times 5 = 1{,}289$ bucăți.

9.3.2.2. Regresii unifactoriale neliniare

În acest subcapitol vor fi prezentate succint câteva dintre formele de regresii unifactoriale neliniare folosite destul de frecvent în practică:

- regresii polinomiale;
- regresii exponențiale;
- regresii hiperbolice;
- regresii logaritmice.

9.3.2.2.1. Regresii polinomiale

O funcție de regresie polinomială îmbracă forma:

$$y_{x_i} = a_0 + a_1 x_i + a_2 x_i^2 + \ldots + a_p x_i^p \qquad (9.15.)$$

unde p este gradul polinomului asociat funcției.

Modalitatea de determinare a parametrilor unei funcții de regresie polinomială este asemănătoare celei utilizate pentru funcțiile de regresie liniare (de altfel o funcție de regresie liniară poate fi considerată o funcție de regresie polinomială de gradul unu). În continuare vom prezenta modul de calcul al parametrilor unei funcții polinomiale de gradul doi, ce are forma:

$$y_{x_i} = a_0 + a_1 x_i + a_2 x_i^2 \qquad (9.16.)$$

Pentru o serie simplă, funcția care exprimă suma pătratelor diferențelor dintre valorile teoretice şi valorile empirice este dată de relația:

$$S_{(a_0, a_1, a_2)} = \sum_{i=1}^{N} \left(y_{x_i} - y_i \right)^2 = \sum_{i=1}^{N} \left(a_0 + a_1 x_i + a_2 x_i^2 - y_i \right)^2 \qquad (9.17.)$$

Pentru calculul parametrilor a_0, a_1 şi a_2 se pot folosi ecuațiile lui Fermat:

$$\begin{cases} \dfrac{\partial S}{\partial a_0} = 0 \\ \dfrac{\partial S}{\partial a_1} = 0 \\ \dfrac{\partial S}{\partial a_2} = 0 \end{cases} \tag{9.18}$$

Derivata parțială a funcției S în raport cu argumentul a_0 are expresia:

$$\frac{\partial S}{\partial a_0} = \sum_{i=0}^{N} \frac{\partial\left[\left(a_0 + a_1 x_i + a_2 x_i^2 - y_i\right)^2\right]}{\partial a_0} =$$

$$= \sum_{i=0}^{N}\left[2 \times \frac{\partial\left(a_0 + a_1 x_i + a_2 x_i^2 - y_i\right)}{\partial a_0} \times \left(a_0 + a_1 x_i + a_2 x_i^2 - y_i\right)\right] =$$

$$= \sum_{i=0}^{N}\left[2 \times 1 \times \left(a_0 + a_1 x_i + a_2 x_i^2 - y_i\right)\right] = 2\left(\sum_{i=1}^{N} a_0 + \sum_{i=1}^{N} a_1 x_i + \sum_{i=1}^{N} a_2 x_i^2 - \sum_{i=1}^{N} y_i\right) =$$

$$= 2\left(N \cdot a_0 + a_1 \sum_{i=1}^{N} x_i + a_2 \sum_{i=1}^{N} x_i^2 - \sum_{i=1}^{N} y_i\right) \tag{9.19.}$$

Pentru derivata parțială a funcției S în raport cu argumentul a_1 se obține expresia:

$$\frac{\partial S}{\partial a_1} = \sum_{i=0}^{N} \frac{\partial\left[\left(a_0 + a_1 x_i + a_2 x_i^2 - y_i\right)^2\right]}{\partial a_1} =$$

$$= \sum_{i=0}^{N}\left[2 \times \frac{\partial\left(a_0 + a_1 x_i + a_2 x_i^2 - y_i\right)}{\partial a_1} \times \left(a_0 + a_1 x_i + a_2 x_i^2 - y_i\right)\right] =$$

$$= \sum_{i=0}^{N}\left[2 \times x_i \times \left(a_0 + a_1 x_i + a_2 x_i^2 - y_i\right)\right] = 2\sum_{i=0}^{N}\left(a_0 x_i + a_1 x_i^2 + a_2 x_i^3 - y_i x_i\right)$$

$$2\left(\sum_{i=1}^{N} a_0 x_i + \sum_{i=1}^{N} a_1 x_i^2 + \sum_{i=1}^{N} a_2 x_i^3 - \sum_{i=1}^{N} x_i\, y_i\right) =$$

$$= 2\left(a_0 \sum_{i=1}^{N} x_i + a_1 \sum_{i=1}^{N} x_i^2 + a_2 \sum_{i=1}^{N} x_i^3 - \sum_{i=1}^{N} x_i\, y_i\right) \tag{9.20.}$$

În raport cu argumentul a_2, derivata parțială a funcției S are expresia:

$$\frac{\partial S}{\partial a_1} = \sum_{i=0}^{N} \frac{\partial \left[\left(a_0 + a_1 x_i + a_2 x_i^2 - y_i \right)^2 \right]}{\partial a_2} =$$

$$= \sum_{i=0}^{N} \left[2 \times \frac{\partial \left(a_0 + a_1 x_i + a_2 x_i^2 - y_i \right)}{\partial a_2} \times \left(a_0 + a_1 x_i + a_2 x_i^2 - y_i \right) \right] =$$

$$= \sum_{i=0}^{N} \left[2 \times x_i^2 \times \left(a_0 + a_1 x_i + a_2 x_i^2 - y_i \right) \right] = 2 \sum_{i=0}^{N} \left(a_0 x_i^2 + a_1 x_i^3 + a_2 x_i^4 - y_i x_i^2 \right)$$

$$2 \left(\sum_{i=1}^{N} a_0 x_i^2 + \sum_{i=1}^{N} a_1 x_i^3 + \sum_{i=1}^{N} a_2 x_i^4 - \sum_{i=1}^{N} x_i^2 y_i \right) =$$

$$= 2 \left(a_0 \sum_{i=1}^{N} x_i^2 + a_1 \sum_{i=1}^{N} x_i^3 + a_2 \sum_{i=1}^{N} x_i^4 - \sum_{i=1}^{N} x_i^2 y_i \right) \qquad (9.21.)$$

Dacă se introduc expresiile derivatelor parțiale în ecuațiile lui Fermat se obține:

$$\begin{cases} \dfrac{\partial S}{\partial a_0} = 2 \left(N \cdot a_0 + a_1 \sum\limits_{i=1}^{N} x_i + a_2 \sum\limits_{i=1}^{N} x_i^2 - \sum\limits_{i=1}^{N} y_i \right) = 0 \\ \dfrac{\partial S}{\partial a_1} = 2 \left(a_0 \sum\limits_{i=1}^{N} x_i + a_1 \sum\limits_{i=1}^{N} x_i^2 + a_2 \sum\limits_{i=1}^{N} x_i^3 - \sum\limits_{i=1}^{N} x_i \, y_i \right) = 0 \\ \dfrac{\partial S}{\partial a_2} = 2 \left(a_0 \sum\limits_{i=1}^{N} x_i^2 + a_1 \sum\limits_{i=1}^{N} x_i^3 + a_2 \sum\limits_{i=1}^{N} x_i^4 - \sum\limits_{i=1}^{N} x_i^2 y_i \right) = 0 \end{cases} \qquad (9.22.)$$

Rezultă astfel următorul sistem de ecuații care poate fi folosit în determinarea parametrilor a_0, a_1 şi a_2:

$$\begin{cases} N \cdot a_0 + a_1 \sum\limits_{i=1}^{N} x_i + a_2 \sum\limits_{i=1}^{N} x_i^2 = \sum\limits_{i=1}^{N} y_i \\ a_0 \sum\limits_{i=1}^{N} x_i + a_1 \sum\limits_{i=1}^{N} x_i^2 + a_2 \sum\limits_{i=1}^{N} x_i^3 = \sum\limits_{i=1}^{N} x_i \, y_i \\ a_0 \sum\limits_{i=1}^{N} x_i^2 + a_1 \sum\limits_{i=1}^{N} x_i^3 + a_2 \sum\limits_{i=1}^{N} x_i^4 = \sum\limits_{i=1}^{N} x_i^2 y_i \end{cases} \qquad (9.22.)$$

Aceste relații pot fi adaptate şi pentru variabile reprezentate prin distribuții heterograde dacă se folosesc relațiile de echivalență utilizate în cazul regresiei liniare, la care se adaugă încă trei:

- expresia $\sum\limits_{i=1}^{N} x_i^3$ este echivalentă cu expresia $\sum\limits_{i=1}^{N} x_i^{'3} \cdot n_i^x$;
- expresia $\sum\limits_{i=1}^{N} x_i^4$ este echivalentă cu expresia $\sum\limits_{i=1}^{N} x_i^{'4} \cdot n_i^x$;
- expresia $\sum\limits_{i=1}^{N} x_i^2 \cdot y_i$ este echivalentă cu expresia $\sum\limits_{i=1}^{K_x} \sum\limits_{j=1}^{K_y} x_i^{'2} \cdot y_j^{'} \cdot n_{ij}^{xy}$.

Se obține astfel următorul sistem de ecuații care poate fi folosit în determinarea parametrilor unei regresii polinomiale de gradul doi la care se utilizează o distribuție heterogradă:

$$\begin{cases} a_0 \sum_{i=1}^{K_x} \sum_{j=1}^{K_y} n_{ij}^{xy} + a_1 \sum_{i=1}^{K_x} x_i^{'} n_i^x + a_2 \sum_{i=1}^{K_x} x_i^{'2} n_i^x = \sum_{j=1}^{K_y} y_j^{'} n_j^y \\ a_0 \sum_{i=1}^{K_x} x_i^{'} n_i^x + a_1 \sum_{i=1}^{K_x} x_i^{'2} n_i^x + a_2 \sum_{i=1}^{K_x} x_i^{'3} n_i^x = \sum_{i=1}^{K_x} \sum_{j=1}^{K_y} x_i^{'} y_j^{'} n_{ij}^{xy} \\ a_0 \sum_{i=1}^{K_x} x_i^{'2} n_i^x + a_1 \sum_{i=1}^{K_x} x_i^{'3} n_i^x + a_2 \sum_{i=1}^{K_x} x_i^{'4} n_i^x = \sum_{i=1}^{K_x} \sum_{j=1}^{K_y} x_i^{'2} y_j^{'} n_{ij}^{xy} \end{cases} \quad (9.24.)$$

9.3.2.2.2. Regresii exponențiale

O funcție de regresie exponențială are forma:

$$y_{x_i} = a_e \cdot b_e \quad (9.25.)$$

Ecuația funcției, destul de complexă, induce unele dificultăți în aplicarea directă a metodei celor mai mici pătrate. Din acest motiv, în practică, se preferă logaritmarea expresiei funcției:

$$\lg y_{x_i} = \lg\left(a_e \cdot b_e^x\right) = \lg a_e + \lg b_e^x = \lg a_e + x \lg b_e \quad (9.26.)$$

Dacă se fac următoarele transformări:

$y_{x_i}^{'} = \lg y_{x_i}$; $a_e^{'} = \lg a_e$; $b_e^{'} = \lg b_e$;

se ajunge la o funcție de regresie liniară de forma:

$$y_{x_i}^{'} = a_e^{'} + b_e^{'} \cdot x_i \quad (9.27.)$$

pentru care parametrii $a_e^{'}$ și $b_e^{'}$ pot fi determinați printr-o modalitate prezentată anterior. Odată calculate valorile teoretice $y_{x_i}^{'}$ acestea pot fi transformate, prin antilogaritmare, în valorile teoretice y_{x_i}: $y_{x_i} = anti\log\left(y_{x_i}^{'}\right)$

(9.29.)

9.3.2.2.3. Regresii hiperbolice

O funcție de regresie hiperbolică are ecuația:

$$y_{x_i} = a_h + b_h \cdot \frac{1}{x_i} \quad (9.29.)$$

Modalitatea de determinare a parametrilor a_h și b_h este similară celei aplicată în cazul regresiei liniare. Ecuațiile lui Fermat au, în această situație, forma:

$$\begin{cases} N \cdot a_h + b_h \cdot \sum_{i=1}^{N} \frac{1}{x_i} = \sum_{i=1}^{N} y_i \\ a_h \cdot \sum_{i=1}^{N} \frac{1}{x_i} + b_h \cdot \sum_{i=1}^{N} \frac{1}{x_i^2} = \sum_{i=1}^{N} \frac{1}{x_i} \cdot y_i \end{cases} \quad (9.30.)$$

Ecuaţiile utilizate pentru o serie simplă pot fi adaptate şi pentru variabile reprezentate prin distribuţii heterograde. Într-o astfel de situaţie se pot folosi relaţiile de echivalenţă folosite pentru regresia liniară, la care se adaugă încă trei:

- expresia $\sum_{i=1}^{N} \frac{1}{x_i}$ este echivalentă cu expresia $\sum_{i=1}^{K_x} \frac{1}{x_i^{'}} \cdot n_i^x$;
- expresia $\sum_{i=1}^{N} \frac{1}{x_i^2}$ este echivalentă cu expresia $\sum_{i=1}^{K_x} \frac{1}{x_i^{'2}} \cdot n_i^x$;
- expresia $\sum_{i=1}^{N} \frac{1}{x_i} \cdot y_i$ este echivalentă cu expresia $\sum_{i=1}^{K_x} \sum_{j=1}^{K_y} \frac{1}{x_i^{'}} \cdot y_j^{'} \cdot n_{ij}^{xy}$.

Se obţine astfel sistemul de ecuaţii pe baza cărora pot fi determinaţi parametrii unei regresii hiperbolice la care se utilizează o distribuţie heterogradă:

$$\begin{cases} a_h \sum_{i=1}^{K_x} \sum_{j=1}^{K_y} n_{ij}^{xy} + b_h \cdot \sum_{i=1}^{K_x} \frac{1}{x_i^{'}} \cdot n_i^x = \sum_{j=1}^{K_y} y_j^{'} y_j^{y} \\ a_h \cdot \sum_{i=1}^{K_x} \frac{1}{x_i^{'}} \cdot n_i^x + b_h \cdot \sum_{i=1}^{K_x} \frac{1}{x_i^{'2}} \cdot n_i^x = \sum_{i=1}^{K_x} \sum_{j=1}^{K_y} \frac{1}{x_i} \cdot y_j^{'} n_{ij}^{xy} \end{cases} \quad (9.31.)$$

9.3.2.2.4. Regresii logaritmice

O funcţie de regresie logaritmică are expresia:

$$y_{x_i} = a_l + b_l \cdot \lg x_i \quad (9.32.)$$

Parametrii a_l şi b_l pot fi determinaţi printr-un raţionament similar celui folosit în cazul regresiei liniare. Pentru regresia logaritmică ecuaţiile lui Fermat au forma:

$$\begin{cases} N \cdot a_l + b_l \cdot \sum_{i=1}^{N} \lg x_i = \sum_{i=1}^{N} y_i \\ a_l \cdot \sum_{i=1}^{N} \lg x_i + b_l \cdot \sum_{i=1}^{N} (\lg x_l)^2 = \sum_{i=1}^{N} (y_i \cdot \lg x_i) \end{cases} \quad (9.33.)$$

Ecuaţiile regresiei logaritmice stabilite pentru o serie simplă pot fi adaptate pentru o distribuţie heterogradă. În acest scop pot fi utilizate relaţiile de echivalenţă formulate pentru regresia liniară, la care se adaugă încă trei:

- expresia $\sum_{i=1}^{N} \lg x_i$ este echivalentă cu expresia $\sum_{i=1}^{K_x} [(\lg x_i) \cdot n_i^x]$;

- expresia $\sum_{i=1}^{N}(\lg x_i)^2$ este echivalentă cu expresia $\sum_{i=1}^{K_x}\left[(\lg x_i)^2 \cdot n_i^x\right]$;
- expresia $\sum_{i=1}^{N} y_i \cdot \lg x_i$ este echivalentă cu expresia $\sum_{i=1}^{K_x}\sum_{j=1}^{K_y}(\lg x_i)\cdot y_j^{'} n_{ij}^{xy}$.

Se obține astfel un sistem de ecuații care poate fi folosit în calculul parametrilor unei regresii logaritmice la care se utilizează o distribuție heterogradă:

$$\begin{cases} a_l \sum_{i=1}^{K_x}\sum_{j=1}^{K_y} n_{ij}^{xy} + b_l \cdot \sum_{i=1}^{K_x}(\lg x_i^{'}) n_{ix} = \sum_{i=1}^{K_y} y_j^{'} \cdot n_j^y \\ a_l \cdot \sum_{i=1}^{K_x}(\lg x_i^{'})\cdot n_{ix} + b_l \cdot \sum_{i=1}^{K_x}(\lg x_i^{'})^2 \cdot n_{ix} = \sum_{i=1}^{K_x}\sum_{j=1}^{K_y}(\lg x_i)\cdot y_j^{'} n_{ij}^{xy} \end{cases} \quad (9.34.)$$

9.3.3. Regresii multifactoriale

Regresiile multifactoriale sunt folosite, de regulă, în cazul unor fenomene desfăşurate în condiții complexe, pentru care factorii relevanți de influență nu pot fi exprimați printr-o singură variabilă independentă. La fel ca în cazul regresiei unifactoriale, funcțiile folosite în regresia multifactorială pot îmbrăca diferite forme: liniare, polinomiale, exponențiale, hiperbolice, logaritmice etc. De exemplu, o funcție de regresie multifactorială liniară cu două variabile are forma:

$$y_{x_i} = a_{m_0} + a_{m_1} x_{1i} + a_{m_2} x_{i2} \quad (9.35.)$$

Funcția S care exprimă suma pătratelor diferențelor dintre valorile teoretice şi valorile empirice are, în acest caz, expresia:

$$S_{(a_{m_0}, a_{m_1}, a_{m_2})} = \sum_{i=1}^{N}\left(y_{x_i} - y_i\right)^2 = \sum\left(a_{m_0} + a_{m_i} \cdot x_{1i} + a_{m_2} x_{2i} - y_i\right)^2 \quad (9.36.)$$

Parametrii a_{m_0}, a_{m_1} şi a_{m_2} ai funcției de regresie pot fi determinați pe baza ecuațiilor lui Fermat.

$$\begin{cases} \frac{\partial S}{\partial a_{m_0}} = 0 \\ \frac{\partial S}{\partial a_{m_1}} = 0 \\ \frac{\partial S}{\partial a_{m_2}} = 0 \end{cases} \quad (9.36.)$$

Derivata parțială a funcției S în raport cu argumentul a_{m_0} are expresia:

$$\frac{\partial S}{\partial a_{m_0}} = \sum_{i=1}^{N} \frac{\partial\left[\left(a_{m_0} + a_{m_1} \cdot x_{1i} + a_{m_2} \cdot x_{2i}\right)^2\right]}{\partial a_{m_0}} =$$

$$= \sum_{i=1}^{N} 2 \cdot \left[\frac{\partial\left(a_{m_0} + a_{m_1} \cdot x_{1i} + a_{m_2} \cdot x_{2i}\right)}{\partial a_{m_0}} \cdot \left(a_{m_0} + a_{m_1} \cdot x_{1i} + a_{m_2} \cdot x_{2i}\right)\right] =$$

$$= \sum_{i=1}^{N} 2 \cdot 1 \cdot \left(a_{m_0} + a_{m_1} \cdot x_{1i} + a_{m_2} \cdot x_{2i}\right) = \qquad (9.38.)$$

$$= 2\sum_{i=1}^{N} a_{m_0} + \sum_{i=1}^{N} a_{m_1} \cdot x_{1i} + \sum_{i=1}^{N} a_{m_2} \cdot x_{2i} - \sum_{i=1}^{N} y_i =$$

$$= 2\left(N \cdot a_{m_0} + a_{m_1} \sum_{i=1}^{N} x_{1i} + a_{m_2} \sum_{i=1}^{N} x_{2i} - \sum_{i=1}^{N} y_i\right)$$

Pentru derivata parțială a funcției *S* în raport cu argumentul a_{m_1} se obține expresia:

$$\frac{\partial S}{\partial a_{m_1}} = \sum_{i=1}^{N} \frac{\partial\left[\left(a_{m_0} + a_{m_1} \cdot x_{1i} + a_{m_2} \cdot x_{2i} - y_i\right)^2\right]}{\partial a_{m_1}} =$$

$$= \sum_{i=1}^{N}\left[2 \times \frac{\partial\left(a_{m_0} + a_{m_1} \cdot x_{1i} + a_{m_2} \cdot x_{2i} - y_i\right)}{\partial a_{m_1}}\left(a_{m_0} + a_{m_1} \cdot x_{1i} + a_{m_2} \cdot x_{2i} - y_i\right)\right] =$$

$$= \sum_{i=1}^{N}\left[2 \cdot x_{1i}\left(a_{m_0} + a_{m_1} \cdot x_{1i} + a_{m_2} \cdot x_{2i} - y_i\right)\right] =$$

$$= 2\sum_{i=1}^{N}\left(a_{m_0} \cdot x_{1i} + a_{m_1} \cdot x_{1i}^2 + a_{m_2} \cdot x_{1i} \cdot x_{2i} - x_{1i} \cdot y_i\right) =$$

$$= 2\left(\sum_{i=1}^{N} a_{m_0} \cdot x_{1i} + \sum_{i=1}^{N} a_{m_1} \cdot x_{1i}^2 + \sum_{i=1}^{N} a_{m_2} \cdot x_{1i} \cdot x_{2i} - \sum_{i=1}^{N} x_{1i} \cdot y_i\right) =$$

$$= 2\left(a_{m_0} \sum_{i=1}^{N} x_{1i} + a_{m_1} \sum_{i=1}^{N} x_{1i}^2 + a_{m_2} \sum_{i=1}^{N} x_{1i} \cdot x_{2i} - \sum_{i=1}^{N} x_{1i} \cdot y_i\right) \qquad (9.38.)$$

În raport cu argumentul a_{m_2}, derivata parțială a funcției *S* are expresia:

$$\frac{\partial S}{\partial a_{m_1}} = \sum_{i=1}^{N} \frac{\partial\left[\left(a_{m_0} + a_{m_1} \cdot x_{1i} + a_{m_2} \cdot x_{2i} - y_i\right)^2\right]}{\partial a_{m_2}} =$$

$$= \sum_{i=1}^{N}\left[2 \times \frac{\partial\left(a_{m_0} + a_{m_1} \cdot x_{1i} + a_{m_2} \cdot x_{2i} - y_i\right)}{\partial a_{m_2}}\left(a_{m_0} + a_{m_1} \cdot x_{1i} + a_{m_2} \cdot x_{2i} - y_i\right)\right] =$$

$$= \sum_{i=1}^{N}\left[2 \cdot x_{2i}\left(a_{m_0} + a_{m_1} \cdot x_{1i} + a_{m_2} \cdot x_{2i} - y_i\right)\right] =$$

$$= 2\sum_{i=1}^{N}\left(a_{m_0} \cdot x_{2i} + a_{m_1} \cdot x_{1i} \cdot x_{2i} + a_{m_2} \cdot x_{2i}^2 - x_{2i} \cdot y_i\right) =$$

$$= 2\left(\sum_{i=1}^{N} a_{m_0} \cdot x_{2i} + \sum_{i=1}^{N} a_{m_1} \cdot x_{1i} \cdot x_{2i} + \sum_{i=1}^{N} a_{m_2} \cdot x_{2i}^2 - \sum_{i=1}^{N} x_{2i} \cdot y_i\right) =$$

$$= 2\left(a_{m_0}\sum_{i=1}^{N} x_{2i} + a_{m_1}\sum_{i=1}^{N} x_{1i} \cdot x_{2i} + a_{m_2}\sum_{i=1}^{N} x_{2i}^2 - \sum_{i=1}^{N} x_{2i} \cdot y_i\right) \qquad (9.39.)$$

Introducând expresiile derivatelor parțiale în ecuațiile lui Fermat se obține:

$$\begin{cases} \dfrac{\partial S}{\partial a_{m_0}} = 2\left(N \cdot a_{m_0} + a_{m_1}\sum_{i=1}^{N} x_{1i} + a_{m_2}\sum_{i=1}^{N} x_{2i} - \sum_{i=1}^{N} y_i\right) = 0 \\ \dfrac{\partial S}{\partial a_{m_1}} = 2\left(a_{m_0}\sum_{i=1}^{N} x_{1i} + a_{m_1}\sum_{i=1}^{N} x_{1i}^2 + a_{m_2}\sum_{i=1}^{N} x_{1i} \cdot x_{2i} - \sum_{i=1}^{N} x_{1i} \cdot y_i\right) = 0 \\ \dfrac{\partial S}{\partial a_{m_2}} = 2\left(a_{m_0}\sum_{i=1}^{N} x_{2i} + a_{m_1}\sum_{i=1}^{N} x_{1i} \cdot x_{2i} + a_{m_2}\sum_{i=1}^{N} x_{2i}^2 - \sum_{i=1}^{N} x_{2i} \cdot y_i\right) = 0 \end{cases} \quad (9.40.)$$

Rezultă astfel un sistem de ecuații prin a cărui rezolvare pot fi obținute valorile parametrilor a_{m_0}, a_{m_1} şi a_{m_1}:

$$\begin{cases} N \cdot a_{m_0} + a_{m_1}\sum_{i=1}^{N} x_{1i} + a_{m_2}\sum_{i=1}^{N} x_{2i} = \sum_{i=1}^{N} y_i \\ a_{m_0}\sum_{i=1}^{N} x_{1i} + a_{m_1}\sum_{i=1}^{N} x_{1i}^2 + a_{m_2}\sum_{i=1}^{N} x_{1i} \cdot x_{2i} = \sum_{i=1}^{N} x_{1i} \cdot y_i \\ a_{m_0}\sum_{i=1}^{N} x_{2i} + a_{m_1}\sum_{i=1}^{N} x_{1i} \cdot x_{2i} + a_{m_2}\sum_{i=1}^{N} x_{2i}^2 = \sum_{i=1}^{N} x_{2i} \cdot y_i \end{cases} \qquad (9.41.)$$

9.4. Indicatori de apreciere a sensului şi intensității legăturilor dintre variabile

În acest subcapitol vor fi prezentate succint cinci mărimi utilizate destul de frecvent în cuantificarea sensului şi intensității legăturilor dintre variabile:

- coeficientul de asociere;
- covarianța;

- coeficientul de corelație liniară simplă;
- coeficientul de determinare;
- raportul de corelație.

9.4.1. Coeficientul de asociere

Coeficientul de asociere este o mărime propusă de statisticianul G.U. Yule pentru evaluarea legăturii dintre două atribute de ordin calitativ. Pentru determinarea coeficientului de asociere este necesar ca populația studiată să fie împărțită, în raport cu cele două atribute, notate cu *A* şi *B*, în patru subgrupe (tabelul 9.5.):

- unitățile care au atât atributul *A* cât şi atributul *B*, al căror număr este notat cu *AB*;
- unitățile care au atributul *A* dar nu au atributul *B* ci opusul acestuia β, al căror număr este notat cu $A\beta$;
- unitățile care nu au atributul *A* ci opusul acestuia α şi care au atributul *B*, al căror număr este notat cu $B\alpha$;
- unitățile care nu au nici atributul *A* nici atributul *B* ci opusele acestora, adică α, respectiv β, al căror număr este notata cu $\alpha\beta$.

Tabelul 9.5. Împărțirea unei populații statistice în raport cu două atribute

Primul atribut / Al doilea atribut	A	α	Total
B	AB	$B\alpha$	B
β	$A\beta$	$\alpha\beta$	β
Total	A	α	N

Valoarea coeficientului de asociere, notat cu Q_{as}, este dată de relația:

$$Q_{as} = \frac{AB \times \alpha\beta - A\beta \times B\alpha}{AB \times \alpha\beta + A\beta \times B\alpha} \qquad (9.42.)$$

Domeniul de variație a coeficientului de asociere este reprezentat de intervalul [–1 ; 1]. O valoare negativă indică o legătură inversă între atributul *A* şi atributul *B* în timp ce o valoare pozitivă semnifică o legătură directă. Intensitatea legăturii este cu atât mai mare cu cât valoarea absolută a coeficientului este mai mare. În tabelul 9.6. sunt prezentate intervalele de valori ale mărimii în raport cu care sunt apreciate sensul şi intensitatea legăturii.

Tabelul 9.6. Aprecierea sensului şi intensităţii unei legături în raport cu coeficientul de asociere

Nr. crt.	Valori ale coeficientului de asociere	Apreciere asupra sensului şi intensităţii legăturii
1.	$Q_{as} = -1$	legătură inversă deterministă
2.	$-1 < Q_{as} < -0{,}9$	legătură inversă foarte pronunţată
3.	$-0{,}9 \le Q_{as} < -0{,}7$	legătură inversă pronunţată
4.	$-0{,}7 \le Q_{as} < -0{,}5$	legătură inversă moderată
5.	$-0{,}5 \le Q_{as} < -0{,}3$	legătură inversă slabă
6.	$-0{,}3 \le Q_{as} < 0$	legătură inversă foarte slabă
7.	$Q_{as} = 0$	nu există legătură între cele două variabile
8.	$0 < Q_{as} \le 0{,}3$	legătură directă foarte slabă
9.	$0{,}3 < Q_{as} \le 0{,}5$	legătură directă slabă
10.	$0{,}5 < Q_{as} \le 0{,}7$	legătură directă moderată
11.	$0{,}7 < Q_{as} \le 0{,}9$	legătură directă pronunţată
12.	$0{,}9 < Q_{as} \le 1$	legătură directă foarte pronunţată
13.	$Q_{as} = 0$	legătură directă deterministă

Exemplul 9.6. Clienţii unei firme au fost grupaţi în raport cu două caracteristici: sexul şi onorarea facturilor (tabelul 9.7.). Se cere să se aprecieze, pe baza coeficientului de asociere, legătura care se poate face între sexul masculin al clienţilor şi calitatea acestora de restanţier.

Tabelul 9.7. Gruparea clienţilor unei firme în raport cu sexul şi cu onorarea facturilor

Sex / Onorarea facturilor	Bărbaţi (A)	Femei (α)	Total
Restanţieri (B)	15	5	20
Buni platnici (β)	45	35	80
Total	60	40	100

Rezolvare: Valoarea coeficientului de asociere reprezintă:

$$Q_{as} = \frac{AB \times \alpha\beta - A\beta \times B\alpha}{AB \times \alpha\beta + A\beta \times B\alpha} = \frac{15 \times 35 - 45 \times 5}{15 \times 35 + 45 \times 5} = 0{,}4\,,$$

ceea ce înseamnă că între calitatea de bărbat şi cea de restanţier există o legătură directă dar slabă.

9.4.2. Covarianţa dintre două variabile

Covarianţa dintre două variabile, x şi y, este o mărime, notată cu $\mathrm{cov}(x,y)$ şi care se poate calcula prin formula:

$$\mathrm{cov}(x,y)=\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y}) \qquad (9.42.)$$

Pe baza valorii covarianţei pot fi apreciate atât sensul cât şi intensitatea legăturii dintre cele două variabile.

Atunci când legătura este inversă, adică variabilele evoluează în sensuri opuse, valorilor peste medie ale unei variabile le vor corespunde, în general, valori sub medie ale celeilalte variabile, astfel încât valoarea covarianţei este negativă. În schimb, atunci când legătura este directă, iar variabilele evoluează în acelaşi sens, valoarea covarianţei este pozitivă întrucât, pentru o unitate statistică, valorile celor două variabile vor fi, în general, fie ambele peste medie, fie ambele sub medie.

Se poate demonstra că valoarea absolută a covarianţei nu poate depăşi produsul dintre abaterile medii pătratice ale celor două variabile. În consecinţă, covarianţa dintre două variabile x şi y are ca domeniu de variaţie intervalul $[-\sigma_x \cdot \sigma_y;+\sigma_x \cdot \sigma_y]$. Valorile absolute ale covarianţei sunt cu atât mai mari cu cât legătura este mai intensă. Pentru o legătură deterministă covarianţa atinge una dintre limitele intervalului în timp ce valoarea nulă este atinsă atunci când între cele două variabile nu există nicio legătură.

Aprecierea intensităţii unei legături pe baza covarianţei dintre variabile, este facilitată de simplitatea modului de calcul. Totuşi, pe baza acestei mărimi nu pot fi făcute încadrări sau comparaţii asupra intensităţii.

9.4.3. Coeficientul de corelaţie liniară simplă

Coeficientul de corelaţie liniară simplă este o mărime, notată cu r_{xy}, prin care pot fi apreciate sensul şi intensitatea unei legături ce poate fi exprimată printr-o funcţie liniară. Valoarea sa poate fi calculată raportând covarianţa la produsul dintre abaterile medii pătratice ale celor două variabile:

$$r_{xy}=\frac{\mathrm{cov}(x,y)}{\sigma_x \cdot \sigma_y} \qquad (9.43.)$$

În condiţiile în care valoarea absolută a covarianţei nu poate fi mai mare decât produsul, domeniul de variaţie al acestei mărimi va fi reprezentat de intervalul [–1;1]. Coeficientul de corelaţie liniară simplă are, la fel ca şi covarianţa, o valoare pozitivă în cazul unei legături directe şi o valoare negativă în cazul unei legături inverse. Valoarea sa absolută este cu atât mai mare cu cât legătura dintre cele două variabile este mai intensă. Fiind o mărime adimensională, coeficientul de corelaţie liniară simplă are, în comparaţie cu covarianţa, avantajul că poate fi folosit pentru încadrarea intensităţii şi pentru comparaţii între serii. Valorile sale au, în ce priveşte sensul şi intensitatea unei legături liniare, aceleaşi semnificaţii (prezentate în tabelul 9.6.) pe care le au valorile coeficientului de asociere.

Valoarea coeficientului de corelaţie liniară simplă poate fi folosită în verificarea ipotezelor statistice asupra unei legături semnificative între două variabile. În acest scop poate fi folosit un procedeu numit *testul Student* ce utilizează o distribuție t în care numărul de grade de libertate este dat de relația:

$$\nu = N - 2 \tag{9.44.}$$

unde N este numărul de unități statistice folosit în studiul legăturii dintre cele două variabile. Testul Student presupune formularea a două ipoteze:

- ipoteza nulă H_0: „coeficientul de corelație liniară simplă diferă semnificativ de zero" (altfel spus, între cele două variabile există o legătură semnificativă);
- ipoteza alternativă H_A: „coeficientul de corelație liniară simplă nu diferă semnificativ de zero" (altfel spus, între cele două variabile nu există o legătură semnificativă).

În continuare, se calculează o mărime numită *valoarea testului statistic Student pentru coeficientul de corelație liniară simplă*, notată cu t_r şi dată de formula:

$$t_r = \frac{r_{xy}}{\sqrt{1 - r_{xy}^2}} \cdot \sqrt{\nu} \tag{9.45.}$$

Această valoare calculată se compară cu o valoare tabelată t_ν^q, obținută în raport cu numărul de grade de libertate şi de nivelul de încredere dorit pentru verificarea ipotezei statistice, rezultând una din următoarele situații:

- dacă $t_r \geq t_\nu^q$ se admite ipoteza nulă;
- dacă $t_r < t_\nu^q$ se respinge ipoteza nulă.

În practică, se obişnuieşte ca valoarea coeficientului de corelație liniară simplă să fie calculată printr-o aşa-numită *formulă simplificată*:

$$r_{xy} = \frac{N\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{\sqrt{\left[N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2\right]\left[N\sum_{i=1}^{N} y_i^2 - \left(\sum_{i=1}^{N} y_i\right)^2\right]}} \tag{9.46.}$$

Exemplul 9.7. Se cere să se aprecieze, pe baza seriei statistice prezentată în tabelul 9.1., sensul şi intensitatea legăturii liniare dintre volumul desfacerilor şi cheltuielile pentru publicitate prin intermediul coeficientului de corelație liniară simplă.

Rezolvare: Coeficientul de corelație liniară simplă poate fi determinat pe baza formulei de calcul simplificat, utilizând valorile intermediare prezentate în tabelul 9.3.:

$$r_{xy}=\frac{N\sum_{i=1}^{N}x_i y_i-\sum_{i=1}^{N}x_i\sum_{i=1}^{N}y_i}{\sqrt{\left[N\sum_{i=1}^{N}x_i^2-\left(\sum_{i=1}^{N}x_i\right)^2\right]\left[N\sum_{i=1}^{N}y_i^2-\left(\sum_{i=1}^{N}y_i\right)^2\right]}}=$$

$$=\frac{5\times 2{,}19-2{,}5\times 4{,}0}{\sqrt{\left(5\times 1{,}45-2{,}5^2\right)\left(5\times 3{,}4-4{,}0^2\right)}}=0{,}95$$

În raport cu această valoare se poate aprecia că între volumul desfacerilor şi cheltuielile pentru publicitate există o legătură directă foarte pronunţată.

Relaţia de calcul simplificat al coeficientului de corelaţie simplă aplicabilă pentru seriile simple poate fi adaptată şi pentru distribuţiile heterograde, pe baza relaţiilor de echivalenţă utilizate în cazul regresiei liniare. Se obţine astfel următoarea formulă de calcul.

$$r_{xy}=\frac{\sum_{i=1}^{K_x}\sum_{j=1}^{K_y}n_{ij}^{xy}\cdot\sum_{i=1}^{K_x}\sum_{j=1}^{K_y}x_i'\cdot y_j'\cdot n_{ij}^{xy}-\sum_{i=1}^{K_x}x_i'\cdot n_i^x\cdot\sum_{j=1}^{K_y}y_j'\cdot n_j^y}{\sqrt{\left[\sum_{i=1}^{K_x}\sum_{j=1}^{K_y}n_{ij}^{xy}\cdot\sum_{i=1}^{K_x}x_i'^2\cdot n_i^x-\left(\sum_{i=1}^{K_x}x_i'\cdot n_i^x\right)^2\right]\left[\sum_{i=1}^{K_x}\sum_{j=1}^{K_y}n_{ij}^{xy}\cdot\sum_{j=1}^{K_y}y_j'^2\cdot n_j^y-\left(\sum_{j=1}^{K_y}y_j'\cdot n_j^y\right)^2\right]}}$$

Exemplul 9.8. Se cere să se aprecieze, pe baza distribuţiei heterograde prezentată în tabelul 9.2., sensul şi intensitatea legăturii dintre vechimea în muncă şi numărul mediu zilnic de rebuturi prin intermediul coeficientului de corelaţie liniară simplă. Se cere, de asemenea, să se verifice, pe baza testului Student, ipoteza unei legături semnificative între cele două variabile, pe baza unui nivel de semnificaţie $\alpha = 0{,}01$.

Rezolvare: Aplicând formula de calcul simplificat, în care sunt introduse valorile intermediare prezentate în tabelul 9.4. rezultă:

$$r_{xy}=\frac{\sum_{i=1}^{K_x}\sum_{j=1}^{K_y}n_{ij}^{xy}\cdot\sum_{i=1}^{K_x}\sum_{j=1}^{K_y}x_i'\cdot y_j'\cdot n_{ij}^{xy}-\sum_{i=1}^{K_x}x_i'\cdot n_i^x\cdot\sum_{j=1}^{K_y}y_j'\cdot n_j^y}{\sqrt{\left[\sum_{i=1}^{K_x}\sum_{j=1}^{K_y}n_{ij}^{xy}\cdot\sum_{i=1}^{K_x}x_i'^2\cdot n_i^x-\left(\sum_{i=1}^{K_x}x_i'\cdot n_i^x\right)^2\right]\left[\sum_{i=1}^{K_x}\sum_{j=1}^{K_y}n_{ij}^{xy}\cdot\sum_{j=1}^{K_y}y_j'^2\cdot n_j^y-\left(\sum_{j=1}^{K_y}y_j'\cdot n_j^y\right)^2\right]}}=$$

$$=\frac{25\times 267{,}1-226\times 30{,}75}{\sqrt{\left(25\times 2.788-226^2\right)\left(25\times 38{,}1425-30{,}75^2\right)}}=-0{,}7047$$

Valoarea coeficientului de corelaţie liniară simplă indică o legătură inversă pronunţată între cele două variabile.

Pentru aplicarea testului Student asupra relevanţei legăturii dintre cele două variabile sunt formulate două ipoteze statistice:

- ipoteza nulă H_0: „r_{xy} diferă semnificativ de zero”;
- ipoteza alternativă H_A: „r_{xy} nu diferă semnificativ de zero”;

Numărul de grade de libertate al distribuției t folosite reprezintă:

$$\nu = \sum_{i=1}^{K_x} \sum_{j=1}^{K_y} n_{ix} - 2 = 25 - 2 = 23$$

În raport cu numărul de grade de libertate şi cu nivelul de încredere α = 0,01 se obține o valoare tabelată t_{ν}^{q} = 2,807. Această valoare se compară cu valoarea testului Student:

$$t_r = \frac{r_{xy}}{\sqrt{1 - r_{xy}^2}} \cdot \sqrt{\nu} = \frac{0{,}7047}{\sqrt{1 - 0{,}7047^2}} \cdot \sqrt{23} = 4{,}7634$$

Întrucât valoarea lui t_r este mai mare decât valoarea tabelată t_{ν}^{q} se poate accepta ipoteza nulă.

9.4.4. Coeficientul de determinare

Coeficientul de determinare este o mărime, notată cu η_d prin care poate fi evaluată intensitatea unei legături între două variabile pentru care a fost stabilită o funcție de regresie liniară sau neliniară. Valoarea sa este dată de relația:

$$\eta_d^2 = \frac{\sigma_{y_{xi}}^2}{\sigma_{y_i}^2} \qquad (9.48.)$$

în care: - $\sigma_{y_{xi}}^2$ este dispersia valorilor teoretice ale variabilei dependente;

-$\sigma_{y_i}^2$ este dispersia valorilor empirice ale variabilei dependente.

În principiu, valorile teoretice sunt doar rezultatul factorilor de influență care au fost considerați relevanți în cadrul regresiei şi care sunt exprimați prin intermediul variabilei independente. În schimb, valorile empirice sunt rezultatul tuturor factorilor de influență, inclusiv a celor care au fost considerați nerelevanți şi care nu au fost exprimați prin variabila independentă.

Raportul dintre dispersia valorilor teoretice şi dispersia valorilor empirice ale variabilei dependente reflectă gradul în care valorile empirice sunt influențate de factorii exprimați prin variabila independentă. Cele două dispersii nu pot avea decât valori pozitive, iar dispersia valorilor teoretice este cel mult egală cu dispersia valorilor empirice, astfel încât domeniul de variație al coeficientului de determinare este reprezentat de intervalul [0;1]. Exprimat în termeni procentuali, coeficientul de determinare reflectă proporția în care valorile variabilei dependente sunt datorate factorilor exprimați prin variabila independentă.

Exemplul 9.9. Se cere să se aprecieze, pe baza seriei statistice prezentată în tabelul 9.1., prin intermediul coeficientului de determinare, impactul pe care îl au cheltuielile pentru publicitate asupra volumului desfacerilor.

Rezolvare: În tabelul 9.8. sunt prezentate valorile intermediare pentru calculul dispersiilor valorilor teoretice şi ale valorilor empirice ale variabilei dependente. Pentru ambele variabile, media aritmetică are aceeaşi valoare:

$$\overline{y} = \frac{\sum_{i=1}^{N} y_i}{N} = \frac{4,0}{5} = 0,8 \text{ mii buc.};$$

sau,

$$\overline{y}_{x_i} = \frac{\sum_{i=1}^{N} y_{x_i}}{N} = \frac{4,0}{5} = 0,8 \text{ mii buc.}$$

Dispersia valorilor teoretice reprezintă:

$$\sigma^2_{y_{x_i}} = \frac{\sum_{i=1}^{N} \left(y_{x_i} - \overline{y}_{x_i}\right)^2}{N} = \frac{0,1805}{5} = 0,0361 \text{ (mii buc.)}^2$$

Tabelul 9.9. Valori intermediare utilizate în calculul coeficientului de determinare

Nr. crt.	x_i [mil. RON]	y_i [mii buc.]	$y_{x_i} = a + bx_i$ [mii buc.]	$(y_i - \overline{y})^2$ [(mii buc.)^2]	$(y_{x_i} - \overline{y}_{x_i})^2$ [(mii buc.)^2]
(0)	(1)	(2)	(3)	(4)	(5)
1.	0,2	0,5	0,515	0,09	0,0812
2.	0,8	1,1	1,085	0,09	0,0812
3.	0,5	0,7	0,80	0,01	0
4.	0,6	0,9	0,895	0,01	0,009
5.	0,4	0,8	0,705	0	0,009
Total	2,5	4,0	4,0	0,2	0,1805
Simbol pentru total	$\sum_{i=1}^{N} x_i$	$\sum_{i=1}^{N} y_i$	$\sum_{i=1}^{N} y_{x_i}$	$\sum_{i=1}^{N} \left(y_i - \overline{y}\right)^2$	$\sum_{i=1}^{N} \left(y_{x_i} - \overline{y}_{x_i}\right)^2$

Coeficientul de determinare are valoarea:

$$\eta^2_d = \frac{\sigma^2_{y_{x_i}}}{\sigma^2_{y_i}} = \frac{0,0361}{0,04} = 0,9025,$$

ceea ce înseamnă că în medie 90,25% din valoarea volumului desfacerilor se datorează cheltuielilor pentru publicitate în timp ce restul de 9,75% se datorează altor factori.

Atunci când variabilele sunt reprezentate prin distribuții heterograde, calculul dispersiei valorilor teoretic presupune ca acestea să fie grupate pe baza grupelor constituite în raport cu variabila independentă.

Exemplul 9.10. Se cere să se aprecieze, pe baza distribuției heterograde prezentată în tabelul 9.3., prin intermediul coeficientului de determinare, impactul pe care îl are vechimea în muncă asupra numărului mediu zilnic de rebuturi.

Rezolvare: În tabelul 9.10. sunt prezentate valorile intermediare utilizate în calculul dispersiei valorilor teoretice. Media aritmetică a valorilor teoretice reprezintă:

$$\overline{y}_{x_i} = \frac{\sum_{i=1}^{K_x} x_i^{'} \cdot n_i^x}{\sum_{i=1}^{K_x} n_{ix}} = \frac{30{,}7504}{25} = 1{,}23 \text{ bucăți}$$

Dispersia valorilor teoretice are valoarea:

$$\sigma_{y_{x_i}}^2 = \frac{\sum_{i=1}^{K_x} \left(y_{x_i}^{'} - \overline{y}_{x_i}\right)^2 \cdot n_i^x}{\sum_{i=1}^{K_x} n_i^x} = \frac{0{,}1588}{25} = 0{,}0064 \text{ buc}^2$$

Tabelul 9.10. Valori intermediare utilizate în calculul dispersiei valorilor teoretice

Nr. crt.	Interval de variație $(x_{i-1} - x_i)$ [ani]	$x_i^{'}$ [ani]	n_i^x	$y_{x_i}^{'} = a + bx_i^{'}$ [buc.]	$y_{x_i}^{'} \cdot n_i^x$ [buc.]	$(y_{x_i}^{'} - \overline{y}_{x_i})^2 \cdot n_i^x$ [buc.^{2}]
(0)	(1)	(2)	(3)	(4)	(5) = (4) × (3)	(6)
1.	(0 ; 4]	2	7	1,3328	9,3296	0,0740
2.	(4 ; 8]	6	3	1,2744	3,8232	0,0059
3.	(8 ; 12]	10	7	1,2160	8,5120	0,0014
4.	(12 ; 16]	14	5	1,1576	5,7880	0,0262
5.	(16 ; 20]	18	3	1,0992	3,2976	0,0513
6.	Total	×	25	×	30,7504	0,1588
7.	Simbol pentru total	×	$\sum_{i=1}^{K_x} n_i^x$	×	$\sum_{i=1}^{K_x} y_{x_i}^{'} \cdot n_i^x$	$\sum_{i=1}^{K_x} \left(y_{x_i}^{'} - \overline{y}_{x_i}\right)^2 \cdot n_i^x$

Valorile intermediare folosite în calculul dispersiei valorilor empirice ale variabilei dependente sunt prezentate în tabelul 9.11. Media aritmetică a valorilor empirice reprezintă:

$$\overline{y} = \frac{\sum_{j=1}^{K_j} y_j^{'} \cdot n_j^y}{\sum_{j=1}^{K_y} n_j^y} = \frac{30{,}75}{25} = 1{,}23 \text{ buc}$$

Dispersia valorilor empirice reprezintă:

$$\sigma_{y_i}^2 = \frac{\sum_{j=1}^{K_y} \left(y_j^{'} - \overline{y}\right)^2 \cdot n_j^y}{\sum_{j=1}^{K_y} n_j^y} = \frac{0{,}32}{25} = 0{,}0128 \text{ buc}^2$$

Coeficientul de determinare are valoarea: $\eta_d^2 = \frac{\sigma_{y_{x_i}}^2}{\sigma_{y_i}^2} = \frac{0{,}0064}{0{,}0128} = 0{,}4963$,

ceea ce înseamnă că în medie doar 49,63% din valoarea numărului mediu de rebuturi este datorată influenței vechimii în muncă, restul de 50,37% datorându-se altor factori.

Tabelul 9.11. Valori intermediare utilizate în calculul dispersiei valorilor empirice

Nr. crt.	Interval de variație $(y_{j-1} - y_j)$ [ani]	n_j^y	$y_j^{'}$ [buc.]	$y_j^{'} \cdot n_j^y$ [buc.]	$(y_j^{'} - \overline{y})^2 \cdot n_j^y$ [buc.^{2}]
(0)	(1)	(2)	(3)	(4) = (3) × (2)	(6)
1.	(1,0 ; 1,1]	4	1,05	4,20	0,1296
2.	(1,1 ; 1,2]	5	1,15	5,75	0,0320
3.	(1,2 ; 1,3]	10	1,25	12,50	0,0040
4.	(1,3 ; 1,4]	4	1,35	5,40	0,0576
5.	(1,4 ; 1,5]	2	1,45	2,90	0,0968
6.	Total	25	×	30,75	0,32
7.	Simbol pentru total	$\sum_{j=1}^{K_y} n_j^y$	×	$\sum_{j=1}^{K_y} y_j^{'} \cdot n_j^y$	$\sum_{j=1}^{K_y} \left(y_j^{'} - \overline{y}\right)^2 \cdot n_j^y$

9.4.5. Raportul de corelație

Raportul de corelație este o mărime, notată cu η_d, care poate fi obținută extrăgând rădăcina pătrată din coeficientul de determinare:

$$\eta_d = \sqrt{\eta_d^2} \qquad (9.49.)$$

La fel ca şi coeficientul de determinare, raportul de corelație are un domeniu de variație reprezentat de intervalul [0 ; 1]. Valoarea sa este cu atât mai mare cu cât intensitatea legăturii dintre cele două variabile este mai mare.

Se poate demonstra că dacă între cele două variabile poate fi stabilită o legătură liniară atunci valoarea raportului de corelație este egală cu valoarea absolută a coeficientului de corelație liniară simplă:

$$\eta_d = \left| r_{xy} \right| \qquad (9.50.)$$

Capitolul 10 - Analiza seriilor de timp

10.1. Coordonate ale analizei seriilor de timp

Analiza seriilor de timp are ca obiect studiul dinamicii fenomenelor colective, prin evidențierea transformărilor suferite de acestea sub impactul factorilor de influență. Pentru un astfel de demers trebuie folosite procedee şi mărimi specifice, care să exprime evoluțiile unor caracteristici.

Adeseori, factorii care influențează un fenomen colectiv se manifestă diferențiat în timp. Din această perspectivă se poate face următoarea clasificare a factorilor de influență:

- factori de influență continuă;
- factori de influență oscilantă;
- factori de influență aleatoare.

1. **Factorii de influență continuă** îşi exercită impactul în mod constant pentru toată durata acoperită de seria în timp. Influența acestor factori dă direcția generală a evoluției, numită *trend*.

2. **Factorii de influență oscilantă** îşi exercită impactul în mod discontinuu, dar cu regularitate, la intervale de timp relativ egale. În funcție de lungimea acestor intervale de timp se pot delimita două categorii de factori de influență oscilantă:

- *factori ciclici*, care se manifestă la intervale de timp (numite *cicluri*) mai mari de un an;
- *factori sezonieri* care se manifestă la intervale de timp (numite *sezoane*) mai mici de un an.

Efectele pe care factorii de influență oscilantă le au asupra fenomenelor colective sunt numite *mişcări ciclice (ondulatorii)* în cazul factorilor ciclici şi *variații sezoniere* în cazul factorilor sezonieri.

3. **Factorii de influență aleatorie** îşi exercită impactul în mod discontinuu şi neregulat. Efectul pe care aceşti factori îl au asupra unui fenomen colectiv este numit *variație reziduală*.

Pentru relevarea efectelor acestor tipuri de factori sunt folosite diferite modele ale fenomenelor colective. În acest subcapitol vom prezenta două astfel de modele, utilizate destul de frecvent în practică:

a) modelul aditiv;
b) modelul multiplicativ.

a) **Modelul aditiv** este descris de ecuația:

$$y_i = y_{T_i} + y_{C_i} + y_{S_i} + y_{R_i} \qquad (10.1.)$$

în care:

- y_i este valoarea caracteristicii y la un moment de timp (sau pentru un interval de timp) i;
- y_{T_i} este trendul inclus în valoarea y_i;

- y_{C_i} este mişcarea ciclică inclusă în valoarea y_i;
- y_{S_i} este variația sezonieră inclusă în valoarea y_i;
- y_{R_i} este variația reziduală inclusă în valoarea y_i;

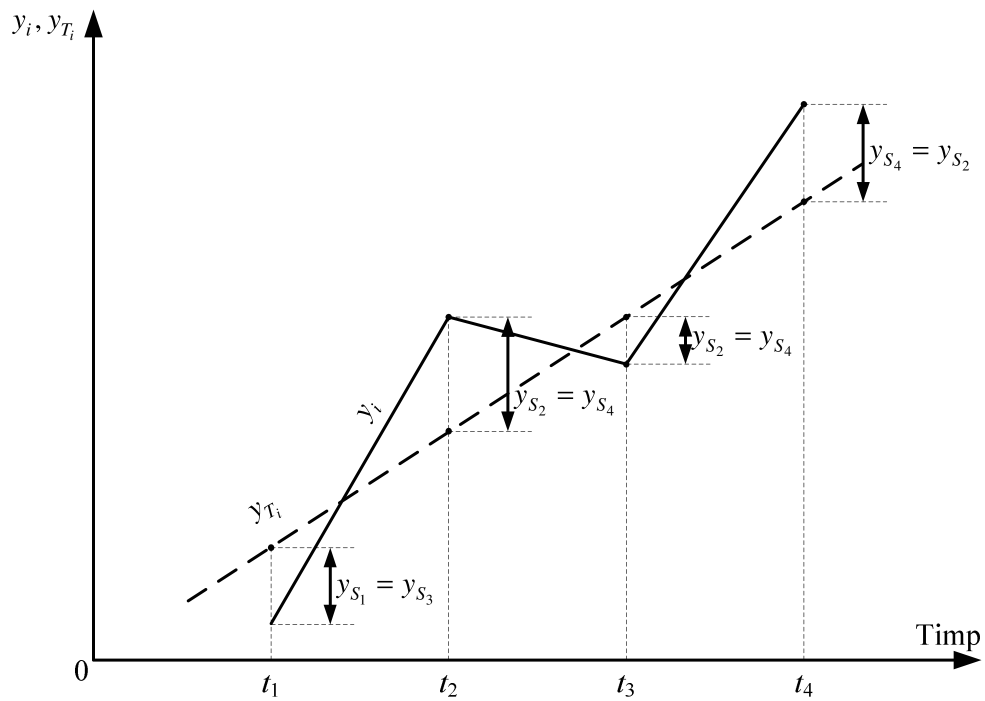


Fig. 10.1. Model aditiv asupra evoluției valorilor unei caracteristici

În practică, delimitarea mişcării ciclice este în general foarte dificilă, necesitând observații îndelungate asupra fenomenului studiat. Din acest motiv, adeseori se face abstracție de mişcarea ciclică, astfel încât ecuația modelului aditiv devine: $y_i = y_{T_i} + y_{S_i} + y_{R_i}$ (10.2.)

Într-o serie de aplicații practice ale modelului aditiv se porneşte de la premisa că variația reziduală poate fi neglijabilă în raport cu evoluția în ansamblu a fenomenului studiat. Dacă se face abstracție şi de acest element rezultă că valoarea caracteristicii studiate este egală cu suma dintre trend şi variația sezonieră: $y_i = y_{T_i} + y_{S_i}$ (10.3.)

Tot din considerente de simplicitate se consideră că unor diviziuni similare ale sezonului le corespund variații sezoniere egale. În figura 10.1. este prezentat modelul aditiv pentru evoluția unei caracteristici timp de două sezoane. Variațiile sezoniere din momentele t_1 şi t_2, care desemnează începuturile de sezoane, sunt egale, aşa cum sunt şi variațiile sezoniere din momentele t_3 şi t_4, care desemnează centrele celor două sezoane.

b) **Modelul multiplicativ** este descris de ecuația:

$$y_i = y_{Ti} * r_{Ci} * r_{Si} * r_{Ri} \quad (10.4.)$$

în care:

- r_{C_i} este o rație ce reflectă efectul factorilor ciclici în momentul de timp (sau intervalul de timp) i;

- r_{S_i} este o rație ce reflectă efectul factorilor sezonieri în momentul de timp (sau intervalul de timp) *i*;
- r_{R_i} este o rație ce reflectă efectul factorilor aleatorii în momentul de timp (sau intervalul de timp) *i*.

Atunci când se face abstracţie de mişcarea ciclică se consideră că $r_{C_i} = 1$, iar ecuația modelului devine:

$y_i = y_{Ti} * r_{Si} * r_{Ri}$ (10.5.)

De asemenea, atunci când se neglijează impactul factorilor aleatori, se consideră că $r_{R_i} = 1$, astfel încât valoarea y_i este dată de produsul dintre trend şi rația ce reflectă variația sezonieră:

$y_i = y_{Ti} * r_{Si}$ (10.6.)

Pentru unele aplicații practice ale modelului multiplicativ se consideră că unor diviziuni similare ale sezonului le corespund valori egale ale ratelor ce reflectă factorii sezonieri.

10.2. Indicatori ai analizei seriilor de timp

În raport cu modul de exprimare, indicatorii utilizați în analiza seriilor de timp pot fi grupați în trei categorii:

- indicatori absoluți;
- indicatori relativi;
- indicatori medii.

10.2.1. Indicatorii absoluți ai seriilor de timp

Indicatorii absoluți sunt mărimi exprimate în unitatea de măsură a caracteristicii studiate, al căror calcul nu implică mijlocirea unor alți indicatori. Printre indicatorii absoluți utilizați relativ frecvent în practică pentru caracterizarea seriilor în timp se numără:

- indicatorul de nivel;
- modificarea absolută.

a) **Indicatorul de nivel** este o mărime, notată cu y_i, care exprimă valoarea caracteristicii *y* la un moment de timp (sau pentru un interval de timp) *i*. Valorile acestei mărimi, care rezultă din observările statistice şi din prelucrările primare ale datelor, se află, practic, la baza calculului tuturor celorlalți indicatori de analiză a seriilor în timp.

b) **Modificarea absolută** este o mărime, notată cu Δ_{ij}, ce exprimă diferența dintre valorile indicatorului de nivel la două momente de timp *i* şi *j*:

$\Delta_{ij} = y_i - y_j$ (10.7.)

Prin intermediul modificării absolute se pot face comparații între stările unui fenomen la două momente de timp diferite apreciindu-se astfel sensul şi amploarea evoluției. Dintre cele două momente de timp, primul, în ordine cronologică, este numit *bază de comparație*, iar al doilea este numit *termen curent*.

În funcție de valoarea modificării absolute se pot stabili sensurile evoluției între cele două momente de timp:

- creştere, pentru o valoare pozitivă;
- scădere, pentru o valoare negativă;
- stagnare, pentru o valoare nulă.

Pentru analiza unei serii în timp se poate folosi un sistem de modificări absolute în care fiecare moment al seriei este folosit drept termen curent. În funcție de modul de alegere a bazei de comparație se pot delimita două tipuri de sisteme de modificări absolute:

sisteme de modificări absolute cu baza fixă; sisteme de modificări absolute cu baza în lanț.

1. Un **sistem de modificări absolute cu baza fixă** presupune ca pentru toți termenii seriei să se folosească o singură bază de comparație, care corespunde, de regulă, primului moment de timp. În acest caz modificarea absolută este dată de relația: $\Delta_{i/1} = y_i - y_1$ (10.8.)

2. Un **sistem de modificări absolute cu baza în lanț** presupune ca fiecare termen al seriei, cu excepția primului, să fie comparat ca termenul anterior. O modificare absolută cu baza în lanț poate fi calculată prin formula: $\Delta_{i/i-1} = y_i - y_{i-1}$ (10.9.)

Indicatorii relativi ai seriilor în timp sunt mărimi adimensionale obținute prin raportarea valorilor a doi indicatori. Printre indicatorii relativi utilizați frecvent în analiza seriilor în timp se numără:

a. indicele dinamicii;

b. ritmul dinamicii.

a) **Indicele dinamicii** este o mărime, notată cu $I_{i/j}$, care exprimă raportul dintre valorile indicatorului de nivel la două momente de timp i şi j:

$$I_{i/j} = \frac{y_i}{y_j} \qquad (10.10.)$$

Interpretarea indicelui dinamicii este oarecum asemănătoare interpretării modificării absolute. Primul moment de timp, în ordine cronologică, este numit bază de comparație, iar al doilea este numit termen curent. Caracteristica studiată înregistrează o creştere, atunci când indicele dinamicii este supraunitar, o scădere, când are o valoare subunitară şi o stagnare pentru o valoare unitară. Pentru analiza unei serii în timp se pot folosi două tipuri de sisteme de indici ai dinamicii:

- sisteme de indici ai dinamicii cu bază fixă;
- sisteme de indici ai dinamicii cu baza în lanț.

1. Într-un **sistem de indici ai dinamicii cu bază fixă** se foloseşte pentru toți termenii seriei în timp o singură bază de comparație. De regulă, aceasta corespunde primului termen al seriei. În acest caz, indicele dinamicii poate fi calculat prin formula:

$$I_{i/1} = \frac{y_i}{y_1} \qquad (10.11.)$$

2. Într-un **sistem de indici ai dinamicii cu bază în lanț** fiecare termen al seriei, cu excepția primului, este comparat cu termenul anterior. Un indice al dinamicii cu baza în lanț este dat de relația:

$$I_{i/i-1} = \frac{y_i}{y_{i-1}} \qquad (10.12.)$$

b) **Ritmul dinamicii** este o mărime, notată cu $R_{i/j}$, care poate fi obținută raportând o modificare absolută la valoarea folosită drept bază de comparație:

$$R_{i/j} = \frac{\Delta_{i/j}}{y_j} = I_{i/j} - 1 \qquad (10.13.)$$

Amploarea evoluției caracteristicii studiate este cu atât mai mare cu cât valoarea absolută a ritmului de creştere (scădere) este mai mare.

Pentru analiza unei serii în timp pot fi folosite sisteme de ritmuri ale dinamicii cu bază fixă sau cu bază în lanț, după cum modificările absolute sunt calculate ca baza fixă sau în lanț.

Adeseori ritmul dinamicii este exprimat într-o formă procentuală. Este cazul ratei inflației care reprezintă ritmul creşterii procentuale a prețurilor.

10.2.3. Indicatori medii ai seriilor în timp

Un indicator mediu exprimă nivelul general, pentru toată seria în timp, al unui indicator absolut sau relativ. printre indicatorii medii utilizați destul de frecvent în practică pentru caracterizarea seriilor în timp se numără:

a) indicatorul mediu de nivel;
b) modificarea absolută medie;
c) indicele mediu al dinamicii;
d) ritmul mediu.

a) **Indicatorul mediu de nivel** este o mărime, notată cu $\bar{y}_{C_r}$, care exprimă valoarea medie, pentru toată perioada acoperită de seria în timp, a indicatorului de nivel y_i. Această mărime poate fi calculată ca o medie aritmetică a valorilor indicatorului de nivel atunci când acestea corespund unor diviziuni egale ca lungime ale perioadei de timp acoperită de serie:

$$y_{C_r} = \frac{\sum_{i=1}^{N} y_i}{N} \quad (10.14.)$$

unde N este numărul termenilor seriei.

În situația în care valorile indicatorului de nivel corespund unor momente de timp aflate la distanțe inegale, indicatorul mediu de nivel este calculat ca o medie aritmetică ponderată cu lungimile intervalelor dintre momentele de timp:

$$y_{C_r} = \frac{\frac{y_1 + y_2}{2} \cdot t_1 + \frac{y_2 + y_3}{2} \cdot t_2 + \ldots + \frac{y_{N-1} + y_N}{2} \cdot t_{N-1}}{t_1 + t_2 + \ldots + t_N} =$$

$$= \frac{y_1 \cdot \frac{t_1}{2} + y_2 \cdot \frac{t_1 + t_2}{2} + \ldots + y_{N-1} \cdot \frac{t_{N-2} + t_{N-1}}{2} + y_N \cdot \frac{t_{N-1}}{2}}{t_1 + t_2 + \ldots + t_N} \qquad (10.15)$$

unde t_1, t_2, … , t_N reprezintă lungimile intervalelor de timp la care se înregistrează valorile y_i.

b) **Modificarea absolută medie** este mărime, notată cu $\overline{\Delta}$, calculată ca o medie aritmetică a tuturor mărimilor absolute cu baza în lanț:

$$\overline{\Delta} = \frac{\Delta_{2/1} + \Delta_{3/2} + \ldots + \Delta_{N/N-1}}{N-1} = \frac{\sum_{i=2}^{N} \Delta_{i/i-1}}{N-1} \qquad (10.16.)$$

Din această formulă de calcul se poate deduce legătura dintre modificarea absolută medie şi modificarea absolută cu bază fixă pentru ultimul termen al seriei:

$$\overline{\Delta} = \frac{(y_2 - y_1) + (y_3 - y_2) + \ldots + (y_{N-1} - y_{N-2}) + (y_N - y_{N-1})}{N-1} =$$

$$= \frac{y_N - y_1}{N-1} = \frac{\Delta_{N/1}}{N-1} \qquad (10.17)$$

c) **Indicele mediu al dinamicii** este o mărime, notată cu $\bar{I}$, calculată ca o medie geometrică a indicilor dinamicii cu baza în lanț determinați pentru întreaga serie:

$$\bar{I} = \sqrt[N-1]{I_{2/1} \times I_{3/2} \times \ldots \times I_{N/N-1}} = \sqrt[N-1]{\prod_{i=2}^{N} I_{i/i-1}} \qquad (10.18.)$$

Formula de calcul a indicelui mediu al dinamicii permite evidențierea legăturii dintre această mărime şi indicele dinamicii cu baza fixă pentru ultimul termen al seriei:

$$\bar{I} = \sqrt[N-1]{\frac{y_2}{y_1} \times \frac{y_3}{y_2} \times \ldots \times \frac{y_{N-1}}{y_{N-2}} \times \frac{y_N}{y_{N-1}}} = \sqrt[N-1]{\frac{y_N}{y_1}} = \sqrt[N-1]{I_{N/1}} \qquad (10.19.)$$

d) **Ritmul mediu al dinamicii** este o mărime, notată cu $\overline{R}$, care poate fi calculată prin relația:

$$\overline{R} = \bar{I} - 1 \qquad (10.19.)$$

Exemplul 10.2. În tabelul 10.1. este prezentată o serie în timp care exprimă volumul vânzărilor realizate de o firmă pentru un sortiment de produs în primele cinci luni ale anului 2006. Se cere să se calculeze următorii indicatori ai acestei serii în timp:

a) indicatorii absoluți;
b) indicatorii relativi;
c) indicatorii medii.

Tabelul 10.1. Volumul vânzărilor înregistrat de o firmă în primele cinci luni ale anului 2006

Nr. crt.	Luna	Volumul vânzărilor (y_i) [mii buc.]
(0)	(1)	(2)
1.	Ianuarie	1,50
2.	Februarie	1,45
3.	Martie	1,60
4.	Aprilie	1,70
5.	Mai	1,75

Rezolvare:

a) Indicatorii absoluți ai seriei în timp

Valorile indicatorului de nivel (altfel spus, valorile lunare ale volumului vânzărilor) sunt prezentate în coloana cu numărul de ordine 2 din tabelul 10.1.

Modificările absolute cu bază fixă, prezentate în coloana cu numărul de ordine 3 din tabelul 10.2. au fost calculate prin formula:

$$\Delta_{i/1} = y_i - y_1$$

Modificările absolute cu baza în lanț, prezentate în coloana cu numărul de ordine 4 din tabelul 10.2. au fost determinate pe baza relației:

$$\Delta_{i/i-1} = y_i - y_{i-1}$$

b) Indicatori relativi ai seriei în timp

Indicii dinamicii cu baza fixă sunt prezentați în coloana cu numărul de ordine 5 din tabelul 10.2. Aceste valori au fost calculate prin formula: $I_{i/1} = \dfrac{y_i}{y_1}$

Indicii dinamicii cu baza în lanț sunt prezentați în coloana cu numărul de ordine 6 din tabelul 10.2. Pentru determinarea acestor valori a fost utilizată relația:

$$I_{i/i-1} = \frac{y_i}{y_{i-1}}$$

Ritmurile dinamicii cu baza fixă sunt prezentate în coloana cu numărul de ordine 7 din tabelul 10.2. Calculul acestora are la bază formula: $R_{i/1} = \dfrac{\Delta_{i/1}}{y_1}$

Ritmurile dinamicii cu bază în lanț sunt prezentate în coloana cu numărul de ordine 8 din tabelul 10.2. În determinarea acestora a fost folosită formula: $R_{i/i-1} = \dfrac{\Delta_{i/i-1}}{y_{i-1}}$

c) Indicatori medii ai seriei de timp

Pentru calculul indicatorului mediu de nivel se consideră că toate cele cinci luni au un număr egal de zile, astfel încât se poate aplica formula:

$$\bar{y}_{Cr} = \frac{\sum_{i=1}^{N} y_i}{N} = \frac{8,00}{5} = 1,60 \text{ mii buc.}$$

Modificarea absolută medie reprezintă: $\bar{\Delta} = \frac{\sum_{i=2}^{N} \Delta_{i/i-1}}{N-1} = \frac{0,25}{5-1} = 0,0625$ mii buc.

Tabelul 10.2. Indicatori absoluți şi relativi ai seriei în timp

Nr. crt.	Luna	Indicator de nivel (y_i) [mii buc.]	Modificări absolute		Indici ai dinamicii		Ritmul dinamicii	
			cu bază fixă ($\Delta_{i/1}$)	cu bază în lanț ($\Delta_{i/i-1}$)	cu bază fixă ($I_{i/1}$)	cu bază în lanț ($I_{i/i-1}$)	cu bază fixă ($R_{i/1}$)	cu bază în lanț ($R_{i/i-1}$)
(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1.	Ianuarie	1,50	×	×	×	×	×	×
2.	Februarie	1,45	– 0,05	– 0,05	0,9667	0,9667	– 0,0333	0,0333
3.	Martie	1,60	0,10	0,15	1,0667	1,1034	0,0667	0,1034
4.	Aprilie	1,70	0,20	0,10	1,1333	1,0625	0,1333	0,0625
5.	Mai	1,75	0,25	0,05	1,1667	1,0294	0,1667	0,0294
6.	Total	8,00	×	0,25	×	×	×	×
7.	Simbol pentru total	$\sum_{i=1}^{N} y_i$	×	$\sum_{i=2}^{N} \Delta_{i/i-1}$	×	×	×	×

Indicele mediu al dinamicii are valoarea:

$$\sqrt[N-1]{\prod_{i=2}^{N} I_{i/i-1}} = \sqrt[5-1]{0,9667 \times 1,1034 \times 1,0625 \times 1,0294} = 1,0393$$

Ritmul mediu al dinamicii reprezintă:

$$\bar{R} = \bar{I} - 1 = 1,0393 - 1 = 0,0393$$

10.3. Determinarea trendului unei serii de timp

10.3.1. Considerații generale asupra determinării trendului unei serii de timp

În general, determinarea trendului unei serii de timp este întreprinsă în scopul evidențierii efectelor unor factori care influențează continuu fenomenul studiat. Pe baza trendului pot fi analizate aspectele esențiale ale unei activități şi pot fi prognozate desfăşurările viitoare ale unor fenomene.

În cadrul analizelor unor fenomene în raport cu factorii care îi influențează în mod continuu se practică procedeul *ajustării seriilor de timp în raport cu trendul*, care constă în determinarea, pentru toate valorile seriilor, a componentelor datorate factorilor de influență continuă. Acest procedeu are mai multe variante:

- tehnica mediilor mobile;
- tehnica ajustării pe baza modificării absolute medii;
- tehnica ajustării pe baza indicelui mediu al dinamicii;
- tehnica ajustării pe baza unei funcții de regresie.

Valorile ajustate în raport cu trendul pot fi folosite în cadrul prognozelor prin extrapolare. Într-o prognoză prin extrapolare asupra manifestării unui fenomen colectiv se porneşte de la premisa că factorii care au influențat fenomenul în trecut vor avea în viitor un impact similar. În privința trendului, extrapolarea constă în determinarea valorilor prognozate prin procedee similare celor care au fost aplicate pentru ajustarea valorilor seriei în timp.

Valorile extrapolate ale trendului sunt combinate cu valorile extrapolate pentru mişcările ciclice şi pentru variațiile sezoniere şi reziduale, rezultând astfel valorile prognozate ale indicatorului de nivel. Adeseori în practică se consideră că impactul factorilor de influență oscilantă şi aleatorie este nesemnificativ în raport cu impactul factorilor de influență continuă, astfel încât valorile prognozate ale indicatorului de nivel $(\hat{y}_i)$ sunt date doar de valorile prognozate ale trendului $(\hat{y}_{T_i})$:

$$\hat{y}_i = \hat{y}_{T_i} \qquad (10.20)$$

Acuratețea unei valori prognozate prin extrapolarea trendului poate fi cunoscută doar după ce perioada pentru care s-a elaborat prognoza s-a încheiat, pe baza unei mărimi numită *eroare de prognoză*, notată cu $\varepsilon_{t_i}^P$ şi dată de relația:

$$\varepsilon_{t_i}^P = y_i - \hat{y}_i = y_i - \hat{y}_{T_i} \qquad (10.21.)$$

În momentul previziunii, eroarea de prognoză poate fi doar estimată în raport cu parametrii procedeului de ajustare. Drept estimator este folosit un indicator numit *abaterea medie pătratică a trendului față de indicatorul de nivel* notat cu $\sigma_{y/T}$ şi calculat ca o medie pătratică a diferențelor dintre valorile indicatorului de nivel şi valorile ajustate în raport cu trendul ale seriei în timp:

$$\sigma_{y/T} = \sqrt{\frac{\sum_{i=1}^{N}\left(y_i - y_{T_i}\right)}{N}} \qquad (10.22.)$$

Acuratețea unei prognoze este cu atât mai mare cu cât abaterea medie pătratică a trendului față de indicatorul de nivel este mai mică.

10.3.2. Ajustarea seriilor de timp prin tehnica mediilor mobile

Determinarea valorilor ajustate prin tehnica mediilor mobile are la bază premisa compensării, pentru mai multe momente succesive, a abaterilor de la trend cauzate de factorii cu influență oscilantă sau aleatorie. În acest fel, media aritmetică a unor termeni succesivi dintr-o serie în timp poate fi considerată un rezultat al factorilor cu influență continuă.

Prin aplicarea procedeului mediilor mobile, valoarea ajustată a unui termen dintr-o serie în timp este dată de media aritmetică a unui număr impar de termeni consecutivi, în care termenul ce trebuie ajustat ocupă poziția centrală.

Exemplul 10.2. Se cere să se ajusteze seria de timp prezentată în tabelul 10.1. prin tehnica mediilor mobile.

Rezolvare: Întrucât seria în timp are doar cinci termeni s-a ales ca mediile aritmetice să se calculeze pe baza a trei termeni. În tabelul 10.3. este prezentat modul de calcul al valorilor ajustate.

Tabelul 10.3. Ajustarea unei serii în timp prin tehnica mediilor mobile

Nr. crt.	Luna	Indice de nivel (y_i) [mii buc.]	Suma termenilor succesivi [mii buc.]	Valori ajustate [mii buc.]
(0)	(1)	(2)	(3)	(4) = (3)/3
1.	Ianuarie	1,50	×	×
2.	Februarie	1,45	4,55	1,5167
3.	Martie	1,60	4,75	1,5833
4.	Aprilie	1,70	5,05	1,6833
5.	Mai	1,75	×	x

Tehnica mediilor mobile este destul de simplă însă aplicarea ei este limitată la termenii pentru care media aritmetică poate fi calculată pe baza numărului stabilit de termeni succesivi (în exemplul 10.2. nu s-au putut ajusta valorile primului şi ultimului termen al seriei întrucât pentru acestea nu s-au putut determina medii aritmetice pe baza a trei termeni succesivi). Această tehnică are, în plus, dezavantajul că nu poate fi folosită în cadrul prognozelor.

10.3.3. Ajustarea seriilor de timp pe baza modificării absolute medii

Ajustarea pe baza modificării absolute medii este indicată pentru seriile în timp ale căror valori au o evoluție apropiată de cea a unei progresii aritmetice. Se poate considera că rata progresiei aritmetice este egală cu modificarea absolută medie astfel încât între valorile trendului pentru doi termeni succesivi ai seriei există relația:

$$y_{T_i}^{ma} = y_{T_{i-1}}^{ma} + \overline{\Delta} \qquad (10.23.)$$

În aplicarea procedeului se consideră că pentru primul termen al unei serii în timp valoarea ajustată coincide cu indicatorul de nivel:

$$y_{T_1}^{ma} = y_1 \qquad (10.24.)$$

Pentru ceilalți termeni, valorile ajustate pot fi determinate prin aplicări succesive ale relației (10.23.) sau prin formula:

$$y_{T_{i+1}}^{ma} = y_{T_i}^{ma} + i \times \overline{\Delta} \qquad (10.25.)$$

Modul de calcul al modificării absolute medii face ca şi pentru ultimul termen al seriei valoarea ajustată să coincidă cu indicatorul de nivel:

$$y_{T_N}^{ma} = y_N \qquad (10.26.)$$

Tehnica de ajustare a trendului pe baza modificării medii absolute poate fi folosită în cadrul prognozelor prin extrapolare atunci când se consideră că evoluția viitoare a fenomenului poate fi încadrată într-o progresie aritmetică. În acest caz, valoarea prognozată a indicatorului de nivel pentru un moment viitor de timp este dată de relația:

$$\hat{y}_{N+k}^{ma} = y_N + k \times \overline{\Delta} \qquad (10.27)$$

în care $N + k$ este indicele numeric atribuit momentului viitor în raport cu distanța în timp la care acesta se află față de ultimul termen al seriei.

Exemplul 10.3. Se cere să se ajusteze, pe baza modificării absolute medii, seria în timp prezentată în tabelul 10.1. Se cere, de asemenea, să se determine, prin extrapolare pe baza modificării absolute medii, valorile prognozate ale volumului vânzărilor în lunile iunie şi iulie făcând abstracție de mişcările ciclice şi de variațiile sezoniere sau reziduale.

Rezolvare: În exemplul 10.1. a fost determinată modificarea absolută medie $\overline{\Delta} = 0{,}0625$ mii buc. Valorile ajustate ale seriei în timp, prezentate în tabelul 10.4., au fost calculate pe baza relației: $y_{T_{i-1}}^{ma} = y_{T_i}^{ma} + 1 \times \overline{\Delta}$

Valoarea prognozată a volumului vânzărilor în luna iunie, pentru care se atribuie indicele numeric $N + k = 6$, reprezintă: $\hat{y}_6^{ma} = y_5 + 1 \times \overline{\Delta} = 1{,}75 + 0{,}0625 = 1{,}8125$ mii buc.

Pentru luna iulie, căreia i se atribuie un indice numeric $N + k = 7$, valoarea prognozată a volumului vânzărilor reprezintă: $\hat{y}_7^{ma} = y_5 + 2 \times \overline{\Delta} = 1{,}75 + 2 \times 0{,}0625 = 1{,}875$ mii buc.

Acuratețea prognozei poate fi estimată pe baza abaterii medii pătratice a trendului față de indicatorul de nivel:

$$\sigma_{y/T}^{ma} = \sqrt{\frac{\sum_{i=1}^{N}\left(y_i - y_{T_i}^{ma}\right)^2}{N}} = \sqrt{\frac{0{,}0135}{5}} = 0{,}052 \text{ mii bucăți.}$$

Tabelul 10.4. Ajustarea seriei în timp pe baza modificării absolute medii

Nr. crt.	Luna	y_i [mii buc.]	$y_{T_i}^{ma}$ [mii buc.]	$y_i - y_{T_i}^{ma}$ [mii buc.]	$(y_i - y_{T_i}^{ma})^2$ [(mii buc.)^2]
(0)	(1)	(2)	(3)	(4) = (2) – (3)	(5) = (4)^2
1.	Ianuarie	1,50	1,5000	–	–
2.	Februarie	1,45	1,5625	– 0,1125	0,0127
3.	Martie	1,60	1,6250	– 0,025	0,0006
4.	Aprilie	1,70	1,6875	0,0125	0,0002
5.	Mai	1,75	1,7500	–	–
Total	×	8,00	8,1250	×	0,0135
Simbol pentru total	×	$\sum_{i=1}^{N} y_i$	$\sum_{i=1}^{N} y_{T_i}^{ma}$	×	$\sum_{i=1}^{N}\left(y_1 - y_{T_i}^{ma}\right)$

10.3.4. Ajustarea seriilor în timp pe baza indicelui mediu al dinamicii

Ajustarea pe baza indicelui mediu al dinamicii este indicată pentru seriile în timp ale căror valori evoluează asemănător unei progresii geometrice. În acest caz se poate considera că rata progresiei geometrice este egală cu indicele mediu al dinamicii astfel încât pentru doi termeni succesivi ai seriei se poate stabili relația:

$$y_{T_i}^{id} = y_{T_{i-1}}^{id} \times \bar{I} \qquad (10.28.)$$

Atunci când procedeul este aplicat se consideră că pentru primul termen al seriei în timp valoarea ajustată coincide cu indicatorul de nivel:

$$y_{T_1}^{id} = y_1 \qquad (10.29.)$$

Pentru termenii următori, valorile ajustate pot fi calculate fie aplicând succesiv relația (10.28), fie prin formula:

$$y_{T_{i+1}}^{id} = y_{T_i}^{id} \times \left(\bar{I}\right)^i \qquad (10.30.)$$

Din modul de calcul al indicelui mediu al dinamicii rezultă că şi pentru ultimul termen al seriei valoarea ajustată coincide cu indicatorul de nivel:

$$y_{T_N}^{id} = y_N \qquad (10.31.)$$

Tehnica de ajustare a seriilor în timp pe baza indicelui mediu al dinamicii poate fi folosită în cadrul prognozelor prin extrapolare atunci când se consideră că evoluția viitoare a fenomenului poate fi încadrată într-o progresie geometrică ce are aceeaşi rată $\bar{I}$. În această situație, valoarea prognozată a indicatorului de nivel pentru un moment viitor de timp poate fi calculată prin formula:

$$y_{N+k}^{id} = y_N \times \left(\bar{I}\right)^k \qquad (10.32.)$$

în care k este indicele numeric atribuit momentului viitor în raport cu distanța în timp la care acesta se află de ultimul termen al seriei.

Exemplul 10.4. Se cere să se ajusteze, pe baza indicelui mediu al dinamicii, seria în timp prezentată în tabelul 10.1. Se cere, de asemenea, să se determine prin extrapolare pe baza indicelui mediu al dinamicii, valorile prognozate ale volumului vânzărilor în lunile iunie şi iulie făcând abstracție de mişcările ciclice şi de variațiile sezoniere sau reziduale.

Rezolvare: În exemplul 10.1. a fost determinat indicele mediu al dinamicii $\bar{I} = 1{,}0393$. Valorile ajustate ale seriei în timp, prezentate în tabelul 10.5., au fost determinate pe baza relației: $y_{T_{i+1}}^{id} = y_{T_i}^{id} \times \left(\bar{I}\right)^i$

Valoarea prognozată a volumului vânzărilor în luna iunie, pentru care s-a atribuit indicele numeric $N + k = 6$ reprezintă: $\hat{y}_6^{id} = y_5 \times \left(\bar{I}\right)^1 = 1{,}75 \times 1{,}0393 = 1{,}8188$ mii buc.

Pentru luna iulie, căreia i s-a atribuit un indice numeric $N + k = 7$, valoarea prognozată a volumului vânzărilor reprezintă: $\hat{y}_7^{id} = y_5 \times \left(\bar{I}\right)^2 = 1{,}75 \times (1{,}0393)^2 = 1{,}8903$ mii buc.

Acuratețea prognozei poate fi estimată pe baza abaterii medii pătratice a trendului față de indicatorul de nivel:

$$\sigma_{y/T_i}^{id} = \sqrt{\frac{\sum_{i=1}^{N}\left(y_i - y_{T_i}^{id}\right)^2}{N}} = \sqrt{\frac{0{,}0126}{5}} = 0{,}0502 \text{ mii bucăți}$$

Tabelul 10.5. Ajustarea seriei în timp pe baza indicelui mediu al dinamicii

Nr. crt.	Luna	y_i [mii buc.]	$y_{T_i}^{id}$ [mii buc.]	$y_i - y_{T_i}^{id}$ [mii buc.]	$(y_i - y_{T_i}^{id})^2$ [(mii buc.)^2]
(0)	(1)	(2)	(3)	(4) = (2) – (3)	(5) = (4)^2
1.	Ianuarie	1,50	1,5000	–	–
2.	Februarie	1,45	1,5590	– 0,1090	0,0119
3.	Martie	1,60	1,6202	– 0,0202	0,0004
4.	Aprilie	1,70	1,6839	0,0161	0,0003
5.	Mai	1,75	1,7500	–	–
Total	×	8,00	8,1131	×	0,0126
Simbol pentru total	×	$\sum_{i=1}^{N} y_i$	$\sum_{i=1}^{N} y_{T_i}^{id}$	×	$\sum_{i=1}^{N}\left(y_1 - y_{T_i}^{id}\right)$

10.3.4. Ajustarea seriilor de timp pe baza funcțiilor de regresie

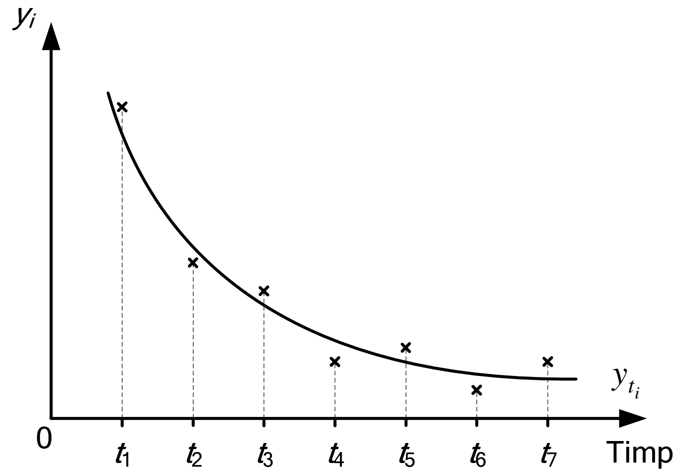


Fig. 10.2. Ajustarea unei serii în timp printr-o funcție de regresie

Ajustarea seriilor în timp pe baza funcțiilor de regresie este considerată cea mai riguroasă dintre tehnicile de determinare a trendului, aplicabilă pentru toate situațiile. Procedeul are la bază exprimarea timpului printr-o variabilă numerică şi reflectarea dependenței față de această variabilă a unei variabile dată de valorile trendului. În acest scop este stabilită o funcție matematică ale cărei valori să fie apropiate de valorile seriei în timp (fig. 10.2.). Practic, această funcție matematică poate fi considerată o funcție de regresie, pentru care timpul are semnificația variabilei independente, trendul are semnificația valorilor teoretice ale variabilei dependente iar indicatorul de nivel are semnificația valorilor empirice ale aceleiaşi variabile dependente.

Dacă se notează cu t_i valorile variabilei independente care exprimă timpul şi cu y_{t_i} valorile teoretice ale variabilei dependente, atunci funcția de regresie f are forma:

$$f(t_i) = y_{t_i} = y_{T_i} \tag{10.33.}$$

Parametrii funcției de regresie rezultă din condiția ca pentru ansamblul observărilor statistice valorile teoretice y_{t_i} să fie cât mai apropiate de cele empirice y_i. Prin aplicarea metodei celor mai mici pătrate se obțin pentru funcțiile de regresii expresii similare celor determinate în cadrul analizei legăturilor dintre variabile:

- pentru o funcție liniară de forma $y_{t_i} = a + bt_i$, parametrii a şi b pot fi obținuți rezolvând sistemul:

$$\begin{cases} N \cdot a + b\sum_{i=1}^{N} t_i = \sum_{i=1}^{N} y_i \\ a\sum_{i=1}^{N} t_i + b\sum_{i=1}^{N} t_i^2 = \sum_{i=1}^{N} t_i \cdot y_i \end{cases} \quad (10.34.)$$

- pentru o funcție polinomială de ordinul doi, de forma $y_{t_i} = a_0 + a_1 \cdot t_i + a_2 \cdot t_i^2$, parametrii a_0, a_1 şi a_2 pot fi obținuți prin intermediul sistemului:

$$\begin{cases} N \cdot a_0 + a_1\sum_{i=1}^{N} t_i + a_2\sum_{i=1}^{N} t_i^2 = \sum_{i=1}^{N} y_i \\ a_0\sum_{i=1}^{N} t_i + a_1\sum_{i=1}^{N} t_i^2 + a_2\sum_{i=1}^{N} t_i^3 = \sum_{i=1}^{N} t_i \cdot y_i \\ a_0\sum_{i=1}^{N} t_i^2 + a_1\sum_{i=1}^{N} t_i^3 + a_2\sum_{i=1}^{N} t_i^4 = \sum_{i=1}^{N} t_i^2 \cdot y_i \end{cases} \quad (10.35.)$$

Valorile numerice ale variabilei independente t_i sunt stabilite în raport cu poziția momentelor sau intervalelor de timp pe care le reprezintă în cadrul perioadei acoperite de seria în timp. Atunci când termenii seriei corespund unor momente de timp aflate la distanțe egale sau unor intervale de timp egale, valorile numerice ale variabilei t_i sunt alese astfel încât diferențele dintre termenii succesivi să fie egale.

Ajustarea trendului pe baza unei funcții de regresie poate fi folosită în prognozele prin extrapolare, atunci când se consideră că evoluția viitoare a fenomenului poate fi încadrată funcției de regresie ce a fost utilizată în cadrul ajustării. În acest caz, momentelor sau intervalelor de timp pentru care se fac prognoze le sunt asociate valori ale variabilei t_i care reflectă distanța în timp față de ultimul termen al seriei.

Pentru seriile la care termenii sunt poziționați la distanțe egale de timp, se obişnuieşte ca valorile variabilei t_i să fie dispuse simetric în raport cu valoarea nulă. În acest fel, sumele valorilor t_i la puteri impare devin nule, ceea ce simplifică foarte mult rezolvarea ecuațiilor lui Fermat. Alegerea acestor valori comportă unele deosebiri în raport cu numărul par sau impar de termeni ai seriei. Din această perspectivă, tehnicile de ajustare a seriilor în timp pe baza funcțiilor de regresie pot fi împărțite în două categorii:

a) tehnici de ajustare pentru seriile în timp cu un număr impar de termeni;

b) tehnici de ajustare pentru seriile în timp cu un număr par de termeni.

a) Pentru **seriile cu un număr impar de termeni**, în scopul simplificării calculelor, se poate atribui o valoare nulă variabilei t_i a termenului central, diferențele dintre doi termeni succesivi fiind egale cu o unitate (fig. 10.3.).

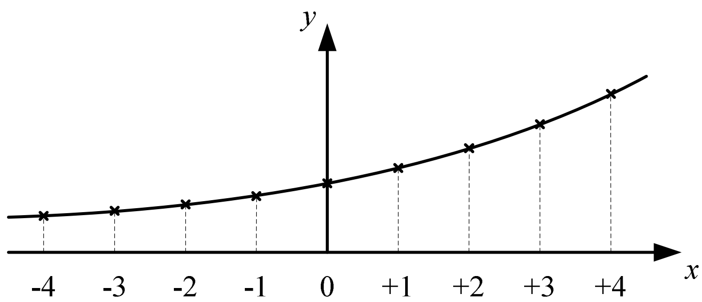


Fig. 10.3. Stabilirea valorilor variabilei t_i pentru o serie cu un număr impar de termeni

Exemplul 10.5. Se cere să se ajusteze, pe baza unei funcții liniare, seria în timp prezentată în tabelul 10.1. Se cere, de asemenea, ca pe baza acestei funcții să se prognozeze volumul vânzărilor în lunile iunie şi iulie, neglijând mişcările ciclice şi variațiile sezoniere sau reziduale.

Rezolvare: Valorile numerice ale variabilei t_i au fost stabilite astfel încât suma acestora să fie nulă. În acest scop, pentru termenul central, care corespunde lunii martie, a fost aleasă o valoare nulă, iar diferența dintre doi termeni succesivi a fost stabilită la o unitate.

Tabelul 10.6. Valori intermediare utilizate în calculul parametrilor funcției liniare de regresie

Nr. crt.	Luna	y_i [mii buc.]	t_i	t_i^2	$t_i \cdot y_i$ [mii buc.]
(0)	(1)	(2)	(3)	(4) = (3)^2	(5) = (3) × (2)
1.	Ianuarie	1,50	– 2	4	– 3,00
2.	Februarie	1,45	– 1	1	– 1,45
3.	Martie	1,60	0	0	0
4.	Aprilie	1,70	+ 1	1	1,70
5.	Mai	1,75	+ 2	4	3,50
Total	×	8,00	–	10	0,75
Simbol pentru total	×	$\sum_{i=1}^{N} y_i$	$\sum_{i=1}^{N} t_i$	$\sum_{i=1}^{N} t_i^2$	$\sum_{i=1}^{N} t_i \cdot y_i$

În tabelul 10.6. sunt prezentate valorile intermediare utilizate în calculul parametrilor funcției liniare de regresie. Aceştia rezultă din ecuațiile lui Fermat:

$$\begin{cases} Na + b\sum_{i=1}^{N} t_i = \sum_{i=1}^{N} y_i \\ a\sum_{i=1}^{N} t_i + b\sum_{i=1}^{N} t_i^2 = \sum_{i=1}^{N} t_i \cdot y_i \end{cases}$$

adică:

$$\begin{cases} 5a + 0 \cdot b = 8 \\ 0 \cdot a + 10 \cdot b = 0{,}75 \end{cases}$$

Rezolvând ecuațiile lui Fermat se obține: a = 1,6 mii buc.; b = 0,075 mii buc.

ceea ce înseamnă că funcția de regresie liniară are expresia: $y_{t_i} = 1{,}6 + 0{,}075 \cdot t_i$

Tabelul 10.7. Ajustarea seriei în timp pe baza funcției liniare de regresie

Nr. crt.	Luna	y_i [mii buc.]	t_i	$y_{T_i}^f = 1{,}6 + 0{,}075 \cdot t_i$ [mii buc.]	$y_i - y_{T_i}^f$ [mii buc.]	$(y_i - y_{T_i}^f)^2$ [(mii buc.)^2]
(0)	(1)	(2)	(3)	(4)	(5) = (2) – (4)	(6) = (5)^2
1.	Ianuarie	1,50	– 2	1,450	0,05	0,0025
2.	Februarie	1,45	– 1	1,525	– 0,075	0,0056
3.	Martie	1,60	0	1,600	–	–
4.	Aprilie	1,70	+1	1,675	0,025	0,0006
5.	Mai	1,75	+2	1,750	–	–
Total	×	8,00	–	8,00	–	0,0087
Simbol pentru total	×	$\sum_{i=1}^{N} y_i$	$\sum_{i=1}^{N} t_i$	$\sum_{i=1}^{N} y_{T_i}^f$	$\sum_{i=1}^{N} \left(y_i - y_{T_i}^f\right)$	$\sum_{i=1}^{N} \left(y_i - y_{T_i}^f\right)^2$

Valorile ajustate ale seriei în timp pe baza acestei funcții sunt prezentate în tabelul 10.7. Pentru prognoza volumului vânzărilor pe baza funcției de regresie valorile t_i sunt stabilite menținându-se diferența de o unitate dintre două luni succesive.

Lunii iunie, care se află la o distanță de o lună de ultimul termen al seriei în timp, i-a fost stabilită valoarea t_6 = 2 + 1 = 3. Atribuind această valoare argumentului funcției de regresie rezultă o valoare prognozată a volumului vânzărilor:

$\hat{y}_{iun}^f = y_{t_i}(3) = 1{,}6 + 0{,}075 \times 3 = 1{,}825$ mii buc. Pentru luna iulie, care se află la o distanță de două luni față de ultimul termen, a fost stabilită valoarea t_7 = 2 + 2 = 4. Pentru această valoare a argumentului funcției de regresie rezultă o valoare prognozată a volumului vânzărilor:

$$\hat{y}_{iul}^{f} = y_{t_i}(4) = 1{,}6 + 0{,}075 \times 4 = 1{,}9 \text{ mii buc.}$$

Acuratețea prognozei poate fi estimată pe baza abaterii medii pătratice a trendului față de indicatorul de nivel:

$$\sigma_{y/T_i}^{f} = \sqrt{\frac{\sum_{i=1}^{N}\left(y_i - y_{T_i}^{f}\right)^2}{N}} = \sqrt{\frac{0{,}0087}{5}} = 0{,}0418 \text{ mii buc.}$$

b) Pentru **seriile cu un număr par de termeni**, simplificarea calculelor poate fi obținută atribuind celor doi termeni centrali valorile de – 1 şi + 1, diferența dintre doi termeni centrali fiind egală cu două unități (fig. 10.4.).

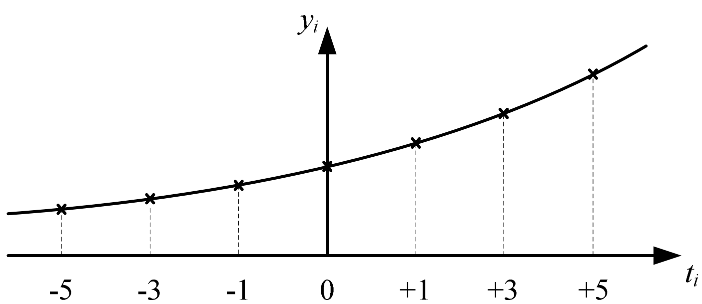


Fig. 10.4. Stabilirea valorilor variabilei t_i pentru o serie cu un număr par de termeni

Exemplul 10.6. În tabelul 10.8. este prezentată evoluția numărului de rebuturi înregistrat de o secție de producție a unei firme în primul semestru al anului 2006.

Tabelul 10.8. Evoluția numărului de rebuturi pentru un sortiment de produs în primul semestru al unui an

Nr. crt.	Luna	Număr de rebuturi [buc.]
(0)	(1)	(2)
1.	Ianuarie	43
2.	Februarie	41
3.	Martie	38
4.	Aprilie	35
5.	Mai	31
6.	Iunie	25

Se cere:

a) să se ajusteze seria în timp prin următoarele procedee:

a_1) ajustare pe baza modificării absolute medii;

a$_2$) ajustare pe baza indicelui mediu al dinamicii;
a$_3$) ajustare pe baza unei funcții liniare de regresie;
a$_4$) ajustare pe baza unei funcții liniare de regresie;

b) să se prognozeze, prin extrapolare pe baza celor patru procedee, numărul rebuturilor înregistrat în lunile iulie şi august făcând abstracție de mişcarea ciclică şi de variațiile sezoniere şi reziduale;
c) să se aprecieze, pe baza abaterii medii pătratice a trendului față de indicatorul de nivel, care dintre cele patru metode de prognoză are o acuratețe mai mare.

Rezolvare:

a) Ajustarea seriei în timp

a$_1$) Ajustare pe baza modificării absolute medii

Tabelul 10.9. Valori utilizate în ajustarea unei serii în timp pe baza modificării absolute medii

Nr. crt.	Luna	Număr de rebuturi (y_i) [buc.]	Modificări absolute [buc.]		$y_{T_i}^{id}$ [buc.]	$y_i - y_{T_i}^{id}$ [buc.]	$(y_i - y_{T_i}^{id})^2$ [buc.^{2}]
			cu bază fixă ($\Delta_{i/1}$)	cu bază în lanț ($\Delta_{i/-1}$)			
(0)	(1)	(2)	(3)	(4)	(5)	(6) = (2) – (5)	(7) = (6)^2
1.	Ianuarie	43	×	×	43,0	–	–
2.	Februarie	41	– 2,0	– 2,0	39,4	1,6	2,56
3.	Martie	38	– 5,0	– 3,0	35,8	2,2	4,84
4.	Aprilie	35	– 8,0	– 3,0	32,2	2,8	7,84
5.	Mai	31	– 12,0	– 4,0	28,6	2,4	5,76
6.	Iunie	25	– 18,0	– 6,0	25,0	–	–
Total	×	213	×	– 18,0	204,0	×	21,0
Simbol pentru total	×	$\sum_{i=1}^{N} y_i$	×	$\sum_{i=2}^{N} \Delta_{i/i-1}$	$\sum_{i=1}^{N} y_{T_i}^{ma}$	×	$\sum_{i=1}^{N} \left(y_i - y_{T_i}^{ma}\right)^2$

În tabelul 10.9. sunt prezentate valorile utilizate în ajustarea seriei în timp pe baza modificării absolute medii. Acest indicator reprezintă:

$$\overline{\Delta} = \frac{\sum_{i=2}^{N} \Delta_{i/i-1}}{N-1} = \frac{-18{,}0}{6-1} = -3{,}6 \text{ buc.}$$

Valorile ajustate ale seriei în timp au fost calculate prin formula:

$$y_{T_{i+1}}^{ma} = y_{T_i}^{ma} + i \cdot \overline{\Delta}$$

a$_2$) Ajustare pe baza indicelui mediu al dinamicii

În tabelul 10.10. sunt prezentate valorile utilizate în ajustarea seriei în timp pe baza indicelui mediu al dinamicii. Această mărime are valoarea:

$$\bar{I} = \sqrt[N-1]{\prod_{i=2}^{N} I_{i/i-1}} = \sqrt[6-1]{0{,}9535 \times 0{,}9268 \times 0{,}9211 \times 0{,}8857 \times 0{,}8065} = 0{,}8972$$

Tabelul 10.10. Valori utilizate în ajustarea unei serii în timp pe baza indicelui mediu al dinamicii

Nr. crt.	Luna	Număr de rebuturi (y_i) [buc.]	Indici ai dinamicii		$y_{T_i}^{id}$ [buc.]	$y_i - y_{T_i}^{id}$ [buc.]	$(y_i - y_{T_i}^{id})^2$ [buc.^{2}]
			cu bază fixă ($I_{i/1}$)	cu bază în lanț ($I_{i/i-1}$)			
(0)	(1)	(2)	(3)	(4)	(5)	(6) = (2) – (5)	(7) = (6)^2
1.	Ianuarie	43	×	×	43,0	–	–
2.	Februarie	41	0,9535	0,9535	38,58	2,42	5,86
3.	Martie	38	0,8837	0,9268	34,61	3,39	11,49
4.	Aprilie	35	0,8537	0,9211	31,06	3,94	15,52
5.	Mai	31	0,7209	0,8857	27,86	3,14	9,86
6.	Iunie	25	0,5814	0,8065	25,0	–	–
Total	×	213	×	×	200,11	×	42,73
Simbol pentru total	×	$\sum_{i=1}^{N} y_i$	×	×	$\sum_{i=1}^{N} y_{T_i}$	×	$\sum_{i=1}^{N} \left(y_i - y_{T_i}^{id}\right)^2$

Valorile ajustate ale seriei au fost determinate prin formula:

$$y_{T_i}^{id} = y_{T_{i-1}}^{id} \times \bar{I}$$

a_3) Ajustare pe baza unei funcții liniare de regresie

Valorile variabilei t_i au fost alese astfel încât suma acestora să fie nulă. În acest scop, celor doi termeni centrali, care corespund lunilor martie şi aprilie, le-au fost atribuite valorile – 1 respectiv + 1, în timp ce diferența pentru doi termeni succesivi a fost stabilită la două unități.

Tabelul 10.11. Valori utilizate în ajustarea unei serii în timp pe baza unei funcții liniare de regresie

Nr. crt.	Luna	y_i [buc.]	t_i	t_i^2	t_i^3	t_i^4	$t_i \cdot y_i$ [buc.]	$t_i^2 \cdot y_i$ [buc.]
(0)	(1)	(2)	(3)	(4) = $(3)^2$	(5) = $(3)^3$	(6) = $(3)^4$	(7) = (3) × (2)	(8) = (4) × (2)
1.	Ianuarie	43	– 5	25	– 125	625	– 215	1 075
2.	Februarie	41	– 3	9	– 27	81	– 123	369
3.	Martie	38	– 1	1	– 1	1	– 38	38
4.	Aprilie	35	+ 1	1	+ 1	1	35	35
5.	Mai	31	+ 3	9	+ 27	81	93	279
6.	Iunie	25	+ 5	25	+ 125	625	125	625
Total	×	213	–	70	–	1 414	– 123	2 421
Simbol pentru total	×	$\sum_{i=1}^{N} y_i$	$\sum_{i=1}^{N} t_i$	$\sum_{i=1}^{N} t_i^2$	$\sum_{i=1}^{N} t_i^3$	$\sum_{i=1}^{N} t_i^4$	$\sum_{i=1}^{N} t_i \cdot y_i$	$\sum_{i=1}^{N} t_i^2 \cdot y_i$

În tabelul 10.11. sunt prezentate valorile intermediare utilizate în determinarea parametrilor funcției liniare de regresie. Valorile acestora reies din ecuațiile lui Fermat.

$$\begin{cases} Na + b\sum_{i=1}^{N} t_i = \sum_{i=1}^{N} y_i \\ a\sum_{i=1}^{N} t_i + b\sum_{i=1}^{N} t_i^2 = \sum_{i=1}^{N} t_i \cdot y_i \end{cases}$$

adică:

$$\begin{cases} 6a + 0 \cdot b = 213 \\ 0 \cdot a + 70 \cdot b = -123 \end{cases}$$

Prin rezolvarea ecuațiilor lui Fermat se obține: a = 35,5 buc.; b = – 1,757 buc.
de unde rezultă că funcția de regresie liniară are expresia:

$$y_{t_i} = 35{,}5 - 1{,}757 \cdot t_i$$

În raport cu ecuația funcției de regresie liniară au fost determinate valorile ajustate ale seriei în timp care sunt prezentate în tabelul 10.12.

Tabelul 10.12. Ajustarea seriei în timp pe baza unei funcții liniare de regresie

Nr. crt.	Luna	y_i [buc.]	t_i	$y_{T_i}^f = 35{,}5 + 1{,}757 \cdot t_i$ [buc.]	$y_i - y_{T_i}^f$ [buc.]	$(y_i - y_{T_i}^f)^2$ [buc.^{2}]
(0)	(1)	(2)	(3)	(4)	(5) = (2) – (4)	(6) = (5)^2
1.	Ianuarie	43	– 5	44,285	– 1,285	1,6512
2.	Februarie	41	– 3	40,771	0,229	0,0524
3.	Martie	38	– 1	37,257	0,743	0,5520
4.	Aprilie	35	+1	33,743	1,257	1,5800
5.	Mai	31	+3	30,229	0,771	0,5944
6.	Iunie	25	+ 5	26,715	– 1,715	2,9412
Total	×	213	–	213,000	–	2,3712
Simbol pentru total	×	$\sum_{i=1}^{N} y_i$	$\sum_{i=1}^{N} t_i$	$\sum_{i=1}^{N} y_{T_i}^f$	$\sum_{i=1}^{N} \left(y_i - y_{T_i}^f\right)$	$\sum_{i=1}^{N} \left(y_i - y_{T_i}^f\right)^2$

a$_4$) Ajustare pe baza unei funcții de regresie polinomială de gradul doi

Pentru determinarea parametrilor unei funcții de regresie polinomială de gradul doi se folosesc valorile variabilei t_i care au fost stabilite pentru funcția de regresie liniară. Valorile parametrilor rezultă din ecuațiile lui Fermat.

$$\begin{cases} N \cdot a_0 + a_1 \sum_{i=1}^{N} t_i + a_2 \sum_{i=1}^{N} t_i^2 = \sum_{i=1}^{N} y_i \\ a_0 \sum_{i=1}^{N} t_i + a_1 \sum_{i=1}^{N} t_i^2 + a_2 \sum_{i=1}^{N} t_i^3 = \sum_{i=1}^{N} t_i \cdot y_i \\ a_0 \sum_{i=1}^{N} t_i^2 + a_1 \sum_{i=1}^{N} t_i^3 + a_2 \sum_{i=1}^{N} t_i^4 = \sum_{i=1}^{N} t_i^2 \cdot y_i \end{cases}$$

Introducând în aceste ecuații valorile intermediare prezentate în tabelul 10.12. se obține sistemul de ecuații:

$$\begin{cases} 6 \cdot a_0 + 0 \cdot a_1 + 70 \cdot a_2 = 213 \\ 0 \cdot a_0 + 70 \cdot a_1 + 0 \cdot a_2 = -123 \\ 70 \cdot a_0 + 0 \cdot a_1 + 1414 \cdot a_2 = 2.421 \end{cases}$$

de unde rezultă: a_0 = 36,7495 buc; a_1 = – 1,7571 buc; a_2 = – 0,1071 buc;

Pe baza ecuației de regresie: $y_{t_i} = 36{,}7495 - 1{,}75{,}71 \cdot t_i - 0{,}1071 \cdot t_i^2$

au fost determinate valorile ajustate ale seriei în timp, care sunt prezentate în tabelul 10.13.

Tabelul 10.13. Ajustarea seriei în timp pe baza unei funcții de regresie liniară de gradul doi

Nr. crt.	Luna	y_i [buc.]	t_i	$y_{T_i}^{fp}$ [buc.]	$y_i - y_{T_i}^{fp}$ [buc.]	$(y_i - y_{T_i}^{fp})^2$ [buc.^{2}]
(0)	(1)	(2)	(3)	(4)	(5) = (2) – (4)	(6) = (5)^2
1.	Ianuarie	43	– 5	42,8575	0,1425	0,0203
2.	Februarie	41	– 3	41,0569	– 0,0569	0,0032
3.	Martie	38	– 1	38,3995	– 0,3995	0,1596
4.	Aprilie	35	+1	34,8853	0,1147	0,0132
5.	Mai	31	+3	30,5143	0,4857	0,2359
6.	Iunie	25	+ 5	25,2865	– 0,2865	0,0821
Total	×	213	–	213,000	–	0,5143
Simbol pentru total	×	$\sum_{i=1}^{N} y_i$	$\sum_{i=1}^{N} t_i$	$\sum_{i=1}^{N} y_{T_i}^{fp}$	$\sum_{i=1}^{N}\left(y_i - y_{T_i}^{fp}\right)$	$\sum_{i=1}^{N}\left(y_i - y_{T_i}^{fp}\right)^2$

b) Prognoza prin extrapolare

b_1) Prognoza pe baza modificării absolute medii

Valorile prognozate pe baza modificării absolute medii pot fi calculate prin relația:

$$\hat{y}_{N+k}^{ma} = y_N + k \cdot \overline{\Delta}$$

Pentru luna iulie, căreia i se atribuie indicele numeric $N + k = 7$, valoarea prognozată a numărului de rebuturi reprezintă:

$$\hat{y}_7^{ma} = y_6 + 1 \cdot \overline{\Delta} = 25 + 1 \cdot (-3{,}6) = 21{,}4 \text{ buc.}$$

Valoarea prognozată a numărului de rebuturi pentru luna august, pentru care se atribuie indicele numeric $N + k = 8$, reprezintă:

$$\hat{y}_8^{ma} = y_6 + 2 \cdot \overline{\Delta} = 25 + 2 \cdot (-3{,}6) = 17{,}8 \text{ buc.}$$

b_2) Prognoza pe baza indicelui mediu al dinamicii

Valorile prognozate pe baza indicelui mediu al dinamicii pot fi determinate prin formula:

$$\hat{y}_{N+k}^{ma} = y_N \cdot \left(\bar{I}\right)^k$$

Pentru luna iulie, căreia i s-a atribuit indicele numeric $N + k = 7$, se prognozează un număr de rebuturi: $\hat{y}_7^{ma} = y_6 \times \left(\bar{I}\right)^1 = 25 \times 0{,}8972 = 22{,}43$ buc.

Valoarea prognozată a numărului de rebuturi din luna august, pentru care s-a atribuit indicele numeric $N + k = 8$, reprezintă: $\hat{y}_8^{ma} = y_6 \times \left(\bar{I}\right)^2 = 25 \times 0{,}8972^2 = 20{,}12$ buc.

b_3) Prognoza pe baza funcției liniare de regresie

Numărul de rebuturi poate fi prognozat pe baza funcției de regresie atribuind argumentului acesteia valori ale variabilei t_i stabilite în raport cu poziția în timp față de ultimul termen al seriei şi respectând diferența de două unități dintre doi termeni succesivi.

Pentru luna iulie s-a atribuit o valoare t_i =5 + 2 = 7, căreia îi corespunde o valoare prognozată a numărului de rebuturi:

$$\hat{y}^{f}_{iul} = y^{f}_{t_i}(7) = 35{,}5 - 1{,}757 \times 7 = 23{,}20 \text{ buc.}$$

Valoarea prognozată a numărului de rebuturi din luna august, pentru care s-a atribuit o valoare t_i = 5 + 2 × 2 = 9, reprezintă:

$$\hat{y}^{f}_{aug} = y^{f}_{t_i}(9) = 35{,}5 - 1{,}757 \times 9 = 19{,}69 \text{ buc.}$$

b₄) Prognoza pe baza unei funcții de regresie polinomială de gradul doi.

Pentru prognoza pe baza funcției de regresie polinomială de gradul doi pot fi folosite drept argument valorile variabilei t_i care au fost stabilite pentru prognoza pe baza unei funcții liniare de regresie.

Numărul de rebuturi prognozat pentru luna iulie reprezintă:

$$\hat{y}^{fp}_{iul} = y^{fp}_{t_i}(7) = 36{,}7495 - 1{,}7571 \times 7 - 0{,}1071 \times 7^2 = 19{,}20 \text{ buc.}$$

Pentru luna august a fost prognozat un număr de rebuturi care reprezintă:

$$\hat{y}^{fp}_{aug} = y^{fp}_{t_i}(9) = 36{,}7495 - 1{,}7571 \times 9 - 0{,}1071 \times 9^2 = 12{,}26 \text{ buc.}$$

c) Aprecierea acurateței prognozelor

Pe baza valorilor intermediare, calculate în cadrul ajustărilor, se pot determina abaterile medii pătratice ale trendului față de indicatorul de nivel pentru cele patru procedee:

- pentru prognoza pe baza modificării absolute medii:

$$\sigma^{ma}_{y/T_i} = \sqrt{\frac{\sum_{i=1}^{N}\left(y_i - y^{ma}_{T_i}\right)^2}{N}} = \sqrt{\frac{21{,}0}{6}} = 1{,}8708 \text{ buc.}$$

- pentru prognoza pe baza indicelui mediu al dinamicii:

$$\sigma^{id}_{y/T_i} = \sqrt{\frac{\sum_{i=1}^{N}\left(y_i - y^{id}_{T_i}\right)^2}{6}} = \sqrt{\frac{42{,}73}{6}} = 2{,}6686 \text{ buc.}$$

- pentru prognoza pa baza unei funcții liniare de regresie:

$$\sigma^{f}_{y/T_i} = \sqrt{\frac{\sum_{i=1}^{N}\left(y_i - y^{f}_{T_i}\right)^2}{6}} = \sqrt{\frac{7{,}3712}{6}} = 1{,}1084 \text{ buc.}$$

- pentru prognoza pe baza unei funcții de regresie polinomială de gradul doi:

$$\sigma^{fp}_{y/T_i} = \sqrt{\frac{\sum_{i=1}^{N}\left(y_i - y^{fp}_{T_i}\right)^2}{N}} = \sqrt{\frac{0{,}5143}{6}} = 0{,}2928 \text{ buc.}$$

Rezultă că prognoza pe baza unei funcții de regresie polinomială de gradul doi are cea mai mare acuratețe dintre procedeele utilizate.

Teste grilă

1. Fenomenele tipice au drept caracteristici:
a. sunt guvernate de aşa numite legi deterministe;
b. în condiții de mediu identice vor duce întotdeauna la aceleaşi rezultate;
c. în condiții de mediu identice pot conduce la rezultate diferite;
d. au in general mecanisme simple, cu un număr redus de factori;
e. au in general mecanisme complexe, cu factori de influență numeroşi, în care intervine hazardul;
f. rezultatele nu pot fi anticipate decât în condiții de incertitudine;
g. rezultatele pot fi anticipate în condiții de certitudine;
h. au o singură formă de manifestare;
i. au mai multe forme de manifestare.

R1: a, b, d, g, h.

2. Fenomenele colective au drept caracteristici:
a. sunt guvernate de aşa numite legi deterministe;
b. în condiții de mediu identice vor duce întotdeauna la aceleaşi rezultate;
c. în condiții de mediu identice pot conduce la rezultate diferite;
d. au in general mecanisme simple, cu un număr redus de factori;
e. au in general mecanisme complexe, cu factori de influență numeroşi, în care intervine hazardul;
f. rezultatele nu pot fi anticipate decât în condiții de incertitudine;
g. rezultatele pot fi anticipate în condiții de certitudine;
h. au o singură formă de manifestare;
i. au mai multe forme de manifestare .

R2: c, e, f, i.

3. Populația statistică este o noțiune reprezentată de:
a. o mulțime de elemente studiate pentru a se cerceta starea la un moment dat sau evoluția în timp a unuia sau mai multor fenomene;
b. un rezultat posibil sau o combinație de rezultate posibile, ale unui fenomen studiat;
c. o aplicație prin care fiecărui element al unui câmp de evenimente îi este asociată o valoare numerică.

R3: a.

4. O variabilă aleatoare este o noțiune reprezentată de:

a. o aplicație prin care fiecărui element al unui câmp de evenimente îi este asociată o valoare numerică;
b. un rezultat posibil sau o combinație de rezultate posibile, ale unui fenomen studiat;
c. însuşirile prin care sunt descrise, în cadrul unei cercetări, unitățile statistice.

R4: a.

5. Statistica aplicată are ca obiect:
a. formularea, pe baza principiilor ştiinței matematicii, a unor tehnici de cercetare statistică;
b. combinarea tehnicilor statistice cu procedee bazate pe inteligența artificială;
c. adaptarea tehnicilor statisticii matematice la condițiile concrete ale domeniilor în care sunt utilizate..

R5: c.

6. Culegerea datelor prin recensăminte are drept caracteristici:
a. presupune investigarea tuturor unităților populației statistice prin care se studiază un fenomen;
b. este expusă erorilor de reprezentativitate;
c. presupune investigarea unui eşantion;
d. presupune investigarea unei părți din populația statistică.

R6: a.

7. Culegerea datelor prin sondaje are drept caracteristici:
a. presupune investigarea tuturor unităților populației statistice prin care se studiază un fenomen;
b. este expusă erorilor de reprezentativitate;
c. presupune investigarea unui eşantion;
d. presupune investigarea unei părți din populația statistică.

R7: b, c, d.

8. O chestionare statistică constă în:
a. un ansamblu de întrebări adresate unor persoane cu privire la percepțiile şi reacțiile acestora față de un fenomen studiat;
b. înregistrarea unor aspecte ale manifestării unui fenomen cercetat;
c. provocarea, în mod artificial dar în condiții cât mai apropiate de cele naturale, a unui proces, pentru a i se putea studia manifestarea.

R8: a.

9. O observație statistică constă în:
a. un ansamblu de întrebări adresate unor persoane cu privire la percepțiile şi reacțiile acestora față de un fenomen studiat;
b. înregistrarea unor aspecte ale manifestării unui fenomen cercetat;
c. provocarea, în mod artificial dar în condiții cât mai apropiate de cele naturale, a unui proces, pentru a i se putea studia manifestarea.

R9: b.

10. Un experiment statistic constă în:
a. un ansamblu de întrebări adresate unor persoane cu privire la percepțiile şi reacțiile acestora față de un fenomen studiat;
b. înregistrarea unor aspecte ale manifestării unui fenomen cercetat;
c. provocarea, în mod artificial dar în condiții cât mai apropiate de cele naturale, a unui proces, pentru a i se putea studia manifestarea.

R10: c.

11. Un panel statistic constă în:
a. interogarea periodică a unui grup de persoane cu privire la un acelaşi fenomen;
b. provocarea, în mod artificial dar în condiții cât mai apropiate de cele naturale, a unui proces, pentru a i se putea studia manifestarea;
c. un ansamblu de chestionări efectuate concomitent.

R11: a.

12. O distribuție homogradă reprezintă:
a. o distribuție de frecvențe la care caracteristica atributivă este calitativă;
b. o distribuție de frecvențe la care caracteristica atributivă este cantitativă;
c. o serie simplă la care caracteristica atributivă este calitativă;
d. o serie simplă la care caracteristica atributivă este cantitativă;
e. o serie de timp la care caracteristica atributivă este cantitativă;
f. o serie de timp la care caracteristica atributivă este calitativă.

R12: a.

13. O distribuție heterogradă reprezintă:
a. o distribuție de frecvențe la care caracteristica atributivă este calitativă;
b. o distribuție de frecvențe la care caracteristica atributivă este cantitativă;
c. o serie simplă la care caracteristica atributivă este calitativă;
d. o serie simplă la care caracteristica atributivă este cantitativă;
e. o serie de timp la care caracteristica atributivă este cantitativă;
f. o serie de timp la care caracteristica atributivă este calitativă.

R13: b.

14. Printre valorile tipice utilizate pentru identificarea trăsăturilor esențiale ale fenomenelor colective se numără:
a. mărimile medii;
b. valoarea mediană;
c. modul;
d. media aritmetică;
e. media armonică;
f. varianța;
g. coeficientul de variație în raport cu abaterea medie pătratică;
h. coeficientul de asimetrie în raport cu modul;
i. coeficientul de asimetrie în raport cu mediana;
j. momentele centrate ale distribuțiilor heterograde;
k. coeficientul pearsonian al boltirii.

R14: a, b, c, d, e.

15. O valoare mediană reprezintă:
a. o mărime ce ocupă locul central într-o serie statistică ordonată;
b. un raport dintre suma valorilor şi numărul de unități statistice;
c. o mărime care exprimă valoarea cu cea mai mare frecvență din cadrul seriei.

R15: a.

16. Modul unei distribuții heterograde reprezintă:
a. o mărime ce ocupă locul central într-o serie statistică ordonată;
b. un raport dintre suma valorilor şi numărul de unități statistice;
c. o mărime care exprimă valoarea cu cea mai mare frecvență din cadrul seriei.

R16: c.

17. Un interval modal al unei distribuții heterograde reprezintă:
a. un interval cu frecvența mai mare decât cea a intervalelor învecinate;
b. un interval aflat într-o poziție centrală;
c. un interval aflat în una din extremitățile seriei.

R17: a.

18. Relația dintre dispersia unei serii statistice şi reprezentativitatea valorilor tipice ale acesteia poate fi formulată astfel :
a. cu cât dispersia seriei este mai mare, cu atât valorile tipice sunt mai puțin reprezentative;
b. cu cât dispersia seriei este mai mică, cu atât valorile tipice sunt mai puțin reprezentative;

c. cu cât dispersia seriei este mai mare , cu atât media aritmetică este mai reprezentativă.

R18: a.

19. O serie statistică este simetrică atunci când:
a. influența factorilor întâmplători asupra fenomenului colectiv studiat se produce cu regularitate;
b. media aritmetică este egală cu modul seriei;
c. coeficientul de asimetrie în raport cu mediana este nul.

R19: a, b, c.

20. O distribuție heterogradă este platykurtică atunci când:
a. curba de frecvențe este asemănătoare, în ceea ce priveşte aplatizarea, unei curbe de distribuție normală;
b. curba de frecvențe este mai ascuțită față de curba unei distribuții normale;
c. curba de frecvențe este mai turtită decât curba unei distribuții normale;
d. coeficientul pearsonian al boltirii este mai mic decât 3;
e. coeficientul pearsonian al boltirii este mai mare decât 3;
f. coeficientul pearsonian al boltirii este egal cu 3.

R20: c, d.

21. O distribuție heterogradă este mezokurtică atunci când:
a. curba de frecvențe este asemănătoare, în ceea ce priveşte aplatizarea, unei curbe de distribuție normală;
b. curba de frecvențe este mai ascuțită față de curba unei distribuții normale;
c. curba de frecvențe este mai turtită decât curba unei distribuții normale;
d. coeficientul pearsonian al boltirii este mai mic decât 3;
e. coeficientul pearsonian al boltirii este mai mare decât 3;
f. coeficientul pearsonian al boltirii este egal cu 3.

R21: a, f.

22. O distribuție heterogradă este leptokurtică atunci când:
a. curba de frecvențe este asemănătoare, în ceea ce priveşte aplatizarea, unei curbe de distribuție normală;
b. curba de frecvențe este mai ascuțită față de curba unei distribuții normale;
c. curba de frecvențe este mai turtită decât curba unei distribuții normale;
d. coeficientul pearsonian al boltirii este mai mic decât 3;
e. coeficientul pearsonian al boltirii este mai mare decât 3;
f. coeficientul pearsonian al boltirii este egal cu 3.

R22: b, e.

23. Inferența statistică reprezintă:
a. trecerea de la valorile certe ale parametrilor unui eşantion la valorile probabile ale parametrilor populației;
b. analiza statistică a parametrilor unui eşantion;
c. asocierea unor distribuții probabilistice pentru valorile parametrilor unei populații.

R23: a.

24. Sondajele aleatoare pot fi definite drept:
a. sondajele la care unitățile statistice ale eşantioanelor sunt alese în mod întâmplător;
b. sondajele la care unitățile statistice sunt stabilite în funcție de trăsăturile populației studiate, relevante în raport cu scopul cercetării statistice;
c. sondajele la care intervalele de încredere sunt stabilite aleatoriu.

R24: a.

25. În cadrul inferenței statistice, atunci când nu se cunoaşte dispersia populației studiate se recurge la estimarea acesteia pe baza:
a. dispersiei eşantionului;
b. mediei aritmetice a populației studiate;
c. volumului eşantionului.

R25: a.

26. Impactul dispersiei populației studiate asupra erorii efective de sondaj poate fi descris astfel:
a. cu cât populația studiată este mai omogenă, cu atât sunt mai mari şansele ca valorile estimate să fie apropiate de cele reale;
b. cu cât populația studiată este mai omogenă, cu atât sunt mai mici şansele ca valorile estimate să fie apropiate de cele reale;
c. cu cât dispersia populației studiate este mai mare, cu atât sunt mai mari şansele ca valorile estimate să fie apropiate de cele reale.

R26: a.

27. Impactul volumului unui eşantion asupra erorii efective de sondaj poate fi descris astfel:
a. cu cât volumul eşantionului este mai mare, cu atât sunt mai mari şansele ca valorile estimate să fie apropiate de cele reale;
b. cu cât volumul eşantionului este mai mic, cu atât sunt mai mari şansele ca valorile estimate să fie apropiate de cele reale;

c. cu cât volumul eşantionului are o pondere mai mare în volumul populației, cu atât sunt mai mici şansele ca valorile estimate să fie apropiate de cele reale.

R27: a.

28. În inferența statistică pentru sondajele de volum redus se utilizează drept distribuții probabilistice:
a. distribuția normală, cu condiția ca populația studiată să urmeze tot o distribuție normală;
b. distribuții t;
c. distribuții în formă de clopot;
d. distribuții în formă de J;
e. distribuții în formă de U.

R28: a, b, c.

29. În inferența statistică pentru sondajele de volum mare se utilizează drept distribuții probabilistice:
a. distribuția normală;
b. distribuții t;
c. distribuții în formă de clopot;
d. distribuții în formă de J;
e. distribuții în formă de U.

R29: a, c.

30. În cadrul verificării ipotezelor statistice, ipoteza nulă reprezintă:
a. o ipoteză care îmbracă forma aprecierii inițiale ;
b. o ipoteză care reprezintă opusul aprecierii inițiale;
c. ipoteza distribuției normale a valorilor estimate.

R30: a.

31. În cadrul verificării ipotezelor statistice, ipoteza alternativă reprezintă:
a. o ipoteză care îmbracă forma aprecierii inițiale ;
b. o ipoteză care reprezintă opusul aprecierii inițiale;
c. ipoteza distribuției normale a valorilor estimate.

R31: b.

32. O legătură cu o singură variabilă independentă este inversă atunci când:
a. cele două variabile evoluează în acelaşi sens;
b. variabilele evoluează în sensuri opuse;
c. legătura are intensitate maximă;
d. legătura este liniară.

R32: b.

33. În cadrul analizei dinamice se consideră că factorii de influenţă continuă îşi exercită impactul:

a. în mod constant pentru toată durata acoperită de seria în timp;
b. în mod discontinuu, dar cu regularitate, la intervale de timp relativ egale;
c. în mod discontinuu şi neregulat.

R33: a.

34. În cadrul analizei dinamice se consideră că factorii de influenţă oscilantă îşi exercită impactul:

a. în mod constant pentru toată durata acoperită de seria în timp;
b. în mod discontinuu, dar cu regularitate, la intervale de timp relativ egale;
c. în mod discontinuu şi neregulat.

R34: b.

35. În cadrul analizei dinamice se consideră că factorii de influenţă aleatorie îşi exercită impactul:

a. în mod constant pentru toată durata acoperită de seria în timp;
b. în mod discontinuu, dar cu regularitate, la intervale de timp relativ egale;
c. în mod discontinuu şi neregulat.

R35: c.

36. Categoria factoriilor de influenţă oscilantă cuprinde:
a. factori ciclici;
b. factori sezonieri;
c. factorii influenţă aleatorie.

R36: a, b.

37. În cadrul analizei dinamice se consideră că trendul este un rezultat al:
a. factorilor de influenţă continuă;
b. factorilor de influenţă oscilantă;
c. factorilor de influenţă aleatorie.

R37: a.

38. În cadrul analizei dinamice se consideră că variaţia reziduală este un rezultat al:
a. factorilor de influenţă continuă;
b. factorilor de influenţă oscilantă;

c. factorilor de influență aleatorie.

R38: c.

39. În cadrul analizei dinamice se consideră că mişcările ciclice (ondulatorii) sunt un rezultat al:
a. factorilor de influență continuă;
b. factorilor de influență oscilantă;
c. factorilor de influență aleatorie.

R39: b.

40. În cadrul analizei dinamice modificarea absolută este o mărime care exprimă;
a. valoarea caracteristicii studiate la un moment de timp (sau pentru un interval de timp);
b. diferența dintre valorile indicatorului de nivel la două momente de timp;
c. raportul dintre valorile indicatorului de nivel la două momente de timp.

R40: b.

41. În cadrul analizei dinamice indicatorul de nivel este o mărime care exprimă:
a. valoarea caracteristicii studiate la un moment de timp (sau pentru un interval de timp);
b. diferența dintre valorile caracteristicii studiate la două momente de timp;
c. raportul dintre valorile caracteristicii studiate la două momente de timp.

R41: c.

42. O valoare pozitivă a modificării absolute exprimă, în cadrul analizei dinamice:
a. creşterea între cele două momente de timp;
b. scăderea între cele două momente de timp;
c. stagnarea între cele două momente de timp.

R42: a.

43. O valoare negativă a modificării absolute exprimă, în cadrul analizei dinamice:
a. creşterea între cele două momente de timp;
b. scăderea între cele două momente de timp;
c. stagnarea între cele două momente de timp.

R43: b.

44. O valoare nulă a modificării absolute exprimă, în cadrul analizei dinamice:
a. creşterea între cele două momente de timp;
b. scăderea între cele două momente de timp;
c. stagnarea între cele două momente de timp.

R44: c.

45. În cadrul analizei dinamice o valoare supraunitară a indicelui dinamicii exprimă:
a. creşterea între cele două momente de timp;
b. scăderea între cele două momente de timp;
c. stagnarea între cele două momente de timp.

R45: a.

46. În cadrul analizei dinamice o valoare subunitară a indicelui dinamicii exprimă:
a. creşterea între cele două momente de timp;
b. scăderea între cele două momente de timp;
c. stagnarea între cele două momente de timp.

R46: b.

47. În cadrul analizei dinamice o valoare supraunitară a indicelui dinamicii exprimă:
a. creşterea între cele două momente de timp;
b. scăderea între cele două momente de timp;
c. stagnarea între cele două momente de timp.

R47: c.

48. Ajustarea seriilor în timp în raport cu trendul constă în:
a. determinarea, pentru toate valorile seriilor, a componentelor datorate factorilor de influență continuă;
b. determinarea, pentru toate valorile seriilor, a componentelor datorate factorilor de influență oscilantă;
c. determinarea, pentru toate valorile seriilor, a componentelor datorate factorilor de influență aleatorie.

R48: a.

49. Într-o prognoză prin extrapolare asupra manifestării unui fenomen colectiv se porneşte de la premisa că:
a. factorii care au influențat fenomenul în trecut vor avea în viitor un impact similar;

b. factorii care au influențat fenomenul în trecut nu vor mai avea nicio influență în viitor;
c. factorii care au influențat fenomenul în trecut vor avea în viitor un impact semnificativ diferit.

R49: a.

50. În cadrul analizei dinamice valoarea ritmului dinamicii se obține:
a. scăzând o unitate din valoarea indicelui dinamicii;
b. raportând o modificare absolută la valoarea folosită drept bază de comparație;
c. adunând o unitate din valoarea indicelui dinamicii.

R50: a, b.

Bibliografie selectivă

1. Biji Mircea, Biji Maria Elena, Lilea Eugenia, Anghelache Constantin, ***Tratat de statistică,*** Editura Economică, Bucuresti, 2003;
2. Curwin Jon, Slater Roger**,** ***Quantitative Methods for Business Decision***, Third Edition, Chapman&Hall, London, 1991;
3. Georgescu-Roegen Nicholas, ***Metoda statistică***, Editura Expert, Bucuresti, 1998;
4. Isac-Maniu Alexandru, Mitruț Constantin, Voineagu Vergil, ***Statistica pentru managementul afacerilor,*** Ediția a doua, Editura Economică, Bucuresti, 2003;
5. Jaba Elisabeta, ***Statistica economică***, Ediția a treia, Editura Economică, Bucuresti, 2003;
6. Lucey Terry, ***Quantitative Techniques*****,** 5th Edition, D.P. Publication, London, 1996.